Sign-Full Random Projections

Ping Li
Cognitive Computing Lab (CCL)
Baidu Research USA
Bellevue, WA 98004, USA
pingli98@gmail.com

Abstract

The method of 1-bit ("sign-sign") random projections has been a popular tool for efficient search and machine learning on large datasets. Given two D-dim data vectors $u, v \in \mathbb{R}^D$, one can generate $x = \sum_{i=1}^D u_i r_i$, and $y = \sum_{i=1}^D v_i r_i$, where $r_i \sim N(0,1)$ iid. Then one can estimate the cosine similarity ρ from $sgn(x)$ and $sgn(y)$. In this paper, we study a series of estimators for "sign-full" random projections. First we prove $E(sgn(x)y) = \sqrt{\frac{2}{\pi}}\rho$, which provides an estimator for ρ. Interestingly this estimator can be substantially improved by normalizing y. Then we study estimators based on $E\left(y_- 1_{x\geq 0} + y_+ 1_{x<0}\right)$ and its normalized version. We analyze the theoretical limit (using the MLE) and conclude that, among the proposed estimators, no single estimator can achieve (close to) the theoretical optimal asymptotic variance, for the entire range of ρ. On the other hand, the estimators can be combined to achieve the variance close to that of the MLE.

In applications such as near neighbor search, duplicate detection, knn-classification, etc, the training data are first transformed via random projections and then only the signs of the projected data points are stored (i.e., the $sgn(x)$). The original training data are discarded. When a new data point arrives, we apply random projections but we do not necessarily need to quantize the projected data (i.e., the y) to 1-bit. Therefore, sign-full random projections can be practically useful. This gain essentially comes at no additional cost.

Introduction

Consider two high-dimensional data vectors, $u, v \in \mathbb{R}^D$. Suppose we generate a D-dim random vector whose entries are iid standard normal, i.e., $r_i \sim N(0,1)$, and compute

$$x = \sum_{i=1}^D u_i r_i, \qquad y = \sum_{i=1}^D v_i r_i$$

We have in expectation $E(xy) = \langle u, v\rangle = \sum_{i=1}^D u_i v_i$. If we generate x and y independently for k times, then $\frac{1}{k}\sum_{j=1}^k x_j y_j \approx E(xy) = \langle u, v\rangle$, and the quality of approximation improves as k increases. This idea of random projections has been widely used for large-scale search and machine learning (Johnson and Lindenstrauss 1984; Vempala 2004; Papadimitriou et al. 1998; Dasgupta 1999; Datar et al. 2004; Li, Hastie, and Church 2006).

A popular variant is the "1-bit" random projections (Goemans and Williamson 1995; Charikar 2002), which we refer to as "sign-sign" random projections, based on the following result of "collision probability"

$$P\left(sgn(x) = sgn(y)\right) = 1 - \frac{\cos^{-1}\rho}{\pi} \tag{1}$$

where $\rho = \frac{\sum_{i=1}^D u_i v_i}{\sqrt{\sum_{i=1}^D u_i^2}\sqrt{\sum_{i=1}^D v_i^2}}$ is the "cosine" similarity between the two original data vectors u and v. The method of sign-sign random projections has become popular, for example, in many search related applications (Henzinger 2006; Manku, Jain, and Sarma 2007; Grimes 2008; Hajishirzi, Yih, and Kolcz 2010; Türe, Elsayed, and Lin 2011; Manzoor, Milajerdi, and Akoglu 2016).

Note that by using only the signs of the projected data, we lose the information about the norms of the original vectors. Thus, in this context, with no loss of generality, we assume that the original data vectors are normalized, i.e., $\sum_{i=1}^D u_i^2 = \sum_{i=1}^D v_i^2 = 1$. In other words, we can assume the projected data to be $x \sim N(0,1)$ and $y \sim N(0,1)$.

Interestingly, one can take advantage of $E(sgn(x)y)$ and several variants to considerably improve 1-bit random projections. This gain essentially comes at no additional cost. Basically, the training data after projections are stored using signs (e.g., $sgn(x)$). When a new data vector arrives, however, we need to generate its random projections (y) but do not necessarily have to quantize them.

Related Work

The idea of "1-bit" projections has been extended to "sign cauchy projections" for estimating χ^2 similarity (Li, Samorodnitsky, and Hopcroft 2013), and to "1-bit minwise hashing" (Broder 1997; Li and König 2010) for estimating the resemblance between sets. More general quantization schemes of random projections have been studied in, for example, (Li and Slawski 2017).

