random projection

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

A random projection replaces the feature vector $\mathbf{x} \in \mathbb{R}^d$ of a data point by $\mathbf{A}\mathbf{x} \in \mathbb{R}^{d'}$, with $d' < d$ and a wide matrix $\mathbf{A}$ whose entries are realizations of independent and identically distributed (i.i.d.) random variables (RVs). No data enter $\mathbf{A}$, which separates a random projection from principal component analysis (PCA) and from an autoencoder, whose maps for dimensionality reduction are learned from a training set. When the entries have mean zero and variance $1/d'$, the expectation of the squared length of $\mathbf{A}\mathbf{x}$ equals the squared length of $\mathbf{x}$, so squared Euclidean distances between data points survive on average. The Johnson–Lindenstrauss lemma (JL lemma) turns this into a guarantee for a whole finite dataset: a target dimension of at least $4 \ln m/(\varepsilon^2/2 - \varepsilon^3/3)$ keeps every pairwise squared Euclidean distance within the factor $1 \pm \varepsilon$. That requirement involves the sample size m and the tolerance ε and not the ambient dimension d, so the saving grows with the length of the original feature vector. The entries may be normal RVs or take the values $\pm 1/\sqrt{d'}$ with probability one half each. A random projection costs one matrix multiplication per data point and no pass over the data.

Definition

A weather archive at Krems an der Donau (Austria) keeps, for each day, the daily maximum air temperature of the preceding ten years. Such a day is a data point whose feature vector $\mathbf{x} \in \mathbb{R}^d$ holds $d = 3650$ numbers, and a search for the days most similar to a given one compares these feature vectors by their Euclidean distance. A random projection shortens every record before the search: it replaces $\mathbf{x}$ by $\mathbf{A}\mathbf{x} \in \mathbb{R}^{d'}$, where $d' < d$ and $\mathbf{A} \in \mathbb{R}^{d' \times d}$ is a wide matrix whose entries are realizations of independent and identically distributed (i.i.d.) random variables (RVs). No data enter $\mathbf{A}$, and that is what separates a random projection from principal component analysis (PCA) and from an autoencoder, whose maps for dimensionality reduction are learned from a training set. The word *projection* is traditional here: the matrix $\mathbf{A}$ maps into $\mathbb{R}^{d'}$, so it is not a projection of $\mathbb{R}^d$ onto one of its subsets.

The probability distribution of these RVs has mean zero and variance $1/d'$. Consider a fixed $\mathbf{u} \in \mathbb{R}^d$. Each entry of $\mathbf{A}\mathbf{u}$ is then an RV with mean zero and variance $\|\mathbf{u}\|_2^2/d'$. Adding the variances of the d' entries gives the expectation

of the squared length of $\mathbf{A}\mathbf{u}$,

$$\mathbb{E}\{\|\mathbf{A}\mathbf{u}\|_2^2\} = \|\mathbf{u}\|_2^2. \tag{1}$$

Read with $\mathbf{u} = \mathbf{x}^{(r)} - \mathbf{x}^{(r')}$, the identity (1) says that the squared Euclidean distance between two data points survives the projection on average. One draw of $\mathbf{A}$ is another matter. For two of the Krems days and $d' = 16$, the squared Euclidean distance after the projection ranged from 0.22 to 2.68 times the one before it over 4000 draws.

That spread narrows as d' grows, and the Johnson–Lindenstrauss lemma (JL lemma), proved by Johnson and Lindenstrauss (1984), turns the narrowing into a guarantee that covers every pair of a finite dataset at once. Consider a dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ of sample size m and a tolerance ε with $0 < \varepsilon < 1$. Whenever the target dimension satisfies

$$d' \geq \frac{4 \ln m}{\varepsilon^2/2 - \varepsilon^3/3}, \tag{2}$$

there is a map $\mathbf{\Phi} \colon \mathbb{R}^d \to \mathbb{R}^{d'}$ with

$$(1-\varepsilon)\left\|\mathbf{x}^{(r)} - \mathbf{x}^{(r')}\right\|_2^2 \leq \left\|\mathbf{\Phi}\left(\mathbf{x}^{(r)}\right) - \mathbf{\Phi}\left(\mathbf{x}^{(r')}\right)\right\|_2^2 \leq (1+\varepsilon)\left\|\mathbf{x}^{(r)} - \mathbf{x}^{(r')}\right\|_2^2, \tag{3}$$

for every $r, r' \in \{1, \ldots, m\}$, and a random projection delivers such a map with positive probability, which repeated draws raise as high as wanted (Dasgupta and Gupta, 2003, Thm. 2.1). The requirement (2) involves the sample size m and the tolerance ε, and not the ambient dimension d. Its denominator is largest at $\varepsilon = 1$, where it equals $1/6$, so no tolerance below one is granted before $d' \geq 24 \ln m$, which is 134 for $m = 256$ days. The sample size enters through its logarithm: at $\varepsilon = 1/2$, a hundredfold larger dataset raises the requirement from 266 to 487. Fig. 1 holds the two-sided bound (3) against what one draw of $\mathbf{A}$ does to the 32640 pairwise squared Euclidean distances of 256 such days.

