Fast approximate solution of large dense linear programs

Leo Liberti, Pierre-Louis Poirion, Ky Vu

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

November 21, 2016

Abstract

We show how random projections can be used to solve large-scale dense linear programs approximately. This is a new application of techniques which are now fairly well known in probabilistic algorithms, but have never yet been systematically applied to the fundamental class of Linear Programming. We develop the necessary theoretical framework, and show that this idea works in practice by showcasing its effect on the quantile regression problem, where the state-of-the-art CPLEX solver fails on large or poorly scaled datasets.

1 Introduction

Linear Programming (LP) lies at the very heart of all of Mathematical Programming (MP): its main solution algorithms (the simplex algorithm and the interior-point method) are implemented as part of extremely advanced solver technology, e.g. IBM-ILOG CPLEX [7]. They can yield valid solutions for very large-scale instances in very short times. So much so, in fact, that "solve an LP" is now routinely found listed as an "elementary step" in many algorithms. We consider LPs in standard form:

$$\left.\begin{array}{rrcl} \min & c^\top x & & \\ & A\,x & = & b \\ & x & \geq & 0, \end{array}\right\} \qquad (1)$$

where $c \in \mathbb{R}^n$ is a given cost vector, x is vector of n (continuous) decision variables, A is an $m \times n$ matrix with $m \leq n$, and $b \in \mathbb{R}^m$.

Our idea is to pre-multiply A and b by a "short and fat" $k \times m$ carefully chosen random matrix T, so as to obtain fewer constraints $TA\,x = Tb$, and then solve the "shortened and fattened" LP

$$\min\{c^\top x \mid TA\,x = Tb \wedge x \geq 0\}. \qquad (2)$$

We shall argue in this paper that there exist matrices T such that Eq. (2) is likely to yield an approximate solution for Eq. (1).

The technique we propose is application-independent, and applies to all LPs (Eq. (1)). Since it is probabilistic with increasing success probability in function of instance size, it is most useful with large LP instances. Moreover, since it involves a pre-multiplication of the constraint matrix of the LP instance by a reasonably dense random matrix, the original LP is transformed into a dense LP having smaller size, but yielding approximately the same solution with "high probability". By this we mean that the probability approaches 1 as $1 - e^{-f}$ where f is an increasing function of the size of the instance. Since most LP solvers exploit sparsity, if the original instance is sparse but large, and the transformed instance is smaller but dense, our technique may not yield a performance improvement (though extremely recent — yet unpublished — experiments on network flow LPs seems to point out its usefulness even in case of huge, sparse instances).

Although the large majority of LP instances are sparse, there are a few important LP applications yielding dense instances, such as blending and diet [17, 5] or quantile regression [11] problems. Whereas most blending instances we have seen were not large-scale, quantile regression instances may be as large and dense as the databases they involve. We shall therefore consider quantile regression as our motivating application, and we shall see that size and ill-scaling of the underlying databases prevent quantile regression computations, even using state-of-the-art solvers such as CPLEX or statistical packages such as R [16].

We remark that there exist works in the literature that leverage random projections to solve quantile regression problems (e.g. [4]) but they exploit a different formulation (minimization of $\|Ax - b\|$) which lacks an essential element of the LP in Eq. (1), namely the restriction to the non-negative orthant $x \geq 0$. Showing that random projections preserve order relations is non-trivial: for example, if $Ax \leq b$ holds and T is an appropriately sized random matrix with entries from a normal distribution, there is no particular reason why $TA\,x \leq Tb$ should hold. This work is mainly a contribution to LP, but also marks important progress as regards the range of applicability of random projections.

As regards size and density of the instances our technique can tackle, our results point out that m in $O(10^3)$ and a constraint matrix density around 10-15% is already enough to observe improvements. A side benefit of our technique is that a poorly scaled quantile regression LP, unsolvable for CPLEX due to ill scaling, transforms into a well-scaled LP. Since many database tables have columns containing absolute magnitudes (e.g. for gross revenue or expenditures, often in the millions or billions) as well as percentages (often encoded as fractions in $[0, 1]$), it is not hard to see that ill scaling is common in quantile regression problems. If the statistical model being considered is linear, it might be possible to partly remove ill-scaling by judicious pre- and post-processing; but polynomial models that multiply data columns generally make such processing impossible. We therefore think that our method also contributes something fundamentally new and useful to applications yielding ill-scaled instances.

The rest of this paper is organized as follows. In Sect. 2 we introduce random projections and the Johnson-Lindenstrauss lemma. In Sect. 3 we present our main theoretical result. In Sect. 4 we comment on the practical applicability of these techniques to LP. In Sect. 5 we introduce the quantile regression problem, and in Sect. 6 we discuss computational experiments.

