Today, we analyze an application of random projection to compute approximate solutions of constrained least-squares problems. This method is often referred to as sketched least-squares.

### 3.1 Problem set up and motivation

Suppose that we are given an observation vector $y \in \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n \times d}$, and that for some convex set $C \subset \mathbb{R}^d$, we would like to compute the constrained least-squares solution

$$x_{LS} := \arg\min_{x \in C} \frac{1}{2} \|y - Ax\|_2^2.$$

(3.1)

In general, this solution may not be unique, but we assume throughout this lecture that uniqueness holds (so that $n \geq d$ necessarily).

Different versions of the constrained least-squares problem arise in many applications:

- In the simplest case of an unconstrained problem ($C = \mathbb{R}^d$), it corresponds to the usual least-squares estimator, which has been widely studied. Most past work on sketching least-squares has focused on this case.
- When $C$ is a scaled form of the $\ell_1$-ball—that is, $C = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq R\}$—then the constrained problem is known as the Lasso. It is widely used for estimating sparse regression vectors.
- The support vector machine for classification, when solved in its dual form, leads to a least-squares problem over a polytope $C$.

Problems of the form (3.1) can also arise as intermediate steps of using Newton’s method to solve a constrained optimization problem.

The original problem can be difficult to solve if the first matrix dimension $n$ is too large. Thus, in order to reduce both storage and computation requirements, a natural idea is to randomly project the original data to a lower-dimensional space. In particular, given a random sketch matrix $S \in \mathbb{R}^{m \times n}$, consider the sketched least-squares problem

$$\hat{x} := \arg\min_{x \in C} \frac{1}{2} \|S(y - Ax)\|_2^2.$$

(3.2)
The row dimensions of the sketched problem are now $m$-dimensional as opposed to $n$-dimensional, which will lead to savings whenever we can choose $m \ll n$. This idea (in the unconstrained case $C = \mathbb{R}^d$) was first proposed and analyzed by Sarlos [9]. The more general analysis to be described here is due to Pilanci and Wainwright [8].

There are different ways to measure the quality of the sketched solution. For a given tolerance parameter $\delta \in (0, 1)$, we say that it is $\delta$-accurate cost approximation if
\[
f(x_{LS}) \leq f(\hat{x}) \leq (1 + \delta)^2 f(x_{LS}).
\] (3.3)

On the other hand, we say that it is a $\delta$-accurate solution approximation if
\[
\left\| \frac{1}{\sqrt{n}} A(x_{LS} - \hat{x}) \right\|_2 \leq \delta.
\] (3.4)

In this lecture, we will focus on the cost notion of performance (3.3).

In particular, our main goal is to give a precise answer to the following question: how large must the projection dimension $m$ be to guarantee that the sketched solution $\hat{x}$ provides a $\delta$-accurate cost approximation?

![Figure 3.1: Illustration of the tangent cone at $x_{LS}$](image)

The answer depends on a natural geometric object associated with the convex program (3.1), namely the tangent cone of $C$ at $x_{LS}$. In particular, it is the set
\[
\mathcal{K}(x_{LS}) = \left\{ \Delta \in \mathbb{R}^d \mid \Delta = t(x - x_{LS}), \ x \in C, t \geq 0 \right\}.
\] (3.5)

Our theory shows that the projection dimension $m$ required depends on the “size” of (a transformed version of) this tangent cone. Note that the structure of this $\mathcal{K}(x_{LS})$ can vary drastically as a function of $x_{LS}$. Figure 3.1(a) shows a favorable case, where $x_{LS}$ lies at a vertex of a polyhedral constraint set; panel (b) shows a less favorable case, where $x_{LS}$ lies on a higher-dimensional face. The worst case for our theory is when $x_{LS}$ is an interior point of $C$, so that $\mathcal{K}(x_{LS}) = \mathbb{R}^d$. Note that this worst case always occurs for an unconstrained problem.
3.2 Projection theorem for constrained least square problem

