0368-3248-01-Algorithms in Data Mining **Fall 2012**

Lecture 6: Assignment 2 answers

Lecturer: Edo Liberty

Warning: *This note may contain typos and other inaccuracies which are usually discussed during class. Please do not cite this note as a reliable source. If you find mistakes, please inform me.*

---

1 Weak random projections

setup

In this question we will construct a simple and weak version of random projections. That is, given two vectors $x, y \in \mathbb{R}^d$ we will find two new vectors $x', y' \in \mathbb{R}^k$ such that from x' and y' we could approximate the value of $||x - y||$. The idea is to define k vectors $r_i \in \mathbb{R}^d$ such that each $r_i(j)$ takes a value in $\{+1, -1\}$ uniformly at random. Setting $x'(i) = r_i^T x$ and $y'(i) = r_i^T y$ the questions will lead you through arguing that $\frac{1}{k}||x' - y'||_2^2 \approx ||x - y||_2^2$.

questions

1. Let $z = x - y$, and $z' = x' - y'$. Show that $z'(\ell) = r_\ell^T z$ for any index $\ell \in [1, \ldots, k]$.
2. Show that $E[\frac{1}{k}||z'||_2^2] = E[(z'(\ell))^2] = ||z||_2^2$.
3. Show that
$$\text{Var}[(z'(\ell))^2] \leq 4||z||_2^4.$$
Hint: for any vector w we have $||w||_4 \leq ||w||_2$.
4. From 3 (even if you did not manage to show it) claim that
$$\text{Var}[\frac{1}{k}||z'||_2^2] \leq 4||z||_2^4/k.$$
5. Use 3 and Chebyshev's inequality do obtain a value for k for which:
$$(1 - \varepsilon)||x - y||_2^2 \leq \frac{1}{k}||x' - y'||_2^2 \leq (1 + \varepsilon)||x - y||_2^2$$
with probability at least $1 - \delta$.

2 Answers

1. This is a consequence of the linearity of the operator.
$$z'(\ell) = x'(\ell) - y'(\ell) = r_\ell^T x - r_\ell^T y = r_\ell^T (x - y) = r_\ell^T z$$

2. Since $||z'||_2^2 = \sum_{i=1}^k z'(i)^2$ and since $z'(i)$ are identically distributed we have that $\mathbb{E}[\frac{1}{k}||z'||_2^2] = \mathbb{E}[\frac{1}{k}\sum_{i=1}^k z'(i)^2] = \mathbb{E}[(z'(\ell))^2]$. Now we compute $\mathbb{E}[(z'(\ell))^2]$.

$$\begin{aligned}
\mathbb{E}[(z'(\ell))^2] &= \mathbb{E}[(\sum_{i=1}^d r_\ell(i)z(i))(\sum_{j=1}^d r_\ell(j)z(j))] && (1)\\
&= \mathbb{E}[\sum_{i=1}^d\sum_{j=1}^d r_\ell(i)r_\ell(j)z(i)z(j)] && (2)\\
&= \sum_{i=1}^d\sum_{j=1}^d \mathbb{E}[r_\ell(i)r_\ell(j)]z(i)z(j) && (3)\\
&= \sum_{i=1}^d z(i)^2 = \|z\|^2 && (4)
\end{aligned}$$

The double summation was reduced to a single sum since $\mathbb{E}[r_\ell(i)r_\ell(j)] = 0$ if $i \neq j$. Also, if $i = j$ we have that $\mathbb{E}[r_\ell(i)r_\ell(j)]z(i)z(j) = z(i)^2$