In the context of random projections, the idea of estimating ρ from $sgn(x)$ and y was explored in (Dong, Charikar, and Li 2008). Similar ideas were also studied in the quantization literature (without using random projections) (Jégou, Douze, and Schmid 2011).

Review of Estimators Based on Full Information after Random Projections

In this context, since we are only concerned with estimating the cosine ρ, we can without loss of generality assume that the original data are normalized, i.e., $\|u\| = \|v\| = 1$. The projected data thus follow a bi-variant normal distribution:

$$\begin{bmatrix} x_j \\ y_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \text{ iid} \qquad j = 1, 2, ..., k.$$

where $\rho = \sum_{i=1}^{D} u_i v_i$. The obvious estimator for ρ is based on the inner product:

$$\hat{\rho}_f = \frac{1}{k}\sum_{j=1}^{k} x_j y_j, \qquad E\left(\hat{\rho}_f\right) = \rho,$$

$$Var\left(\hat{\rho}_f\right) = \frac{V_f}{k}, \qquad V_f = 1 + \rho^2$$

one can find the derivation of variance (i.e., V_f) in (Li, Hastie, and Church 2006). Note that $Var(\hat{\rho}_f)$ is the largest when $|\rho| = 1$. This is disappointing, because when two data vectors are identical, we ought to be able to estimate their similarity with no error. One can improve the estimator by simply normalizing the projected data. See (Anderson 2003) for the derivation.

$$\hat{\rho}_{f,n} = \frac{\sum_{j=1}^{k} x_j y_j}{\sqrt{\sum_{j=1}^{k} x_j^2}\sqrt{\sum_{j=1}^{k} y_j^2}}, \qquad E\left(\hat{\rho}_{f,n}\right) = \rho + O\left(\frac{1}{k}\right)$$

$$Var\left(\hat{\rho}_{f,n}\right) = \frac{V_{f,n}}{k} + O\left(\frac{1}{k^2}\right), \qquad V_{f,n} = \left(1-\rho^2\right)^2$$

In particular, $V_{f,n} = 0$ when $|\rho| = 1$, as desired. One can further improve $\hat{\rho}_{f,n}$ but not too much. The theoretical limit (i.e., the Cramér-Rao bound) of the asymptotic variance (Lehmann and Casella 1998) can be obtained by the maximum likelihood estimator (MLE), which is the solution of the following cubic equation:

$$\rho^3 - \rho^2 \sum_{j=1}^{k} x_j y_j + \rho\left(-1 + \sum_{i=1}^{k} x_j^2 + \sum_{j=1}^{k} y_j^2\right) - \sum_{j=1}^{k} x_j y_j = 0$$

This cubic equation can have multiple real roots with a small probability (Li, Hastie, and Church 2006), which decreases exponentially fast with increasing k. The MLE is asymptotically unbiased and its asymptotic variance becomes:

$$E\left(\hat{\rho}_{f,m}\right) = \rho + O\left(\frac{1}{k}\right),$$

$$Var\left(\hat{\rho}_{f,m}\right) = \frac{V_{f,m}}{k} + O\left(\frac{1}{k^2}\right), \qquad V_{f,m} = \frac{\left(1-\rho^2\right)^2}{1+\rho^2}$$

Estimator Based on Sign-Sign Random Projections

From $Pr\left(sgn(x_j) = sgn(y_j)\right) = 1 - \frac{1}{\pi}\cos^{-1}\rho$, we have an asymptotically unbiased estimator and its variance:

$$\hat{\rho}_1 = \cos\pi\left(1 - \frac{1}{k}\sum_{j=1}^{k} 1_{sgn(x_j)=sgn(y_j)}\right),$$

$$E\left(\hat{\rho}_1\right) = \rho + O\left(\frac{1}{k}\right), \qquad Var\left(\hat{\rho}_1\right) = \frac{V_1}{k} + O\left(\frac{1}{k^2}\right),$$

$$V_1 = \cos^{-1}\rho\left(\pi - \cos^{-1}\rho\right)\left(1-\rho^2\right)$$

As later will be shown in Lemma 2, we have when $|\rho| \to 1$,

$$V_1 = 2\sqrt{2}\pi\left(1-|\rho|\right)^{3/2} + o\left(\left(1-|\rho|\right)^{3/2}\right),$$

This rate is slower than $O\left((1-|\rho|)^2\right)$, which is the rate at which $V_{f,n}$ and $V_{f,m}$ approach 0. Figure 1 compares the estimators in terms of $\frac{V_1}{V_{f,m}}$, $\frac{V_f}{V_{f,m}}$, and $\frac{V_{f,n}}{V_{f,m}}$. Basically, $V_{f,n} < V_f$ always which means we should always use the normalized estimator. Note that $V_1 < V_f$ if $|\rho| > 0.5902$.