2 Random projections

The term "random projections" is usually employed to describe a certain class of linear maps (commonly represented by matrices) which satisfy the claim of the celebrated Johnson-Lindenstrauss Lemma (JLL) [10], i.e.

> given n points in Euclidean space, [...] the smallest $k = k(n)$ so that these points can be moved into k-dimensional Euclidean space via a transformation which expands or contracts all pairwise distances by a factor of at most $1 + \epsilon$ [is] $k \leq C(\epsilon) \log n$ [for some universal constant C].

This is proved by choosing a random orthogonal projection on $\mathbb{R}^k$ and showing that with positive probability it satisfies the required condition. A more modern statement of the JLL is provided in [6].

2.1 Theorem ([6])
For any $0 < \epsilon < 1$ and any integer n, let k be $O(\frac{4}{\epsilon^2/2 - \epsilon^3/3} \log n)$. Then for any set X of n points in $\mathbb{R}^m$, there is a map $f : \mathbb{R}^m \to \mathbb{R}^k$ such that for all $x, y \in X$ we have

$$(1-\epsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1+\epsilon)\|x - y\|^2. \quad (3)$$

Furthermore, f can be found in randomized polynomial time.

The proof goes on to show that if $f(x) = T\,x$ (for $x \in X$), where T is a $k \times m$ matrix each component of which is sampled randomly from a normal distribution $\mathcal{N}(0, \frac{1}{\sqrt{k}})$ [8], then f satisfies Eq. (3) with probability at least $\frac{1}{n}$ (the probability can be increased arbitrarily by repeating the sampling sufficiently many times — in the set-up presented in [6], $O(n)$ times are prescribed).

Random projections have been used to design theoretically efficient algorithms for nearest neighbours [8, 9], to speed up clustering in high-dimensional spaces [3] as well as the efficiency of linear Support Vector Machines (SVM) [14]. It has also been used, interestingly, to perform quantile regression [4] in a way that is different from what we suggest in this paper, namely the problem is reduced to $\min \|Ax - b\|_1$ (as observed above, this eschews the difficulty of preserving the orthant $x \geq 0$). Our technique can be used to approximately solve *any* LP; we showcase its usefulness *in particular* to an LP formulation of the quantile regression problem.

We remark that the random projector T is a fully dense matrix, in general. This property can be relaxed by approximating the normal distribution by the discrete distribution which picks $+1$ or -1 with equal probabilities $\frac{\sqrt{s}}{2s}$ and 0 with probability $\sqrt{s}(1 - \frac{1}{s})$: we find $s = 3$ in [1] and $s = \sqrt{m}$ or $s = \frac{m}{\log m}$ in [13].

3 Approximately preserving LP optimality

For any LP $\mathscr{P}$ (with minimization direction in the objective function), we denote by $v(\mathscr{P})$ the optimal objective function value of $\mathscr{P}$. We denote infeasibility by $v(\mathscr{P}) = +\infty$ and unboundedness by $v(\mathscr{P}) = -\infty$. We denote the system of constraints of $\mathscr{P}$ by $\mathsf{con}(\mathscr{P})$ and its feasible set by $\mathcal{F}(\mathscr{P})$. If $\mathcal{F}(\mathscr{P}) \neq \varnothing$, we write $\mathsf{feas}(\mathscr{P})$. If the constraint matrix A in $\mathsf{con}(\mathscr{P})$ is $m \times n$, and T is a $k \times n$ matrix, we write $T\mathscr{P}$ to denote the LP obtained by pre-multiplying $\mathsf{con}(\mathscr{P})$ by T, namely $T\mathscr{P} \equiv \min\{c^\top x \mid TA\,x = Tb \wedge x \geq 0\}$.

In this section we are going to argue that, if $\mathcal{F}(\mathscr{P})$ is non-empty and bounded, and T is an appropriately sized random projector, we have $v(\mathscr{P}) \approx v(T\mathscr{P})$ *with overwhelming probability* (w.o.p.). The approximation relation $\approx$ will be defined in Thm. 3.3. By "w.o.p." we mean that the probability of the event referred to is $1 - \psi(k)$, where $\psi(k) \to 0$ as $k \to \infty$. Typically, $\psi(k)$ is $O(e^{-k})$. We encode the boundedness assumption of $\mathscr{P}$ as the statement that there exists a constant $\theta \geq 1$ and an optimal solution x^* of $\mathscr{P}$ such that $\sum\limits_{j \leq n} x^*_j \leq \theta$.

We first show that one can project just a subset of the constraints of an LP.