Consider the set
\[
AK(x_{LS}) \cap S^{n-1} = \{ v \in \mathbb{R}^n \mid \exists \Delta \in K(x_{LS}), v = A\Delta, \|v\|_2 = 1 \},
\]
(3.6)
corresponding the intersection of the transformed tangent cone \(AK(x_{LS})\) with the Euclidean sphere. We define its Gaussian width in the usual way
\[
\mathcal{W}(AK(x_{LS}) \cap S^{n-1}) := \mathbb{E}\left[ \sup_{u \in AK(x_{LS}) \cap S^{n-1}} (u, g) \right] \quad \text{where } g \sim N(0, I_{n \times n}).
\]
(3.7)

**Theorem 3.2.1.** Consider the sketched least-squares solution (3.2) based on a Gaussian random matrix \(S \in \mathbb{R}^{m \times n}\). Given a projection dimension \(m \geq \frac{C}{\delta^2} \mathcal{W}^2(AK(x_{LS}) \cap S^{n-1})\), it is a \(\delta\)-accurate cost approximation with probability at least \(1 - c_0 \exp\{-c_1 m \delta^2\}\).

Here \((C, c_0, c_1)\) are universal constants independent of all problem parameters.

**Proof.** In order to simplify notation, we omit the dependence of \(K(x_{LS})\) on \(x_{LS}\), and write \(K\) instead. We define the following two quantities:
\[
Z_1(AK) = \inf_{v \in AK(x_{LS}) \cap S^{n-1}} \frac{\|Sv\|^2}{m}
\]
(3.8a)
and
\[
Z_2(AK, u) = \sup_{v \in AK(x_{LS}) \cap S^{n-1}} \left| u^T \left( \frac{S^T S}{m} - I \right) v \right|,
\]
where \(u \in S^{n-1}\) is some fixed vector. (3.8b)

For a random sketch, note that \(Z_1\) and \(Z_2\) are also random variables, but parts of our theory hold deterministically for any choice of \(S\). The following lemma shows that the ratio \(Z_2/Z_1\) controls the quality of \(\hat{x}\) as a cost approximation for \(f(x) = \frac{1}{2} \|Ax - y\|_2^2\).

**Lemma 3.2.1.** For any sketch matrix \(S \in \mathbb{R}^{m \times n}\), we have
\[
f(\hat{x}) \leq \left(1 + \frac{Z_2(AK, u_{LS})}{Z_1(AK)} \right)^2 f(x_{LS}) \quad \text{where } u_{LS} := \frac{y - Ax_{LS}}{\|y - Ax_{LS}\|_2}.
\]
(3.9)