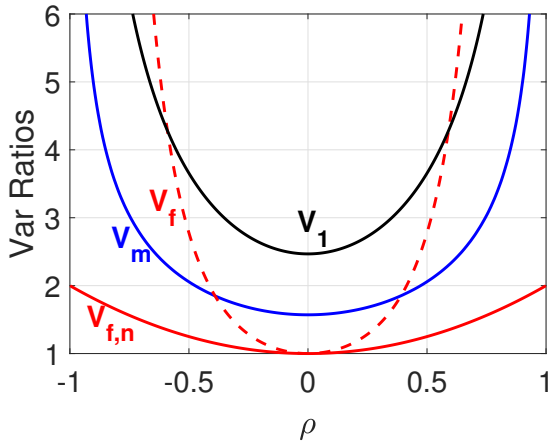

Figure 1: Ratios (lower the better) of variance factors: $\frac{V_1}{V_{f,m}}$, $\frac{V_f}{V_{f,m}}$, $\frac{V_m}{V_{f,m}}$, $\frac{V_{f,n}}{V_{f,m}}$. They are always larger than 1, as $V_{f,m}$ is the theoretically smallest variance factor. Note that V_m is the variance factor for the MLE of sign-full random projections.

Estimators for Sign-Full Random Projections

In many practical scenarios such as near-neighbor search and near-neighbor classification, we can store signs of the projected data (i.e., $sgn(x_j)$) and discard the original high-dimensional data. When a new data vector arrives, we generate its projected vector (i.e., y). At this point we actually have the option to choose whether we would like to use the full information or just the signs (i.e., $sgn(y_j)$) to estimate the similarity ρ. If we are able to find a better (more accurate) estimator by using the full information of y_j, there is no reason why we have to only use the sign of y_j.

We first study the maximum likelihood estimator (MLE), in order to understand the theoretical limit.

Theorem 1 *Given k iid samples $(sign(x_j),\ y_j)$, $j = 1, 2, ..., k$, with $x_j, y_j \sim N(0,1)$, $E(x_j y_j) = \rho$, the maximum likelihood estimator (MLE, denoted by $\hat{\rho}_m$) is the solution to the following equation:*

$$\sum_{j=1}^{k} \frac{\phi\left(\frac{\rho}{\sqrt{1-\rho^2}} sgn(x_j) y_j\right)}{\Phi\left(\frac{\rho}{\sqrt{1-\rho^2}} sgn(x_j) y_j\right)} sgn(x_j) y_j = 0 \qquad (2)$$

$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ and $\Phi(t) = \int_{-\infty}^{t} \phi(t)dt$ are respectively the pdf and cdf of standard normal.

$$E\left(\hat{\rho}_m\right)=\rho+O\left(\frac{1}{k}\right),\qquad Var\left(\hat{\rho}_m\right)=\frac{V_m}{k}+O\left(\frac{1}{k^2}\right),$$

$$\begin{aligned}\frac{1}{V_m}=&E\left\{\frac{\rho}{(1-\rho^2)^{7/2}}\frac{\phi\left(\frac{\rho}{\sqrt{1-\rho^2}}sgn(x_j)y_j\right)}{\Phi\left(\frac{\rho}{\sqrt{1-\rho^2}}sgn(x_j)y_j\right)}sgn(x_j)y_j^3\right\}\\&+E\left\{\frac{1}{(1-\rho^2)^3}\frac{\phi^2\left(\frac{\rho}{\sqrt{1-\rho^2}}sgn(x_j)y_j\right)}{\Phi^2\left(\frac{\rho}{\sqrt{1-\rho^2}}sgn(x_j)y_j\right)}y_j^2\right\}\\&-E\left\{\frac{3\rho}{(1-\rho^2)^{5/2}}\frac{\phi\left(\frac{\rho}{\sqrt{1-\rho^2}}sgn(x_j)y_j\right)}{\Phi\left(\frac{\rho}{\sqrt{1-\rho^2}}sgn(x_j)y_j\right)}sgn(x_j)y_j\right\}\end{aligned}\tag{3}$$

□.

As the MLE equation (2) is quite sophisticated, we study this estimator mainly for theoretical interest, for example, for evaluating other estimators. We can evaluate the expectations in (3) by simulations. Figure 1 already plots $\frac{V_m}{V_{f,m}}$, to compare $\hat{\rho}_m$ with three estimators: $\hat{\rho}_1$, $\hat{\rho}_f$, $\hat{\rho}_{f,n}$. The figure shows that $\hat{\rho}_m$ indeed substantially improves $\hat{\rho}_1$.