3.1 Lemma
Let $\mathscr{P}$ be an LP in standard form $\min\{c^\top x \mid Ax = b \wedge x \geq 0\}$, T be a $k \times n$ random projector, and $\mathscr{Q}$ be an LP such that $\mathsf{con}(\mathscr{Q})$ is the system

$$Ax = b \ \wedge \ A'x \leq b' \ \wedge \ x \geq 0,$$

where A is $m \times n$, A' is $m' \times n$ and T is $k \times n$. If $\mathsf{feas}(\mathscr{P}) \leftrightarrow \mathsf{feas}(T\mathscr{P})$ w.o.p., then there exists a matrix T' such that $\mathsf{feas}(\mathscr{Q}) \leftrightarrow \mathsf{feas}(T'\mathscr{Q})$.

Proof. It suffices to take $T' = \begin{pmatrix} T & 0 \\ 0 & I_{m'} \end{pmatrix}$. □ □

We remark that Lemma 3.1 *assumes* that random projections can preserve feasibility and optimality of an LP: if that is the case, then the lemma says that one can also just project a part of the constraints, rather than all of them. The objective of the rest of this section is to prove that that assumption holds, i.e. random projections *do* preserve LP feasibility and optimality w.o.p.

The fact that if $Ax = b \wedge x \geq 0$ is feasible, then $TAx = Tb \wedge x \geq 0$ is also feasible should be obvious, as T simply achieves a weighted aggregation of feasible constraints (there is no probability involved in this statement). The fact that if $Ax = b \wedge x \geq 0$ is infeasible, then the projected version is also infeasible is not true in general, but it was shown in [12] that it holds w.o.p., since random projections preserve separating hyperplanes and distances from a point to a cone w.o.p.

3.2 Theorem ([12])
Given an LP $\mathscr{P}$ in standard form (Eq. (1)) and an appropriately sized random projector T, the statement $\mathsf{feas}(\mathscr{P}) \leftrightarrow \mathsf{feas}(T\mathscr{P})$ *holds w.o.p.*

Now let $\mathscr{P}$ be the given LP, with bounded non-empty feasible set and a primal solution x^* such that $\sum_j x_j^* \leq \theta$ for some $\theta \geq 1$. We enforce this assumption by means of a single inequality constraint $\sum_j x_j \leq \theta$, so that $\mathscr{P}$ is the following problem:

$$P \equiv \min\{c^\top x \mid Ax = b \ \wedge \ \sum_{j\leq n} x_j \leq \theta \ \wedge \ x \geq 0\}.$$

By Lemma 3.1 and Thm. 3.2, we can extend the definition of $T\mathscr{P}$ to cover LPs other than in standard form as follows:

$$T\mathscr{P} \equiv \min\{c^\top x \mid TA\,x = Tb \ \wedge \ \sum_{j\leq n} x_j \leq \theta \ \wedge \ x \geq 0\}.$$

Our main theorem follows.

3.3 Theorem
Given $\delta > 0$, we have $v(\mathscr{P}) - \delta \leq v(T\mathscr{P}) \leq v(\mathscr{P})$ w.o.p.

We remark that most of the technical complexity behind the statement of Thm. 3.3 is hidden in the "w.o.p.". A more precise statement is that given $\delta > 0$ there is an ϵ, expressed as a function of δ, that ensures that $v(\mathscr{P}) - \delta \leq v(T\mathscr{P}) \leq v(\mathscr{P})$ with probability exceeding $1 - 4n\, e^{-C(\epsilon^2 - \epsilon^3)k}$, where k is as in Thm. 2.1 and C is a universal constant.

We only sketch the proof; our sketch hides all of the technical details behind the "w.o.p." statement, but maintains the proof structure. The boundedness constraint $\sum_j x_j \leq \theta$ is used to show that a separation argument based on that constraint has overwhelming probability of being projected (we already mentioned that random projections preserve separating hyperplanes [12]).

Proof sketch. For $v(T\mathscr{P}) \leq v(\mathscr{P})$ (the "easy" part of the proof), it suffices to remark that $TA\,x = Tb$ is a weighted aggregation of the constraints $Ax = b$, and hence any feasible solution of $\mathscr{P}$ is also feasible in $T\mathscr{P}$. Moreover, $\mathscr{P}$ and $T\mathscr{P}$ have the same objective function, so $T\mathscr{P}$ is actually a relaxation of $\mathscr{P}$, whence $v(T\mathscr{P}) \leq v(\mathscr{P})$. For the converse, we first reduce $\mathscr{P}$ to a Linear Feasibility Problem (LFP), in function of a given parameter $\delta \geq 0$, as follows:

$$\left.\begin{array}{rcl} c^\top x &=& v(\mathscr{P}) - \delta \\ Ax &=& b \\ x &\geq& 0. \end{array}\right\} \tag{4}$$

