Agenda

• Structured random linear maps, two ways.
Johnson-Lindenstrauss lemma

• For any $\varepsilon \in (0, 1/2)$, point set $S \subset \mathbb{R}^d$ of size $|S| = n$, and $k \in \mathbb{N}$ satisfying $k \geq \frac{c \log n}{\varepsilon^2}$, there exists a linear map $f: \mathbb{R}^d \to \mathbb{R}^k$ satisfying

$$(1 - \varepsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2 \quad \forall x, y \in S$$
Main probabilistic lemma

• Ways to construct $f$:
  1. Orthogonal projection to a unif. random $k$-dim. subspace, scaled by $\sqrt{d/k}$.
  2. Map $x \mapsto \frac{1}{\sqrt{k}}Ax$ where entries of $A$ are iid $N(0, 1)$.
  3. As above, except $P(A_{i,j} = +1) = P(A_{i,j} = -1) = 1/2$ (“Rademacher random variables”).

• J-L property: for any unit vector $x \in \mathbb{R}^d$,
  \[
  P (\|f(x)\|_2^2 \geq 1 + \varepsilon) \leq \exp \left( \Omega(k\varepsilon^2) \right)
  \\
  P (\|f(x)\|_2^2 \leq 1 - \varepsilon) \leq \exp \left( \Omega(k\varepsilon^2) \right)
  \]

• Variance of $\|f(x)\|_2^2$ is roughly $O(1/k)$.

• Apply to $(x - y)/\|x - y\|_2$ for each pair $x, y \in S$ with union bound.
Fast random linear maps?

• Linear maps previously discussed apparently require $O(kd)$ time to apply to an arbitrary vector $x \in \mathbb{R}^d$.

• Are there linear maps that
  • (1) satisfy the J-L property, and
  • (2) can be applied in time faster than $O(kd)$?
Sparse linear maps

• Suppose $M \in \mathbb{R}^{k \times d}$ is sparse: i.e., $\text{nnz}(M) \ll kd$.
  • Here, we have the canonical orthonormal bases in mind.

• Then can apply $M$ in time $O(\text{nnz}(M)) \ll O(kd)$ time.

• How sparse can $M$ be and still satisfy J-L property?
Bernoulli-Gaussian random matrices

- Entries of random matrix $M$ are iid,
  \[ M_{i,j} = \frac{1}{\sqrt{\theta k}} A_{i,j} B_{i,j} \]
  where $A_{i,j} \sim \mathcal{N}(0,1)$ and $B_{i,j} \sim \text{Bernoulli}(\theta)$,
  ($A_{i,j}$ and $B_{i,j}$ also independent).

- Scaling ensures that $\mathbb{E}\|Mx\|_2^2 = 1$ for every unit vector $x$.

\[ \mathbb{E}(\text{nnz}(M)) = \theta kd \]
J-L property for sparse random linear maps

- Distribution of $\|Mx\|_2^2$:

$$
\|Mx\|_2^2 = \sum_{i=1}^{k} \left( \sum_{j=1}^{d} \frac{1}{\sqrt{\theta k}} A_{i,j} B_{i,j} x_j \right)^2 
\text{dist} \frac{1}{\theta k} \sum_{i=1}^{k} \left( \sum_{j=1}^{d} B_{i,j} x_j^2 \right) Z_i^2
$$

where $Z_1, Z_2, \ldots, Z_k$ are iid $N(0,1)$.

- If $x = (1,0,\ldots,0)$, then $\text{var}(\|Mx\|_2^2) = \frac{3-\theta}{\theta k}$.

- Therefore, variance is $O(1/k)$ only if $\theta = \Omega(1)$. 
Sparse random linear maps of dense vectors

- Sparse random linear map appears to fail on account of sparse $x$’s.
- If $x$ was dense (e.g., $x = (1,1,\ldots,1)/\sqrt{d}$), then get “averaging” effect:

  $$
  \|Mx\|_2^2 \xrightarrow{\text{dist}} \frac{1}{\theta k} \sum_{i=1}^{k} \left( \sum_{j=1}^{d} B_{i,j} x_j^2 \right) Z_i^2
  $$

- If $|x_i| \leq \alpha$ for all $i$, then $\text{var}(\|Mx\|_2^2) = O((1 + \alpha^2/\theta)/k)$.