Next, we seek estimators which are much simpler than $\hat{\rho}_m$. Ideally, we look for estimators which can be written as "inner products". In this paper, we study **four** such estimators. We first present a Lemma which will be needed for deriving these estimators and proving their properties. The results can also be easily validated by numerical integrations.

Lemma 1

$$\int_0^\infty te^{-t^2/2}\Phi\left(\frac{\rho t}{\sqrt{1-\rho^2}}\right)dt=\frac{1+\rho}{2}\tag{4}$$

$$\int_0^\infty t^3e^{-t^2/2}\Phi\left(\frac{\rho t}{\sqrt{1-\rho^2}}\right)dt=\frac{1}{2}\left(2+3\rho-\rho^3\right)\tag{5}$$

$$\begin{aligned}&\int_0^\infty t^2e^{-t^2/2}\Phi\left(\frac{\rho t}{\sqrt{1-\rho^2}}\right)dt\\&=1_{\rho\geq0}\sqrt{\frac{\pi}{2}}-\sqrt{\frac{1}{2\pi}}\left(\tan^{-1}\frac{\sqrt{1-\rho^2}}{\rho}-\rho\sqrt{1-\rho^2}\right)\end{aligned}\tag{6}$$

where we denote that $\tan^{-1}\left(\frac{1}{0}\right)=\tan^{-1}\left(\frac{1}{0+}\right)=\frac{\pi}{2}$. □

The first estimator we present is based on the (odd) moments of $(sgn(x_j)y_j)$ as shown in Theorem 2.

Theorem 2

$$E(sgn(x_j)y_j)=\sqrt{\frac{2}{\pi}}\rho,\tag{7}$$

$$E\left((sgn(x_j)y_j)^3\right)=\frac{1}{\sqrt{2\pi}}\left(6\rho-2\rho^3\right)\tag{8}$$

Proof Sketch: *Note that:* $E\left((sgn(x_j)^2y_j)^2\right)=1$, *and* $E\left((sgn(x_j)y_j)^4\right)=3$. *Because* (x_j,x_j) *is bi-variate normal, we have* $x_j|y_j\sim N\left(\rho y_j,\ (1-\rho^2)\right)$, *and*

$$\begin{aligned}&E\left(sgn(x_j)y_j)\right)=E\left(y_jE\left(sgn(x_j)|y_j\right)\right)\\=&E\left(y_jPr\left(x_j|y_j\geq0\right)-y_jPr\left(x_j|y_j<0\right)\right)\\=&E\left(y_j\left(1-2\Phi\left(\frac{-\rho y_j}{\sqrt{1-\rho^2}}\right)\right)\right)\\=&E\left(y_j\left(2\Phi\left(\frac{\rho y_j}{\sqrt{1-\rho^2}}\right)-1\right)\right)\\=&2\int_{-\infty}^\infty t\phi(t)\Phi\left(\frac{\rho t}{\sqrt{1-\rho^2}}\right)dt\\=&4\int_0^\infty t\phi(t)\Phi\left(\frac{\rho t}{\sqrt{1-\rho^2}}\right)dt-2\int_0^\infty t\phi(t)dt\\=&4\frac{1+\rho}{2}\frac{1}{\sqrt{2\pi}}-2\frac{1}{\sqrt{2\pi}}=\sqrt{\frac{2}{\pi}}\rho,\qquad\textit{result in Lemma 1}\end{aligned}$$

Similarly, using result from Lemma 1

$$\begin{aligned}E\left(sgn(x_j)y_j^3\right)=&4\int_0^\infty t^3\phi(t)\Phi\left(\frac{\rho t}{\sqrt{1-\rho^2}}\right)dt\\&-2\int_0^\infty t^3\phi(t)dt=\frac{1}{\sqrt{2\pi}}\left(6\rho-2\rho^3\right).\square\end{aligned}$$

Theorem 2 leads to a simple estimator $\hat{\rho}_g$ and its variance:

$$\hat{\rho}_g=\frac{1}{k}\sum_{j=1}^k\sqrt{\frac{\pi}{2}}sgn(x_j)y_j,\qquad E\left(\hat{\rho}_g\right)=\rho\tag{9}$$

$$Var\left(\hat{\rho}_g\right)=\frac{V_g}{k},\qquad V_g=\frac{\pi}{2}-\rho^2\tag{10}$$

The variance does not vanish when $|\rho|\to1$. Interestingly, the variance can be substantially reduced by applying a normalization step on y_j, as shown in Theorem 3.