We note that, by definition of $v(\mathscr{P})$, Eq. (4) is infeasible for all $\delta > 0$. By Thm. 3.2 and Lemma 3.1,

$$\left.\begin{array}{rcl} c^\top x &=& v(\mathscr{P}) - \delta \\ TA\,x &=& Tb \\ x &\geq& 0. \end{array}\right\} \tag{5}$$

is infeasible w.o.p. for all $\delta > 0$. Hence

$$\left.\begin{array}{rcl} c^\top x &<& v(\mathscr{P}) - \delta \\ TA\,x &=& Tb \\ x &\geq& 0. \end{array}\right\} \tag{6}$$

is also infeasible w.o.p. This means that $c^\top x \ge v(\mathscr{P}) - \delta$ holds w.o.p. for any $x \in \mathcal{F}(T\mathscr{P})$, which proves that $v(\mathscr{P}) - \delta \le v(T\mathscr{P})$ w.o.p. □

4 Practical applicability

Many applications of random projections to computer science described in the literature are theoretical in nature (see e.g. [1, 9, 15]). The empirical study [18] was designed to test the direct applicability of the JLL (Sect. 2) in view of determining the value of certain parameters best suited to speed up the k-means clustering algorithm. The application to neuroscience in [2] is almost philosophical in nature, illustrating how the brain itself *might* conceivably be doing something similar to random projections in order to compress information. Although random projections are regularly used empirically [3, 14, 4], we found it somewhat surprising that such an apparently easy-to-use tool is not employed more often in practice.

Our experience provides evidence that random projections are only apparently easy to use, specially with respect to optimization. We found it extremely difficult to obtain good results, for both technical as well as theoretical reasons. In this section we discuss these difficulties and the way we tackled them.

4.1 The optimal value and the optimum

We remark that the main theorem of Sect. 3 only shows that random projections preserve optimality (w.o.p.) *in the sense of the optimal objective value*, rather then the actual optima. By running preliminary computational experiments, we found that the optima of $T\mathscr{P}$ were often infeasible w.r.t. the constraints of $\mathscr{P}$. We remark that Thm. 3.2 only states that $\mathsf{feas}(\mathscr{P}) \leftrightarrow \mathsf{feas}(T\mathscr{P})$ w.o.p.; while obviously every solution of $\mathscr{P}$ is feasible in $T\mathscr{P}$ (which just lists aggregated constraints of $\mathscr{P}$), nothing is said about the solutions of $T\mathscr{P}$. These experiments motivated us to prove the following result.

4.1 Proposition
Let $\mathscr{P}$ be a feasible LP in standard form (Eq. (1)) with non-empty relative interior, and T be an appropriately sized random projector. If $\mathsf{feas}(\mathscr{P})$ and x^* is uniformly sampled in $\mathcal{F}(T\mathscr{P})$ (which we assume to be equipped with a uniform probability measure μ), then the event $x^* \in \mathcal{F}(\mathscr{P})$ has probability zero.

Proof. Let $F = \mathcal{F}(\mathscr{P})$ and $TF = \mathcal{F}(T\mathscr{P})$. For each $v \in \mathsf{ker}(T)$ we let

$$F_v = \{x \ge 0 \mid Ax - b = v\} \cap TF \tag{7}$$

(note that $F_0 = F$). We aim to show that $\mathsf{Prob}(x^* \in F) = 0$, and proceed by contradiction: suppose $\mathsf{Prob}(x^* \in F) = p > 0$. We shall prove that there is a $\delta > 0$ and a family $\mathcal{V}$ of infinitely many $v \in \mathsf{ker}(T)$ such that

$$\mathsf{Prob}\left(x^* \in \bigcup_{v \in \mathcal{V}} F_v\right) \ge \sum_{v \in \mathcal{V}} \delta = +\infty,$$