Can have $\theta = \alpha^2$ and still have $\text{var}(\|Mx\|_2^2) = O(1/k)$. 
Idea: pre-densification

• Compose two linear maps:
  1. **Densification**: \( x \mapsto y := Qx \)
  2. **Dimensionality reduction**: \( y \mapsto \frac{1}{\sqrt{k}} (A \odot B)y \in \mathbb{R}^k \)

• **Question**: How can we orthogonally transform a unit vector \( x \mapsto y \) so that \( y \) is guaranteed to be dense (e.g., \( |x_i| = \tilde{O}(1/\sqrt{d}) \) for all \( i \))?
Densification via uniformly random rotation

• Pick an orthonormal basis $u_1, u_2, \ldots, u_d$ for $\mathbb{R}^d$ uniformly at random.
• Map $x \in S^{d-1}$ to $y := (\langle u_1, x \rangle, \langle u_2, x \rangle, \ldots, \langle u_d, x \rangle)$.
• With high probability,
  \[ |y_i| \leq O\left(\sqrt{(\log d)/d}\right) \text{ for all } i = 1, 2, \ldots, d. \]

But computing this takes $O(d^2)$ time!
Randomized Hadamard transform

• Instead of a completely unstructured orthogonal transformation, Ailon and Chazelle (2009) suggest a highly structured one:

\[ x \mapsto \frac{1}{\sqrt{d}} HDx \]

where \( H = H_d \) is the \( d \times d \) Hadamard matrix (deterministic), and \( D \) is diagonal where \( D_{1,1}, D_{2,2}, \ldots, D_{d,d} \) are iid Rademacher r.v.’s.

• Only randomizes over \( 2^d \) possible orthogonal transformations, but this is sufficient for our purposes.
Hadamard matrices

• Recursive definition (for \(d\) a power of two):

\[
H_1 = +1, \quad H_d = \begin{bmatrix} +H_{d/2} & +H_{d/2} \\ +H_{d/2} & -H_{d/2} \end{bmatrix}
\]

• Example: \(d = 4\)

\[
H_4 = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix}
\]

• **Fact:** \(\frac{1}{\sqrt{d}} H_d\) is an orthogonal matrix. Therefore so is \(\frac{1}{\sqrt{d}} H_d D\).
Hadamard transform via divide-and-conquer

• To compute $H_d x$:
  • Partition $x = (x', x'')$, so $x', x'' \in \mathbb{R}^{d/2}$.
  • Recursively compute $H_{d/2} x'$ and $H_{d/2} x''$.
  • Compute $H_{d/2} x' + H_{d/2} x''$ and $H_{d/2} x' - H_{d/2} x''$.
  • Return $H_d x = \begin{bmatrix} H_{d/2} x' + H_{d/2} x'' \\ H_{d/2} x' - H_{d/2} x'' \end{bmatrix}$

• Total time is $O(d \log d)$. 
Analysis of Randomized Hadamard transform

• Let \( y := \frac{1}{\sqrt{d}} HDx \) for an arbitrary fixed unit vector \( x \in S^{d-1} \).

• Want: Tail bound for \( \|y\|_{\infty} := \max_{i \in \{1,2,\ldots,d\}} |y_i| \).

\[
y_i = \frac{1}{\sqrt{d}} \sum_{j=1}^{d} H_{i,j} D_{j,x_j} \overset{\text{dist}}{=} \frac{1}{\sqrt{d}} \sum_{j=1}^{d} x_j \sigma_j
\]

where \( \sigma_1, \sigma_2, \ldots, \sigma_d \) are iid Rademacher random variables.

• **Fact:** MGF of \( \sum_{j=1}^{d} x_j \sigma_j \) is dominated by that of \( N(0, \sum_{j=1}^{d} x_j^2) = N(0,1) \).