**Proof.** Define the error vector \(\hat{e} = \hat{x} - x_{LS}\). By the triangle inequality, we have
\[
\|A\hat{x} - y\|_2 \leq \|Ax_{LS} - y\|_2 + \|A\hat{e}\|_2 = \|Ax_{LS} - y\|_2 \left(1 + \frac{\|A\hat{e}\|_2}{\|Ax_{LS} - y\|_2}\right).
\]
By definition, we have \(\|A\hat{x} - y\|_2 = \sqrt{2f(\hat{x})}\) and \(\|Ax_{LS} - y\|_2 = \sqrt{2f(x_{LS})}\), and hence
\[
f(\hat{x}) \leq f(x_{LS}) \left(1 + \frac{\|A\hat{e}\|_2}{\|Ax_{LS} - y\|_2}\right)^2.
\]
(3.10)
Consequently, in order to prove the lemma, it suffices to upper bound the ratio \(\frac{\|A\hat{e}\|_2}{\|Ax_{LS} - y\|_2}\) by \(Z_2/Z_1\).
Note that \( \hat{x} \) is optimal and \( x_{LS} \) is feasible for the sketched problem (3.2). Thus, the first-order conditions for optimality guarantee that
\[
\langle \nabla g(\hat{x}), -\hat{e} \rangle \geq 0 \implies \langle S(A\hat{x} - y), -SA\hat{e} \rangle \geq 0
\]
\[
\implies \frac{1}{m} \|SA\hat{e}\|_2^2 \leq \frac{1}{m} \langle S(Ax_{LS} - y), -SA\hat{e} \rangle = (Ax_{LS} - y)^T \left( -\frac{S^TS}{m} + I \right) \hat{A} - (Ax_{LS} - y, A\hat{e}). \tag{3.11}
\]
On the other hand, by the optimality of \( x_{LS} \) and feasibility of \( \hat{x} \) for the original problem (3.1), we have
\[
\langle \nabla f(x_{LS}), \hat{e} \rangle \geq 0 \implies \langle Ax_{LS} - y, A\hat{e} \rangle \geq 0 \tag{3.12}
\]
Combining inequalities (3.11) and (3.12) yields
\[
\frac{1}{m} \|SA\hat{e}\|_2^2 \leq (Ax_{LS} - y)^T \left( -\frac{S^TS}{m} + I \right) \hat{A} - (Ax_{LS} - y, A\hat{e}) \leq 0
\]
Dividing through by Euclidean norms to normalize then yields
\[
\|A\hat{e}\|_2 \frac{\|SA\hat{e}\|_2^2}{\|Ax_{LS} - y\|_2^2} \leq \frac{(Ax_{LS} - y)^T}{\|Ax_{LS} - y\|_2} \left( \frac{S^TS}{m} - I \right) \frac{A\hat{e}}{\|A\hat{e}\|_2} \|Ax_{LS} - y\|_2 \|A\hat{e}\|_2.
\]
Recalling the definitions of \( Z_1 \) and \( Z_2 \), we have
\[
\|SA\hat{e}\|_2^2 \frac{\|Ax_{LS} - y\|_2}{\|A\hat{e}\||_2^2} \geq Z_1 \|A\hat{e}\|_2^2, \quad \text{and}
\]
\[
\frac{(Ax_{LS} - y)^T}{\|Ax_{LS} - y\|_2} \left( \frac{S^TS}{m} - I \right) \frac{A\hat{e}}{\|A\hat{e}\|_2} \|Ax_{LS} - y\|_2 \|A\hat{e}\|_2 \leq Z_2 \|Ax_{LS} - y\|_2 \|A\hat{e}\|_2.
\]
Putting together the pieces yields the desired bound
\[
\frac{\|A\hat{e}\|_2}{\|Ax_{LS} - y\|_2} \leq \frac{Z_2(\mathcal{A},u_{LS})}{Z_1(\mathcal{A})}.
\]

For a given \( \delta \in (0, 1/2) \), define the event \( \mathcal{E}((\delta)) := \left\{ Z_1 \geq 1 - \delta, Z_2 \leq \delta/2 \right\} \). Conditioned on this event, we have
\[
\frac{f(\hat{x})}{f(x_{LS})} \leq \left( 1 + \frac{Z_2(\mathcal{A},u_{LS})}{Z_1(\mathcal{A})} \right)^2 \leq \left( 1 + \frac{\delta^2/2}{1-\delta} \right)^2 \leq (1 + \delta)^2,
\]
which is our desired form of \( \delta \)-accurate cost approximation. Consequently, in order to complete the proof, we need to state a lower bound on the projection dimension that ensures that \( \mathbb{P}[\mathcal{E}(\delta)] \geq 1 - c_1 e^{-c_2 m \delta^2} \).

**Lemma 3.2.2.** For a Gaussian sketch matrix, there is a universal constant \( C \) such that \( m \geq C e^{(W^2(\mathcal{A}) \cap S^{n-1})} \) implies that \( \mathbb{P}[\mathcal{E}(\delta)] \geq 1 - c_1 e^{-c_2 m \delta^2} \).

We only sketch the proof of this lemma here, since it follows essentially from our main result in Lecture #2. Define the random variable
\[
Z(\mathcal{A}) = \sup_{v \in \mathcal{A} \cap S^{n-1}} \left| v^T \left( \frac{S^TS}{m} - I_n \right) v \right|. \tag{3.13}
\]
From Lecture #2, given our stated lower bound on the projection dimension, we have $Z(AK) \leq \delta/C'$ with probability at least $1 - c_1 e^{-c_2 m \delta^2}$. Here the constant $C' > 0$ can be made as large we please by choosing the constant $C$ in Lemma 3.2.2 larger.