3. To compute $\mathrm{Var}[(z'(\ell))^2]$ we start with computing $\mathbb{E}[(z'(\ell))^4]$.

$$\begin{aligned}
\mathbb{E}[(z'(\ell))^4] &= \mathbb{E}[(\sum_{i=1}^d r_\ell(i)z(i))(\sum_{j=1}^d r_\ell(j)z(j))(\sum_{k=1}^d r_\ell(k)z(k))(\sum_{m=1}^d r_\ell(m)z(m))]\\
&= \mathbb{E}[\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d\sum_{m=1}^d r_\ell(i)r_\ell(j)r_\ell(k)r_\ell(m)z(i)z(j)z(k)z(m)]\\
&= \sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d\sum_{m=1}^d \mathbb{E}[r_\ell(i)r_\ell(j)r_\ell(k)r_\ell]z(i)z(j)z(k)z(m)\\
&= \sum_{i=1}^d x(i)^4 + \binom{4}{2}\sum_{i<j} z(i)^2z(j)^2
\end{aligned}$$

The last transition requires an explanation. The expectation of $r_\ell(i)r_\ell(j)r_\ell(k)r_\ell$ when the power of one of the terms $r_\ell(i)$ is odd is zero. Thus, we are only left with terms of the form $x(i)^4$ and $x(i)^2x(j)^2$. The coefficient of $x(i)^4$ is 1 since there is only one what to obtain it. The coefficient of $x(i)^2x(j)^2$ is $\binom{4}{2}$ since two of the indexes should be i and the two others j. There are $\binom{4}{2} = 6$ to get it. In what comes next we use the fact that:

$$\sum_{i<j} z(i)^2z(j)^2 = [\sum_{i=1}^d\sum_{j=1}^d z(i)^2z(j)^2 - \sum_{i=1}^d z(i)^4]/2$$

Picking up where we left off:

$$\begin{aligned}
\mathbb{E}[(z'(\ell))^4] &= \sum_{i=1}^d x(i)^4 + 6\sum_{i<j} z(i)^2z(j)^2\\
&= \sum_{i=1}^d x(i)^4 + 3[\sum_{i=1}^d\sum_{j=1}^d z(i)^2z(j)^2 - \sum_{i=1}^d z(i)^4]\\
&= 3\|z\|_2^4 - 2\|z\|_4^2
\end{aligned}$$

Finally we have that

$$\begin{aligned}
\mathrm{Var}(z'(\ell)^2) &= \mathbb{E}[(z'(\ell))^4] - \mathbb{E}[(z'(\ell))^2]^2\\
&= 3\|z\|_2^4 - 2\|z\|_4^2 - (\|z\|_2^2)^2 = 2(\|x\|_2^4 - \|x\|_4^4) \leq 2\|x\|_2^4
\end{aligned}$$

4. Since $z'(\ell)$ are independent variables we have that

$$\mathrm{Var}[\frac{1}{k}\|z'\|^2] = \mathrm{Var}[\frac{1}{k}\sum_{\ell=1}^{k} z'(\ell)^2] = \frac{1}{k^2}\sum_{\ell=1}^{k}\mathrm{Var}[z'(\ell)^2] = \frac{1}{k}\mathrm{Var}[z'(\ell)^2] \leq 2\|x\|_2^4/k$$

5. From Chebishev's inequality we have that

$$\Pr[|\frac{1}{k}\|z'\|^2 - \mathbb{E}[\frac{1}{k}\|z'\|^2]| \geq t] \leq \frac{\mathrm{Var}[\frac{1}{k}\|z'\|^2]}{t^2}$$

Substituting $\mathbb{E}[\frac{1}{k}\|z'\|^2] = \|z\|^2$, $t = \varepsilon\|z\|^2$ and $\mathrm{Var}[\frac{1}{k}\|z'\|^2] \leq 2\|x\|_2^4/k$ we get:

$$\Pr[|\frac{1}{k}\|z'\|^2 - \|z\|| \geq \varepsilon\|z\|] \leq \frac{2\|x\|_2^4/k}{\varepsilon^2\|z\|^4} = \frac{2}{k\varepsilon^2}$$

By setting $k \geq \frac{2}{\varepsilon^2\delta}$ we get that $\Pr[|\frac{1}{k}\|z'\|^2 - \|z\|| \geq \varepsilon\|z\|] \leq \delta$ which means that $\|z\|(1-\varepsilon) \leq \frac{1}{k}\|z'\|^2 \leq \|z\|(1+\varepsilon)$ with probability at least $1-\delta$.