• **Conclusion:** \( P(\|y\|_{\infty} > \sqrt{\frac{\log(d/\delta)}{d}}) \leq \delta \).
Overall random linear map

- Composition of two linear maps:
  1. **Densification**: $x \mapsto y := \frac{1}{\sqrt{d}} HDx \in \mathbb{R}^d$
  2. **Dimensionality reduction**: $y \mapsto \frac{1}{\sqrt{\theta k}} (A \odot B)y \in \mathbb{R}^k$

- Overall expected running time: $O(d \log d + \theta kd)$.
  Can use $\theta \approx \frac{\log d}{d}$, in which case this is $O((d + k) \log d) = \tilde{O}(d + k)$.
Technical details

• Need tail bound for $\|My\|_2^2$ conditioned on event that $y$ is dense.
• Once details are worked out, guarantees are somewhat weaker than for (say) Gaussian random matrix.
• But morally behaves much like Gaussian random matrix, except faster to compute.
Sparsity-respecting linear maps

• With Gaussian random matrix $M$, computing $Mx$ could be done in time $O(k \times \text{nnz}(x))$, provided that $x$ is represented in some sparse format (e.g., list of non-zero entries); **no explicit dependence on $d$**.

• But randomized Hadamard transform explicitly takes $\tilde{O}(d)$ time.

• Can we do better with a structured linear map?
Sparse linear map via random hash function

- Dasgupta, Kumar, and Sarlos (2010): use hash function to define a sparse linear map

\[ M = CD \]

- \( D \) is diagonal where \( D_{1,1}, D_{2,2}, \ldots, D_{d,d} \) are iid Rademacher r.v.’s.
- \( C \in \{0, 1\}^{k \times d} \) determined by random hash function \( h: [d] \to [k] \):

\[
C_{i,j} = \begin{cases} 
1, & h(j) = i \\
0, & h(j) \neq i
\end{cases}
\]
Computation

• Computation of $z := CDx$:
  • Initialize $z := 0$.
  • For each $j$ such that $x_j \neq 0$:
    • Compute $i := h(j)$.
    • $z_i := z_i + D_{j,j} x_j$.
  • So total time is $O(\text{nnz}(x))$ (assume $y$ pre-initialized).
Analysis

• Unfortunately, can’t expect such a sparse matrix to work well on sparse vectors.

• For any unit vector $x \in S^{d-1}$, $\|CDx\|_2^2 - 1 \leq O\left(\frac{1}{\sqrt{k}} + \|x\|_\infty^2\right)$ w.h.p.

• Fortunately, there is a better way to densify.
Densification via replication

• Can pre-densify a vector by replicating it (say, $r$ times):
  
  \[ x \mapsto y := \frac{1}{\sqrt{r}} (x, x, \ldots, x) \in \mathbb{R}^{dr}. \]

• For any unit vector $x \in S^{d-1}$, the resulting vector $y$ is a unit vector (though in $\mathbb{R}^{dr}$) with $\|y\|_\infty \leq \frac{1}{\sqrt{r}}$.

• Can combine with previous sparse random linear map, so computation time $O(r \times \text{nnz}(x))$.

• Use $k = \tilde{O}(1/\epsilon^2)$ and $r = \tilde{O}(1/\epsilon)$ so that $\|CDx\|_2^2 - 1 \leq \epsilon$ w.h.p.
Analysis (details)

• For any unit vector $x \in S^{d-1}$, $|\|CDx\|_2^2 - 1| \leq \tilde{O}\left(\frac{1}{\sqrt{k}} + \|x\|_\infty^2\right)$ w.h.p.

• Basic idea of analysis:
  • $\|CDx\|_2^2 - 1$ is a quadratic form in the iid Rademacher random variables $D_{1,1}, D_{2,2}, ..., D_{d,d}$ (collect these in a random vector $\sigma$):
    $$\sigma^\top \text{diag}(x)(C^\top C - I)\text{diag}(x)\sigma$$
    where $\text{diag}(x)$ is diagonal matrix with entries of $x$ on diagonal.
  • Expected value (w.r.t. $\sigma$) is $\mathbb{E}_\sigma(\|CDx\|_2^2 - 1) = 0$.
  • Behaves like quadratic forms of Gaussian random vectors: more on this later.
    • Deviation will be small when $\text{diag}(x)(C^\top C - I)\text{diag}(x)$ is “small”.
    • For example, $(i, j)$-th entry of $C^\top C - I$ is zero when $h(i) \neq h(j)$. 