Theorem 3 *As* $k\to\infty$, *the following asymptotic normality holds:*

$$\sqrt{k}\left(\sqrt{\frac{\pi}{2}}\frac{\sum_{j=1}^k sgn(x_j)y_j}{\sqrt{k}\sqrt{\sum_{j=1}^k y_j^2}}-\rho\right)\overset{D}{\Longrightarrow}N\left(0,\ V_{g,n}\right)\tag{11}$$

$$V_{g,n}=V_g-\rho^2\left(3/2-\rho^2\right)\tag{12}$$

where $V_g=\frac{\pi}{2}-\rho^2$ *as in (10).*

Proof Sketch: *Let* $Z_k=\frac{\sum_{j=1}^k sgn(x_j)y_j}{\sqrt{k}\sqrt{\sum_{j=1}^k y_j^2}}$. *As* $k\to\infty$, *then*

$$\frac{1}{k}\sum_{j=1}^k y_j^2\to E\left(y_j^2\right)=1,\ \ a.s.$$

$$Z_k=\frac{\frac{1}{k}\sum_{j=1}^k sgn(x_j)y_j}{\sqrt{\frac{1}{k}k}\sqrt{\frac{1}{k}\sum_{j=1}^k y_j^2}}\to\sqrt{\frac{2}{\pi}}\rho=g,\ \ a.s.$$

We express the deviation $Z_k - g$ as

$$Z_k - g = \frac{\frac{1}{k}\sum_{j=1}^k sgn(x_j)y_j - g + g}{\sqrt{\frac{1}{k}\sum_{j=1}^k y_j^2}} - g$$
$$= \frac{\frac{1}{k}\sum_{j=1}^k sgn(x_j)y_j - g}{\sqrt{\frac{1}{k}\sum_{j=1}^k y_j^2}} + g\frac{1-\sqrt{\frac{1}{k}\sum_{j=1}^k y_j^2}}{\sqrt{\frac{1}{k}\sum_{j=1}^k y_j^2}}$$
$$= \frac{1}{k}\sum_{j=1}^k sgn(x_j)y_j - g + g\frac{1-\frac{1}{k}\sum_{j=1}^k y_j^2}{2} + O_P(1/k)$$

Thus, to analyze the asymptotic variance, it suffices to study:

$$E\left(sgn(x)y - g + g\frac{1-y^2}{2}\right)^2$$
$$= E\left(sgn(x)y - g(1+y^2)/2\right)^2$$
$$= E(y^2) + g^2E(1+y^4+2y^2)/4 - gE(sgn(x)(y+y^3))$$
$$= 1 + g^2(1+3+2)/4 - gE(sgn(x)(y+y^3))$$
$$= 1 + 3/2g^2 - g^2 - gg_3 = 1 - \frac{1}{\pi}\left(5\rho^2 - 2\rho^4\right)$$

where we recall $g_3 = E\left(sgn(x)y^3\right) = \frac{1}{\sqrt{2\pi}}\left(6\rho - 2\rho^3\right)$ □.

Theorem 3 leads to the following estimator $\hat{\rho}_{g,n}$:

$$\hat{\rho}_{g,n} = \sqrt{\frac{\pi}{2}}\left(\frac{\sum_{j=1}^k sgn(x_j)y_j}{\sqrt{k}\sqrt{\sum_{j=1}^k y_j^2}}\right),\ E\left(\hat{\rho}_{g,n}\right) = \rho + O\left(\frac{1}{k}\right),$$
$$Var\left(\hat{\rho}_{g,n}\right) = \frac{V_{g,n}}{k} + O\left(\frac{1}{k^2}\right)$$

This normalization always helps, because $V_{g,n} \leq V_g$.

We can develop more estimators based on Theorem 4.

Theorem 4

$$E\left(y_-1_{x<0} + y_+1_{x\geq 0}\right) = \frac{1+\rho}{\sqrt{2\pi}} \tag{13}$$

$$E\left(y_-1_{x<0} + y_+1_{x\geq 0}\right)^2 \tag{14}$$
$$= 1_{\rho\geq 0} - \frac{1}{\pi}\left(\tan^{-1}\left(\frac{\sqrt{1-\rho^2}}{\rho}\right) - \rho\sqrt{1-\rho^2}\right)$$

$$E\left(y_-1_{x\geq 0} + y_+1_{x<0}\right) = \frac{1-\rho}{\sqrt{2\pi}} \tag{15}$$

$$E\left(y_-1_{x\geq 0} + y_+1_{x<0}\right)^2 \tag{16}$$
$$= 1_{\rho<0} + \frac{1}{\pi}\left(\tan^{-1}\left(\frac{\sqrt{1-\rho^2}}{\rho}\right) - \rho\sqrt{1-\rho^2}\right) \square$$