Thus, let us suppose that the bound $Z(AK) \leq \delta/C'$ holds. As long as $C' > 1$, we are guaranteed that $Z_1 \geq 1 - \delta$. On the other hand, showing that $Z_2 \leq \delta/2$ requires a little more work, involving splitting into two cases, depending on whether the inner product $\langle u_{LS}, v \rangle$ is positive or negative. We refer to the paper [8] for details.

\[ \square \]

3.3 Some illustrative examples

Let’s consider some examples to illustrate.

**Unconstrained least squares:** When $K = \mathbb{R}^d$, then the transformed tangent cone $AK$ is the range space of matrix $A$. Using our calculations of Gaussian widths, we see that a projection dimension $m \gtrsim \text{rank}(A)/\delta^2$ is sufficient to give a $\delta$-accurate cost approximation. This result (modulo some superfluous logarithmic factors) was first shown by Sarlos [9].

**Noiseless compressed sensing:** Suppose that our goal is to recover a sparse vector $x^* \in \mathbb{R}^d$ based on observing the random projection $z = Szx^*$. We can recover known bounds for exact recovery as a special case of our general result, one with $A = I_d$ (and hence $n = d$).

Consider the $\ell_1$-constrained least squares problem $\min_{\|x\|_1 \leq R} \frac{1}{2} \|x - x^*\|_2^2$, say with the radius chosen as $R = \|x^*\|_1$. By construction, we have $x_{LS} = x^*$ for this problem. The sketched analog is given by

\[
\min_{\|x\|_1 \leq R} \frac{1}{2} \|S(x^* - x)\|_2^2 = \min_{\|x\|_1 \leq R} \frac{1}{2} \|z - Sx\|_2^2,
\]

where $z = Szx^*$ is observed. (3.14)

Suppose that $\|x^*\|_0 = k$, meaning that $x^*$ is a $k$-sparse vector. It turns out that (with high probability) the sketched problem (3.14) has $\hat{x} = x^*$ has its unique solution whenever $m \gtrsim k \log(ed/k)$. This is a standard result in compressed sensing [5, 4], and can be derived as a corollary of our general theorem.

We begin by noting that if $x_{LS}$ provides a $\delta$-accurate cost approximation, then we are guaranteed that

\[
\|x^* - \hat{x}\|_2^2 \leq (1 + \delta)^2 \|x^* - x_{LS}\|_2^2 = 0,
\]

which implies that $\hat{x} = x^*$. So for this special noiseless problem, a cost approximation of any quality automatically yields exact recovery.

How large a projection dimension $m$ is needed for a cost approximation (and hence exact recovery)? We need to bound the Gaussian width $\mathbb{W}(AK \cap S^{n-1})$. The following lemma provides an outer bound on this set. We use the notation $\mathbb{B}_p(r) = \{ x \in \mathbb{R}^d \mid \|x\|_p \leq r \}$ for the $\ell_p$-norm ball of radius $r$, and clconv to denote the closed convex hull.

**Lemma 3.3.1.** For any vector $x^*$ that is $k$-sparse, we have

\[
K(x^*) \cap S^{n-1} \overset{(a)}{\subseteq} \mathbb{B}_1(2\sqrt{k}) \cap \mathbb{B}_2(1) \overset{(b)}{\subseteq} \text{clconv} \left\{ \mathbb{B}_0(4k) \cap \mathbb{B}_2(1) \right\}.
\] (3.15)
Let us use this lemma to bound the Gaussian width. From inclusion (a) in equation (3.15), we have
\[
\mathcal{W}(K(x^*) \cap S^{d-1}) \leq \mathbb{E}\left[ \sup_{\|x\|_1 \leq 2\sqrt{k}} \langle x, g \rangle \right]
\]
\[
\leq 2\sqrt{k} \mathbb{E}\|g\|_\infty \quad \text{a very crude bound!}
\]
\[
(i) \leq 2\sqrt{2k \log d}
\]
where inequality (i) follows from Holder’s inequality; and step (ii) follows from our bound on the expected maximum of \(d\) independent Gaussian (shown in Lecture #2). This bound shows that it suffices to take \(m \gtrsim k \log d\).

