Ramdomized Projection Methods

Martin J. Strauss

mstrauss@research.att.com

Ramdomized Projection Methods

Given several high-dimensional vectors, want...

- to represent them concisely
- to preserve approximate distances
- to succeed with high probability *on all inputs*

We'll present a basic algorithm. Avoid questions:

- How are the vectors presented? (All at once, no changes.)
- *Which* distance is preserved? (Euclidean, not edit, nor ...)

Advantages

- Save space—store only small "sketches"
- Save time
  - operate only on sketches
  - avoid unnecessary disk accesses

The Problem

Given vectors

$$\begin{cases} \mathbf{A} = \langle \mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2, \ldots \rangle \\ \mathbf{B} = \langle \mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2, \ldots \rangle, \end{cases}$$

build *sketches* $h(\mathbf{A})$ and $h(\mathbf{B})$; recover $\|\mathbf{A} - \mathbf{B}\|^2 = \sum_i (\mathbf{A}_i - \mathbf{B}_i)^2$

$$\begin{tikzcd}
\mathbf{A} \arrow[rr] \arrow[ddr, dashed] & & h(\mathbf{A}) \arrow[ddl] & \text{sketches} \\
\mathbf{B} \arrow[rr] \arrow[dr, dashed] & & h(\mathbf{B}) \arrow[dl] & \\
& \|\mathbf{A} - \mathbf{B}\|^2 & &
\end{tikzcd}$$

Estimating Norms [JL84, AMS96]

Goal: $\boxed{\|\mathbf{A}\|_{\sim}^{2} = (1 \pm \epsilon)\, \|\mathbf{A}\|^{2}.}$

$R_i = \pm 1$ random variables.

Put $X = h(\mathbf{A}) = \sum \mathbf{A}_i R_i$.

$$\begin{aligned}
E[X^2] &= E\left(\sum \mathbf{A}_i R_i\right)^2 \\
&= E\left(\sum_i \mathbf{A}_i^2 R_i^2\right) + E\left(\sum_{i \neq j} \mathbf{A}_i \mathbf{A}_j R_i R_j\right) \\
&= \sum_i \mathbf{A}_i^2 .
\end{aligned}$$

Similarly, $E[X^4] \leq 3E^2[X^2]$.

Estimate $E[X^2]$ from samples of X^2.

Reducing Distortion

Reduce distortion through averaging:

- Given $E[X_j] = \mu$ and $\mathrm{var}(X_j) = \sigma^2$, iid.
- Put $Y = \mathrm{avg}_{8/\epsilon^2} X_j$.
- Then $\mathrm{var}(Y) = \epsilon^2\sigma^2/8$.
- Chebychev:

$$\begin{aligned}\Pr\left(|Y-\mu|^2 \geq \epsilon^2\sigma^2\right) &\leq \mathrm{var}(Y)/(\epsilon^2\sigma^2)\\ &\leq 1/8.\end{aligned}$$

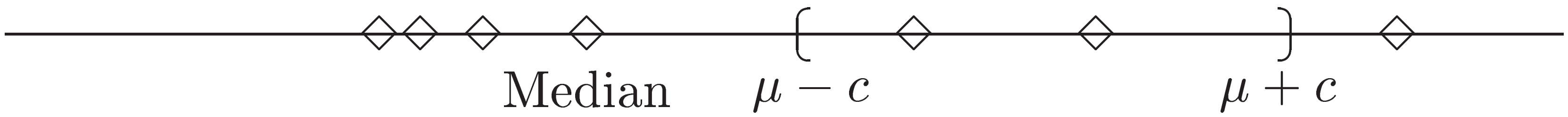
Here $\mu = \|\mathbf{A}\|^2$ and $\sigma \leq O(\mu)$.

Get $\Pr\left(|Y-\mu| \geq \epsilon\mu\right) < 1/8$.

Improving Success Probability

Improve success probability through median:

- Let $Z = \text{median}_{O(\log(1/\delta))} Y_j$, iid.
- $|Z - \mu| > c$ iff $|Y_j - \mu| > c$ for $\geq 1/2$ the Y_j's
- Expect $\leq 1/8$
- Chernoff: $\Pr(|Z - \mu| > c) < \delta$.



Results

- Get $\|\mathbf{A} - \mathbf{B}\|_{\sim}$ from $h(\mathbf{A} - \mathbf{B}) = h(\mathbf{A}) - h(\mathbf{B})$.
- $\|\mathbf{A}\|_{\sim}^{2} = (1 \pm \epsilon)\,\|\mathbf{A}\|^{2}$ with probability $1 - \delta$.
- Sketch size: $O(\log(1/\delta)/\epsilon^{2})$ numbers.
- Reconstruction time: similar to size.