leading to a contradiction. The case where random projectors are actually useful is when $k \ll m$, so we can assume $\mathsf{dim}(\mathsf{ker}(T)) \ge 1$. So $\mathsf{ker}(T)$ must contain at least one segment $[-u, u]$. Moreover, since the relative interior of F is non-empty, we can choose $\|u\|$ small enough so that $[-u, u]$ also belongs to the set $\{Ax - b \mid x \ge 0\}$. For the same reason, there is $\bar{x} > 0$ such that $A\bar{x} = b$. Let $\hat{x}$ be such that $A\hat{x} = -u$. Then, since $\bar{x} > 0$, there is a constant $M > 0$ large enough so that $2\hat{x} \le M\bar{x}$. For all $N \ge M$ and for all $x \in F$, we let $x'_N = \frac{\bar{x}+x}{2} - \frac{1}{N}\hat{x}$. This yields $Ax'_N = b - \frac{1}{N}A\bar{x} = b + \frac{u}{N}$ and $x'_N = \frac{x}{2} + (\frac{\bar{x}}{2} - \frac{\hat{x}}{N}) \ge 0$. We therefore obtain $\frac{\bar{x}+F}{2} - \frac{1}{N}\hat{x} \subseteq F_{u/N}$ (where $F_{u/N}$ is as in Eq. (7) with v replaced by $\frac{1}{N}u$), which implies that, for all $N \geq M$,

$$\mathsf{Prob}(x^* \in F_{u/N}) = \mu(F_{u/N}) \geq \mu\left(\frac{\bar{x}+F}{2}\right) \geq C\mu(F) = Cp > 0$$

for some constant $C > 0$, which proves the claim. □ □

One might argue that we are interested in extreme points of TF rather than a uniformly sampled point in TF, since this is what we obtain from solving $T\mathscr{P}$ in practice. We were able to prove that this does not help either.

4.2 Solution recovery

We were able to circumvent the setback discussed in Sect. 4.1 by considering the dual LP of $\mathscr{P}$ and $T\mathscr{P}$ (where T is an appropriately sized random projector):

$$\mathscr{D} \equiv \max\{b^\top y \mid A^\top y \leq c\} \tag{8}$$
$$T\mathscr{D} \equiv \max\{(Tb)^\top y \mid (TA)^\top y \leq c\}. \tag{9}$$

Most existing LP solvers will output both primal and dual solution pairs (x^*, y^*) to any given LP in standard form Eq. (1). If we solve $T\mathscr{P}$ and get a primal/dual solution pair (x', y'), however, we know by Prop. 4.1 that x' is almost never a solution of $\mathscr{P}$.

4.2 Proposition
$y^* = T^\top y'$ is a feasible solution of $\mathscr{D}$.

Proof. The proof is simply $(TA)^\top y' = (A^\top T^\top)y' = A^\top(T^\top y') = A^\top y^* \leq c$, from which it follows, by Eq. (9), that y' is a solution of $T\mathscr{D}$. □ □

It is somewhat more involved to find a dual optimal solution of $\mathscr{D}$. By Thm. 3.3, we know we can find the optimal objective function value $c^* = v(\mathscr{P})$ w.o.p. by simply solving $T\mathscr{P}$. We consider the following auxiliary LP $\mathscr{Q}$, defined for any uniformly chosen random vector $\alpha \in (0,1)^n$:

$$\mathscr{Q} \equiv \min\{\alpha^\top x \mid c^\top x = c^* \wedge Ax = b \wedge x \geq 0\}, \tag{10}$$

with associated projected LP $T\mathscr{Q}$ defined by replacing $Ax = b$ by $TAx = Tb$ (as per Lemma 3.1 applied to a pair of inequalities $c^\top x \leq c^*$ and $c^\top x \geq c^*$). By Prop. 4.2 we obtain a dual solution y' to $T\mathscr{Q}$ whose primal x' satisfies $cx' = c^*$ and is therefore optimal for $T\mathscr{P}$.

4.3 Lemma
Let y' be an optimal dual solution for $T\mathscr{Q}$, where T is a $k \times m$ random projector. Then almost surely $y^* = A^\top y'$ satisfies exactly k constraints of the system $A^\top y \leq \alpha$ at equality.

Proof. Since y' is dual optimal in $T\mathscr{Q}$ it is a basic solution, so it satisfies at least k constraints of $(TA)^\top y \leq c$ at equality, and hence by the proof of Prop. 4.2 this also holds for y^* w.r.t. $A^\top y \leq c$. Let I be the set of indices of rows of $A^\top$ corresponding to equalities: then we have $A_I^\top y^* = \alpha_I$, i.e. $A_I^\top T^\top y^* = \alpha_I$, where $M = A_I^\top T^\top$ is an square invertible $k \times k$ matrix and $y^* = M^{-1}\alpha_I$. Suppose there were more than k constraints satisfied at equality by y^*. Then there would be an index $j \notin I$ such that $A_j^\top M^{-1}\alpha_I = \alpha_j$. But since α is sampled uniformly at random, this happens with probability 0. □ □

We can now use Lemma 4.3 to reconstruct a primal solution of $\mathscr{Q}$, which is by definition also optimal for $\mathscr{P}$: we consider the set J of column indices of A (i.e. row indices of $A^\top$) such that $A_j^\top y < \alpha$. By the