This leads to another estimator, denoted by $\hat{\rho}_s$:

$$\hat{\rho}_s = 1 - \frac{\sqrt{2\pi}}{k}\sum_{j=1}^k \left[y_{j-}1_{x_j\geq 0} + y_{j+}1_{x_j<0}\right] \tag{17}$$

$$E\left(\hat{\rho}_s\right) = \rho, \qquad Var\left(\hat{\rho}_s\right) = \frac{V_s}{k}$$

$$V_s = 2\pi \times 1_{\rho<0} + 2\tan^{-1}\left(\frac{\sqrt{1-\rho^2}}{\rho}\right) \tag{18}$$
$$- 2\rho\sqrt{1-\rho^2} - (1-\rho)^2$$

After a careful check, the estimator proposed in (Dong, Charikar, and Li 2008) is equivalent to $\hat{\rho}_s$, despite the different expressions. (Dong, Charikar, and Li 2008) used hyper-spherical projection which is equivalent to random projection for high-dimensional original data. The variance expression in (Dong, Charikar, and Li 2008) appeared different but it is indeed the same as the variance of $\hat{\rho}_s$ if we let original data dimension be large.

We can again try to normalize the projected data for the hope of obtaining an improved estimator:

Theorem 5

$$\sqrt{k}\left(\frac{\sum_{j=1}^k y_{j-}1_{x_j\geq 0} + y_{j+}1_{x_j<0}}{\sqrt{k}\sqrt{\sum_{j=1}^k y_j^2}} - \frac{1-\rho}{\sqrt{2\pi}}\right) \tag{19}$$
$$\xrightarrow{D} N\left(0,\ V_{s,n}\right)$$

$$V_{s,n} = V_s - \frac{(1-\rho)^2}{4\pi}\left(1 - 2\rho - 2\rho^2\right) \tag{20}$$

where V_s is in (18). □.

This leads to the following estimator:

$$\hat{\rho}_{s,n} = 1 - \frac{\sum_{j=1}^k \sqrt{2\pi}\left[y_{j-}1_{x_j\geq 0} + y_{j+}1_{x_j<0}\right]}{\sqrt{k}\sqrt{\sum_{j=1}^k y_j^2}} \tag{21}$$

$$E\left(\hat{\rho}_{s,n}\right) = \rho + O\left(\frac{1}{k}\right),\ Var\left(\hat{\rho}_{s,n}\right) = \frac{V_{s,n}}{k} + O\left(\frac{1}{k^2}\right)$$

where $V_{s,n}$ is in (20). The resultant estimator $\hat{\rho}_{s,n}$ has the property that the variance approaches 0 as $\rho \to 1$. The normalization step however does not always help. From (20), we have $V_s \geq V_{s,n}$ if $\rho \leq \frac{\sqrt{3}-1}{2} \approx 0.3660$. On the other hand, as shown in Figure 2, the normalization step only increases the variance slightly if $\rho > 0.3660$.

Figure 2 plots the rations: $\frac{V_m}{V_1}$, $\frac{V_g}{V_1}$, $\frac{V_{g,n}}{V_1}$, $\frac{V_s}{V_1}$, $\frac{V_{s,n}}{V_1}$, to compare those five estimators in terms of their improvements relative to the 1-bit estimator $\hat{\rho}_1$. As expected, the MLE $\hat{\rho}_m$ achieves the smallest asymptotic variance and $\frac{V_m}{V_1} = \frac{2}{\pi}$ at $\rho = 0$ and $\frac{V_m}{V_1} \approx 0.36$ at $|\rho| \to 1$. This means in the high similarity region, using $\hat{\rho}_m$ can roughly reduce the required number of samples (k) by a factor of 3. Overall, $\hat{\rho}_{s,n}$ is computationally simple and its variance is very close to the variance of the MLE, at least for $\rho \geq -0.4$.

We summarize some numerical values in Lemma 2.