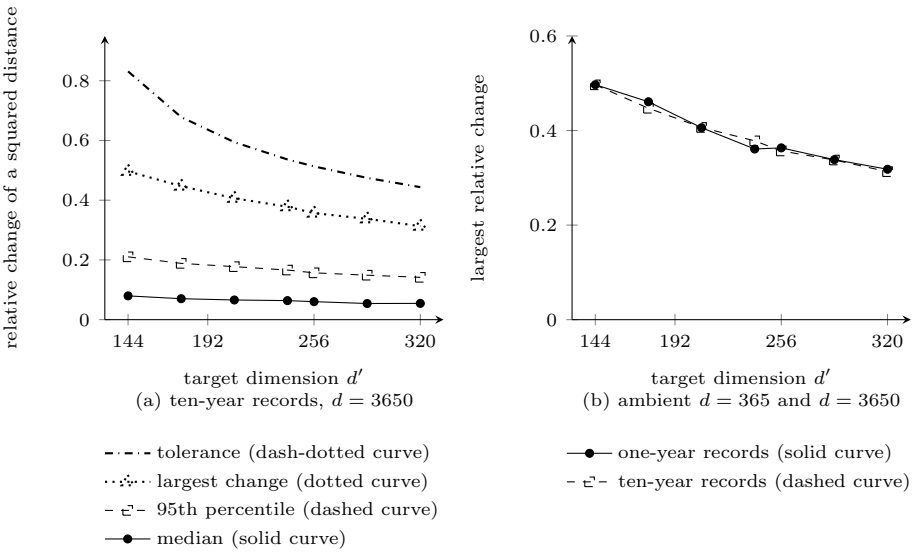


Figure 1: The relative change of the 32640 pairwise squared Euclidean distances of 256 days at Krems under a random projection, against the target dimension d', each value the median over 40 draws of $\mathbf{A}$. (a) For the ten-year records, $d = 3650$: the median over the pairs (solid curve, filled circles), their 95th percentile (dashed curve, open squares), the largest change (dotted curve, open triangles), and the tolerance ε that the requirement (2) grants (dash-dotted curve). (b) The largest change for the one-year records, $d = 365$ (solid curve, filled circles), and for the ten-year ones, $d = 3650$ (dashed curve, open squares), a tenfold difference in d that leaves the curve where it was. Data generated by `pythondemos/randomprojection.py`

Two probability distributions for the entries of $\mathbf{A}$ are in common use. Normal RVs with mean zero and variance $1/d'$ give the guarantee in the form stated above (Shalev-Shwartz and Ben-David, 2014, Lemma 23.4). Entries equal to $1/\sqrt{d'}$ or $-1/\sqrt{d'}$, each with probability one half, meet the same requirement (2) (Achlioptas, 2001; Dasgupta and Gupta, 2003, Sect. 3), and the products in $\mathbf{Ax}$ then reduce to additions and subtractions of the features. Either way the cost is dd' multiplications per data point, paid once: there is no iteration and no training.

For the Krems archive and $d' = 256$, the ten-year record of a day shrinks from 3650 numbers to 256, a factor of 14.3. In a typical draw of $\mathbf{A}$, no pair of the 32640 squared Euclidean distances changes by more than 36 percent and half of them change by less than 6 percent, against the 51 percent the requirement (2) grants at that target dimension. A method that reads the data points only through their pairwise Euclidean distances, such as k-nearest neighbors (k-NN), therefore works on nearly the dataset it would have worked on before. When both the map and the reconstruction are linear, PCA is the choice with the

smallest average squared error over the training set, and a random projection is not. When the step back to $\mathbb{R}^d$ may be nonlinear, the ranking can turn around: feature vectors with few nonzero features are recovered better from a random projection than from PCA (Shalev-Shwartz and Ben-David, 2014, Sect. 23.5).

Cross References

dimensionality reduction, feature learning, JL lemma, PCA, autoencoder, random matrix, wide matrix, Euclidean distance, feature vector, k-NN

References

1. Achlioptas (2001). Database-friendly random projections. In: PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 274–281. ACM.

2. Dasgupta and Gupta (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms, 22(1), pp. 60–65.

3. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

4. Johnson and Lindenstrauss (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary Math., 26, pp. 189–206.