LP complementarity conditions applied to $\mathscr{D}$, $(A_j^\top y - \alpha)x_j = 0$, which implies that $x_j = 0$ for all $j \in J$. We let $\bar{J}$ be the complement of J w.r.t. $\{1,\ldots,n\}$, and consider the subsystem $A_{\bar{J}}x_{\bar{J}} = b$ obtained by removing the columns of A indexed by J. This is an $m \times k$ linear system having rank k. It suffices to pre-multiply $A_{\bar{J}}$ by any $k \times m$ full rank matrix B to obtain the square invertible system $BA_{\bar{J}}x_{\bar{J}} = Bb$, which yields a partial solution $x^*_{\bar{J}}$. Finally, re-inserting the zero components indexed by J yields a primal solution x^* of $\mathscr{P}$.

5 The quantile regression LP

Quantile regression is a form of regression analysis. Linear regression indicates a regression towards the mean, but other types of regressions are possible, e.g. towards the median (which is also the $\frac{1}{2}$-quantile), or towards other arbitrary τ-quantiles, for any $\tau \in [0,1]$. Rather than properly defining quantile regression in a forum for optimizers, we give an introduction "by analogy with linear regression", hoping that everyone is familiar with the latter.

Suppose we have a random variable $\mathcal{B}$, and that we assume that $\mathcal{B}$ depends linearly on a vector $\mathcal{A}$ of independent random variables $\mathcal{A}_1,\ldots,\mathcal{A}_p$, so that $\mathcal{B} \sim \sum_{j\le p} x_j\mathcal{A}_j$. We do not know the coefficient vector $x = (x_1,\ldots,x_p)$, but we have a database table A consisting of p columns $A^1,\ldots,A^p \in \mathbb{R}^q$, the j-th of which is a sample for $\mathcal{A}_j$. We also have a sample b for $\mathcal{B}$. We represent the database table by a $q \times p$ matrix A, its i-th row by A_i and its j-th column by A^j.

Following [19], the *sample mean* for $\mathcal{B}$ can be defined as the optimum of the single-variable optimization problem $\arg\min\limits_{\mu\in\mathbb{R}} \sum\limits_{i\le q}(b_i-\mu)^2$. The sample mean conditional to $\mathcal{B}$ being a linear form in $\mathcal{A}_1,\ldots,\mathcal{A}_p$ can be defined by generalizing the previous problem as $\arg\min\limits_{\nu\in\mathbb{R}^p} \sum\limits_{i\le q}(b_i-\nu A_i)^2$. The loss function related to the sample mean is $\ell(u) = u^2$.

The *sample median* for for $\mathcal{B}$ can be defined analogously, but with a loss function $\ell(u) = \frac{1}{2}|u|$ [19]. It is the optimum of $\arg\min\limits_{\xi\in\mathbb{R}} \frac{1}{2}\sum\limits_{i\le q}|b_i-\xi|$, which can also be written as

$$\arg\min_{\xi\in\mathbb{R}} \sum_{i\le q}\left(\frac{1}{2}\max(b_i-\xi,0) - \frac{1}{2}\min(b_i-\xi,0)\right).$$

Finally, the τ-quantile sample conditional to $\mathcal{B}$ being a linear form in function of the multivariate random variable $\mathcal{A}$ is:

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i\le q}\left(\tau\max(b_i-\beta A_i,0) - (1-\tau)\min(b_i-\beta A_i,0)\right). \tag{11}$$

It is well known (and not too surprising) that Eq. (11) can be reformulated exactly to an LP in standard form. We replace the positive part $\max(b_i-\beta A_i,0)$ by an additional variable u_i^+ and the negative part $\min(b_i-\beta A_i,0)$ by $-u_i^-$ (for all $i\le q$), and obtain $\min\{\tau\mathbf{1}u^+(1-\tau)\mathbf{1}u^- \mid A\beta+u^+-u^-=b \wedge u\ge 0\}$. This can be simply written as the following LP:

$$\left.\begin{array}{rrcl} \min\limits_{\beta,u^+,u^-} & \tau\mathbf{1}^\top u^+ + (1-\tau)\mathbf{1}^\top u^- & & \\ & A\beta+u^+-u^- & = & b \\ & u^+,u^- & \ge & 0. \end{array}\right\} \text{ [QR]} \tag{12}$$

We note that the constraint matrix of Problem [QR] (Eq. (12)) can be written as the $m\times n$ matrix $B = (A\,|\,I_m\,|\,-I_m)$, where $m=q$ and $n=p+2m$. If A is dense, then we can expect B to be reasonably dense. Moreover, since A is a (numeric) database, it can be as large as desired. This turns out to be a good testbed for our technique, as we shall see.