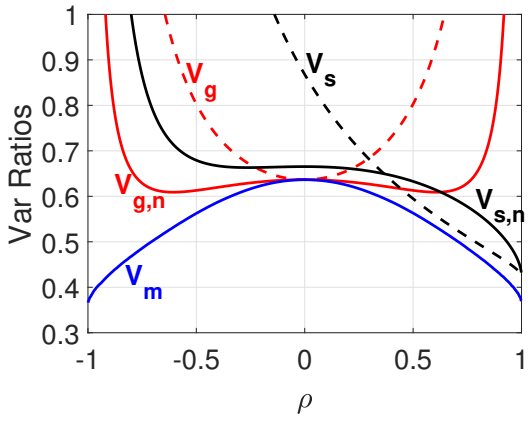

Figure 2: Variance ratios: $\frac{V_m}{V_1}, \frac{V_g}{V_1}, \frac{V_{g,n}}{V_1}, \frac{V_s}{V_1}, \frac{V_{s,n}}{V_1}$, to compare the five estimators, in terms of the relative improvement with respect to the 1-bit estimator $\hat{\rho}_1$. The MLE ($\hat{\rho}_m$, solid blue) achieves the lowest asymptotic variance.

Lemma 2 *At $\rho = 0$,*

$$\frac{V_m}{V_1} = \frac{V_g}{V_1} = \frac{V_{g,n}}{V_1} = \frac{2}{\pi} \approx 0.6366, \quad \frac{V_s}{V_1} = \frac{4}{\pi} - \frac{4}{\pi^2} \approx 0.8680,$$
$$\frac{V_{s,n}}{V_1} = \frac{4}{\pi} - \frac{6}{\pi^2} \approx 0.6653$$

As $|\rho| \to 1$,

$$V_1 = 2\sqrt{2}\pi \left(1 - |\rho|\right)^{3/2} + o\left(\left(1 - |\rho|\right)^{3/2}\right) \tag{22}$$

As $\rho \to 1$,

$$\frac{V_s}{V_1} = \frac{V_{s,n}}{V_1} = \frac{4}{3\pi} \approx 0.4244, \quad \frac{V_g}{V_1} = \infty, \quad \frac{V_{g,n}}{V_1} = \infty \square$$

Recommendation for Estimators

We have studied four estimators (with closed forms) :

$$\hat{\rho}_g = \frac{1}{k} \sum_{j=1}^{k} \sqrt{\frac{\pi}{2}} sgn(x_j) y_j,$$

$$\hat{\rho}_{g,n} = \sqrt{\frac{\pi}{2}} \left(\frac{\sum_{j=1}^{k} sgn(x_j) y_j}{\sqrt{k} \sqrt{\sum_{j=1}^{k} y_j^2}} \right),$$

$$\hat{\rho}_s = 1 - \frac{\sqrt{2\pi}}{k} \sum_{j=1}^{k} \left[y_{j-} 1_{x_j \geq 0} + y_{j+} 1_{x_j < 0} \right],$$

$$\hat{\rho}_{s,n} = 1 - \frac{\sum_{j=1}^{k} \sqrt{2\pi} \left[y_{j-} 1_{x_j \geq 0} + y_{j+} 1_{x_j < 0} \right]}{\sqrt{k} \sqrt{\sum_{j=1}^{k} y_j^2}}$$

The choice depends on application scenarios. Presumably, for a given query, we would like to retrieve data points which have similarity ρ close to 1. However, for practical datasets, typically most data points are not similar at all. From Figure 2, if we hope to use one single estimator, then $\hat{\rho}_{s,n}$ is the overall best. We can also combine two estimators: $\hat{\rho}_s$ and $\hat{\rho}_{g,n}$. Figure 2 shows $\hat{\rho}_s$ is better if $\rho > 0.4437$. For $\rho < 0$, we can always switch to the mirror version of the estimators.

A Simulation Study

We provide a simulation study to verify the theoretical properties of the four estimators for sign-full random projections: $\hat{\rho}_g, \hat{\rho}_{g,n}, \hat{\rho}_s, \hat{\rho}_{s,n}$, as well as $\hat{\rho}_1$ for sign-sign projections.

For a given ρ, we simulate k standard bi-variate normal variables (x_j, y_j) with $E(x_j y_j) = \rho$, $j = 1, ..., k$. Then we choose an estimator $\hat{\rho}$ to estimate ρ. With 10^6 simulations, we plot the empirical mean square errors (MSEs): $MSE(\hat{\rho}) = Bias^2(\hat{\rho}) + Var(\hat{\rho})$, together with the theoretical variance of $\hat{\rho}$. If the empirical MSE curve and the theoretical variance overlap, we know that the estimator is unbiased and the theoretical variance formula is verified.