Now let’s use inclusion (b) in equation (3.15), along with a more careful argument, to obtain a sharper bound. We have
\[
\mathcal{W}(K(x^*) \cap S^{d-1}) \leq 2\mathbb{E}\left[ \max_{|S|=4k} \sup_{\|x\|_2 \leq 1} \langle x_S, g_S \rangle \right] \leq 2\mathbb{E}\left[ \max_{|S|=4k} \|g_S\|_2 \right] \quad \text{where } g \sim N(0, I_d).
\]
Each variable \(\|g_S\|_2\) has mean at most \(\sqrt{|S|} = \sqrt{4k}\), and is concentrated around its mean with sub-Gaussian tails (Lecture #2, concentration of Lipschitz functions). Consequently, we have
\[
\mathbb{E}\left[ \max_{|S|=4k} \sup_{\|x\|_2 \leq 1} \langle x_S, g_S \rangle \right] \leq \sqrt{4k} + c \sqrt{\log \frac{d}{4k}} = c' \sqrt{k \log \frac{ed}{k}}.
\]
This calculation shows that it is sufficient to take \(m \gtrsim k \log(\frac{ed}{k})\).

It remains to prove Lemma 3.3.1.

**Proof.** We prove the first inclusion here. For any \(x \in C\), the difference vector \(\Delta = x - x^*\) (and scaled copies of it) belong to the tangent cone \(K(x^*)\). Let us write the support set of \(x^*\) as \(S\) and zero part \(S^c\). Since \(x \in C\), we have
\[
\|x\|_1 = \|x^* + \Delta\|_1 = \|x_S^* + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq \|x_S^*\|_1.
\]
By triangle inequality,
\[
\|x_S^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq \|x_S^* + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq \|x_S^*\|_1.
\]
Combining the two bounds yields \(\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\), and hence
\[
\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 2\|\Delta_S\|_1 \leq 2\sqrt{k} \|\Delta\|_2,
\]
as claimed.

For the proof of inclusion (b) in equation (3.15), we refer the reader to Loh and Wainwright [7].

### 3.4 Other random projection methods

Up until now, we have showed that through Gaussian projection we can reduce the problem dimension to its intrinsic dimension which characterized by a geometric object: Gaussian Width of the tangent cone at
\[ x_{\text{LS}} \]. If we only care about the data storage issue, this projection does help us. However, from a complexity of computing point of view, the cost of doing Gaussian sketching to get \( Sy, SA \) where \( A \in \mathbb{R}^{n \times d} \), is of order \( \mathcal{O}(mnd) \). The computing complexity of doing the sketching can be even larger than solving the original least square problem directly. Then why bother using the sketching technique?

The key is that our theory holds for many other types of random sketches apart from Gaussian. In fact, our theory gives bounds for any matrix \( S \) for which we can control the quantity

\[
Z(A\mathcal{K}(x_{\text{LS}}) \cap S^{n-1}) = \sup_{v \in A\mathcal{K}(x_{\text{LS}}) \cap S^{n-1}} \left| \frac{1}{m} \| Sv \|_2^2 - 1 \right|.
\]

What type of matrices can be used to speed up the projection step?

- **Fast JL transforms**: These methods are based on structured forms of random matrices for which the projection can be computed quickly [1, 2].

- **Sparse JL transforms**: These methods are based on constructing projection matrices \( S \) with a very large number of zero entries, so that the matrix multiplication can be sped up by using routines for sparse matrices [6, 3].

One class of fast JL transforms based on structured matrices take the following form:

1. Start with an orthonormal matrix \( H \in \mathbb{R}^{n \times n} \) for which it is possible to perform fast matrix multiplication. Examples include the discrete Fourier matrix (multiplies carried out by fast Fourier transform), or the Hadamard matrix (multiplies carried out with the fast Hadamard transform).

2. Form the projection matrix \( S = H_m D \) where
   - the matrix \( H_m \in \mathbb{R}^{m \times n} \) is obtained by sampling \( m \) rows from \( H \) uniformly at random (without replacement)
   - the matrix \( D \) is a diagonal matrix with i.i.d. \( \{-1, 1\} \) entries
Bibliography