We also remark that, although the LP in Eq. (12) is not in standard form, unrestricted variables such as β are easy to deal with, as $\|Ax-b\|_2^2$ is projected to $\|TAx-Tb\|_2^2$ w.o.p. directly by Thm. 2.1 whenever x is unrestricted. As already mentioned, the reason why random projections do not apply to LPs directly, and why our results are important, is that the projection w.o.p. of a given orthant is non-trivial to prove.

6 Computational experiments

We report results on two different sets of computational experiments. The first consists of randomly generated instances where we are able to compute the quantile exactly. These are used to check that the projected LP actually gives an approximate solution that is close to the correct one. This empirical verification is necessary since random projection results are asymptotic, and may fail for "small" instances. The second consists of two difficult instances, on which we were unable to compute the quantile precisely.

All our results have been obtained by running the IBM ILOG CPLEX 12.6.2 [7] on a twin-core 3.1GHz Intel Core i7 CPU virtually appearing as a quad-core to the operating system (OSX 10.11.6), configured with 16GB RAM. All non-CPLEX related computation was carried out using Python 2.7 (and hence slower than it could be, since Python is not a compiled language). We note moreover that much faster implementations exist for matrix multiplications than the default ones provided by Python; in this sense, the time gain in using our techniques could be even larger.

In all experiments we used the CPLEX *barrier* algorithm. In both test sets, projected LPs have been solved with the *crossover* option active, which guarantees the solution to be exact.

We used an evaluation of the universal constant C involved in the asymptotic term $O(\log n)$ for the projected dimension k (see Thm. 2.1) of $C = 1.8$ (derived from a mixture of experimentation and advice found in [18]). We also pursued a couple of other (minor) software engineering tuning actions, which would take too long to explain here. We remark that in each case, we only ever sampled the random projector T once.

To obtain computational results, we proceed as follows. We first solve the original LP $\mathscr{P}$ (11) to optimality to obtain the primal/dual solution pair (x^*, y^*) and $f^* = v(P)$. We then solve the projected LP $T\mathscr{P}$ to optimality to obtain the solution pair $(\bar{x}, \bar{y})$ and the optimal value $\bar{f}$. By Prop. 4.1 we ignore $\bar{x}$, and use the results of Sect. 4.2 to recover an approximate primal solution x' from the projected dual $\bar{y}$. We note that $x^* = (\beta^*, u^*)$ and $x' = (\beta', u')$. Our main quality measure compares β^* to β' (both vectors in $\mathbb{R}^p$) to extract the *mean component error* (MCE):

$$\mathsf{MCE}(\beta^*, \beta') = \frac{1}{p}\|\beta^* - \beta'\|_2. \tag{13}$$

We would also like to ascertain whether the projected solution is within the JLL multiplicative error factor ϵ (see Thm. 2.1). Since this error is related to the columns of A by $\epsilon \geq \left|1 - \frac{\|TA_i - TA_j\|_2}{\|A_i - A_j\|_2}\right|$, we consider the dual solutions y^* and $\bar{y}^*$, and compute the *mean dual projection error* (MDPE):

$$\mathsf{MDPE}(y^*, \bar{y}) = \left|1 - \frac{1}{n}\sum_{j\leq n}\frac{\|\bar{y} - TA_j\|_2}{\|y^* - A_j\|_2}\right|. \tag{14}$$

6.1 Random instances

We generated random $q \times p$ data matrices A and corresponding random columns $b \in \mathbb{R}^q$, as follows: every component of A is sampled uniformly at random from $[0, 1]$, and $b = A\beta + \frac{1}{p+1}\mathcal{N}(0, 1)$ where each component of $\beta \in \mathbb{R}^q$ is sampled uniformly at random from $[0, 1]$. Eight instances were generated for $q \in \{2000, 3000\}$ and $p \in \{100, 500, 1000, 1500\}$, and computed MCE and MDPE for every quantile $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Only aggregated results (over all 40 instances) are presented in Table 1. The

	MCE	MDPE	f^*	$\bar{f}$	CPU	T CPU
Average	2.51E-05	0.09	2.18	1.31	104.62	73.57
Standard deviation	3.75E-05	0.07	3.14	2.04	121.69	38.93

Table 1: Aggregated results on random instances. We also report comparative CPU times (T CPU are the wall-clock seconds of CPU taken to solve $T\mathscr{P}$).

