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
   \[
   \text{Var}[(z'(\ell))^2] \leq 4 ||z||_2^2.
   \]
   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
   \[
   \text{Var}\left[\frac{1}{k} ||z'||_2^2\right] \leq 4 ||z||_2^4 / k.
   \]

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.
   \[ z'(t) = x'(t) - y'(t) = r_t^T x - r_t^T y = r_t^T (x - y) = r_t^T z \]

2. Since \( ||z'||^2 = \sum_{i=1}^{k} z'(i)^2 \) and since \( z'(i) \) are identically distributed we have that \( \mathbb{E}[\frac{1}{k} ||z'||^2] = \mathbb{E}[\frac{1}{k} \sum_{i=1}^{k} z'(i)^2] = \mathbb{E}[(z'(t))^2] \). Now we compute \( \mathbb{E}[(z'(t))^2] \).

   \[ \mathbb{E}[(z'(t))^2] = \mathbb{E}[\left( \sum_{i=1}^{d} r_t(i)z(i) \right) \left( \sum_{j=1}^{d} r_t(j)z(j) \right)] \]
   \[ = \mathbb{E}\left[ \sum_{i=1}^{d} \sum_{j=1}^{d} r_t(i)r_t(j)z(i)z(j) \right] \]
   \[ = \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbb{E}[r_t(i)r_t(j)]z(i)z(j) \]
   \[ = \sum_{i=1}^{d} z(i)^2 = ||z||^2 \]

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

3. To compute \( \text{Var}[(z'(t))^2] \) we start with computing \( \mathbb{E}[(z'(t))^4] \).

   \[ \mathbb{E}[(z'(t))^4] = \mathbb{E}\left[ \left( \sum_{i=1}^{d} r_t(i)z(i) \right) \left( \sum_{j=1}^{d} r_t(j)z(j) \right) \left( \sum_{k=1}^{d} r_t(k)z(k) \right) \left( \sum_{m=1}^{d} r_t(m)z(m) \right) \right] \]
   \[ = \mathbb{E}\left[ \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{m=1}^{d} r_t(i)r_t(j)r_t(k)r_t(m)z(i)z(j)z(k)z(m) \right] \]
   \[ = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{m=1}^{d} \mathbb{E}[r_t(i)r_t(j)r_t(k)r_t(m)]z(i)z(j)z(k)z(m) \]
   \[ = \sum_{i=1}^{d} \sum_{j<i}^{d} \sum_{k=1}^{d} \sum_{m=1}^{d} \mathbb{E}[r_t(i)r_t(j)r_t(k)r_t(m)]z(i)z(j)z(k)z(m) \]
   \[ = \sum_{i=1}^{d} \sum_{j<i}^{d} x(i)^4 + \binom{d}{2} \sum_{i<j} z(i)^2 z(j)^2 \]

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

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

Picking up where we left off:

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

Finally we have that

\[ \text{Var} (z'(\ell)^2) = E[(z'(\ell))^4] - E[(z'(\ell))^2]^2 \]
\[ = 3\|z\|_4^4 - 2\|z\|_2^2 - (\|z\|_2^2)^2 = 2(\|x\|_2^4 - \|x\|_2^4) \leq 2\|x\|_2^4 \]

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

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

5. From Chebishev’s inequality we have that

\[ \Pr[|\|z'\|_2^2 - E[\|z'\|_2^2]| \geq t] \leq \frac{\text{Var}[\frac{1}{k}\|z'\|^2]}{t^2} \]

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

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

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