MCE being $O(10^{-5})$ and p being $O(10^3)$ yields an absolute error in the order $O(10^{-2})$, which must be related with solution vectors β with components in $O(10^{-1})$: this yields an approximation ratio $O(\epsilon)$ (with $\epsilon = 0.1$). The MDPE is also $O(\epsilon)$. By contrast, $\bar{f}/f^* \approx 0.6$.

6.2 Two realistic instances

We now discuss two realistic instances where quantile regression fails using established tools such as CPLEX on the LP (11), or even the statistical software R [16]. The first (`hh1995f`) is the "household table" from the Russia Longitudinal Monitoring Survey 1995 `www.cpc.unc.edu/projects/rlms-hse`, with $q = 3783$, $p = 855$, 0.185 density and very poorly scaled. The second (`my_photos`) is a set of 14596 RGB graphic files scaled to 90×90 pixels: it has $q = 14596$, $p = 24300$, 0.624 density and is well scaled. We set ϵ (Thm. 2.1) to 0.2 for both instances. CPLEX failed by returning (an impossible) "infeasible" on `hh1995f`, and by resource exhaustion on `my_photos`. This prevents us from computing MCE and MDPE measures. We only compute a mean relative feasibility error $\phi(x') = \frac{\|Ax'-b\|/q}{\mathsf{avg}(|b|)}$ of x' with respect to the equality constraints of $\mathscr{P}$, shown in Table 2 together with the $\bar{f} = v(T\mathscr{P})$ and the CPU time taken to solve it.

	ϕ	$\bar{f}$	T CPU
`hh1995f`	0.0005	0.0	3.88
`my_photos`	0.0002	0.0	168.67

Table 2: Projected $\frac{1}{4}$-quantile regression on hard instances.

References

[1] Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. Journal of Computer and System Sciences 66, 671–687 (2003)

[2] Allen-Zhu, Z., Gelashvili, R., Micali, S., Shavit, N.: Sparse sign-consistent Johnson-Lindenstrauss matrices: Compression with neuroscience-based constraints. Proceedings of the National Academy of Sciences 111(47), 1687216876 (2014)

[3] Boutsidis, C., Zouzias, A., Drineas, P.: Random projections for k-means clustering. In: Advances in Neural Information Processing Systems. pp. 298–306. Neural Information Processing Systems Foundation, La Jolla (2010)

[4] Clarkson, K., Drineas, P., Magon-Ismail, M., Mahoney, M., Meng, X., Woodruff, D.: The fast Cauchy transform and faster robust linear regression. SIAM Journal on Computing 45(3), 763–810 (2016)

[5] Dantzig, G.: Linear Programming and Extensions. Princeton University Press, Princeton, NJ (1963)

[6] Dasgupta, S., Gupta, A.: An elementary proof of a theorem by johnson and lindenstrauss. Random Structures and Algorithms 22, 60–65 (2002)

[7] IBM: ILOG CPLEX 12.6 User's Manual. IBM (2014)

[8] Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Symposium on the Theory Of Computing. STOC, vol. 30, pp. 604–613. ACM, New York (1998)

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

[10] Johnson, W., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Hedlund, G. (ed.) Conference in Modern Analysis and Probability. Contemporary Mathematics, vol. 26, pp. 189–206. American Mathematical Society, Providence (1984)

[11] Koenker, R.: Quantile regression. Cambridge University Press, Cambridge (2005)

[12] Ky, V.K., Poirion, P.L., Liberti, L.: Using the Johnson-Lindenstrauss lemma in linear and integer programming. Tech. Rep. 1507.00990v1[math.OC], arXiv (2015)

[13] Li, P., Hastie, T., Church, K.: Very sparse random projections. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining. KDD, vol. 12, pp. 287–296. ACM, New York (2006)

[14] Paul, S., Boutsidis, C., Magdon-Ismail, M., Drineas, P.: Random projections for linear support vector machines. ACM Transactions on Knowledge Discovery from Data 8(4), 22:1–22:25 (2014)

[15] Pilanci, M., Wainwright, M.: Randomized sketches of convex programs with sharp guarantees. In: International Symposium on Information Theory (ISIT). pp. 921–925. IEEE, Piscataway (2014)

[16] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015), `https://www.R-project.org/`

[17] Stigler, G.: The cost of subsistence. Journal of Farm Economics 27(2), 303–314 (1945)

[18] Venkatasubramanian, S., Wang, Q.: The Johnson-Lindenstrauss transform: An empirical study. In: Algorithm Engineering and Experiments. ALENEX, vol. 13, pp. 164–173. SIAM, Providence (2011)

[19] Yu, K., Lu, Z., Stander, J.: Quantile regression: applications and current research areas. The Statistician 52(3), 331–350 (2003)