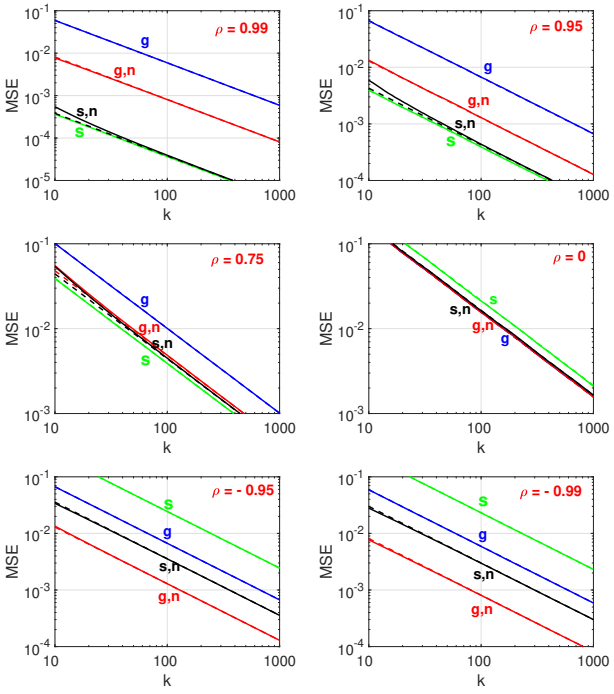

Figure 3: Empirical MSEs (solid curves) for four proposed estimators, together with theoretical (asymptotic) variances (dashed curves), for 6 ρ values. For $\hat{\rho}_g$ and $\hat{\rho}_s$, the solid and dashed curves overlap, confirming that they are unbiased and the variance formulas are correct. For $\hat{\rho}_{g,n}$ and $\hat{\rho}_{s,n}$, the solid and dashed curves overlap when k is not too small.

Figure 3 presents the results for 6 selected ρ values: $0.99, 0.95, 0.75, 0, -0.95, -0.99$. Those simulations verify that both $\hat{\rho}_g$ and $\hat{\rho}_s$ are unbiased, while their normalized versions $\hat{\rho}_{g,n}$ and $\hat{\rho}_{s,n}$ are asymptotically (i.e., when k is not too small) unbiased. The (asymptotic) variance formulas for these four estimators are verified since the solid and dashed curves overlap (when k is not small).

Figure 4 presents the ratios of empirical MSEs (solid curves): $\frac{MSE(\hat{\rho}_1)}{MSE(\hat{\rho}_{s,n})}$ and $\frac{MSE(\hat{\rho}_1)}{MSE(\hat{\rho}_{g,n})}$, together with the theo-

retical asymptotic variance ratios (dashed curves): $\frac{V_1}{V_{s,n}}$ and $\frac{V_1}{V_{g,n}}$. These curves again confirm the asymptotic variance formulas. In addition, they indicate that in the high similarity region, when the sample size k is not too large, the improved gained from using $\hat{\rho}_{s,n}$ can be substantially more than what are predicted by theory. For example, when ρ is close to 1 (e.g., $\rho = 0.99$), theoretically $\frac{V_1}{V_{s,n}} = \frac{3}{4}\pi \approx 2.3562$, the actual improvement can be as much as a factor of 8 (at $k = 10$). This is the additional advantage of $\hat{\rho}_{s,n}$.

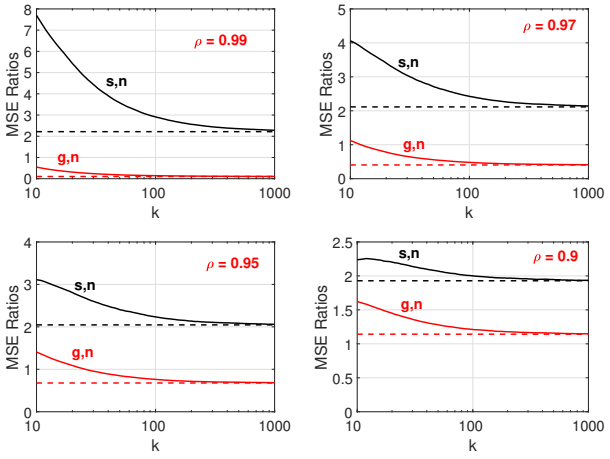

Figure 4: Empirical rations: $\frac{MSE(\hat{\rho}_1)}{MSE(\hat{\rho}_{s,n})}$ and $\frac{MSE(\hat{\rho}_1)}{MSE(\hat{\rho}_{g,n})}$, together with the theoretical asymptotic variance ratios (dashed curves): $\frac{V_1}{V_{s,n}}$ and $\frac{V_1}{V_{g,n}}$. When k is not small, the solid and dashed curves overlap. At high similarity and small k, the improvement would be even more substantial.

An Experimental Study

To further verify the theoretical results, we conduct an experimental study on the ranking task for near-neighbor search on 4 public datasets (see Table 1 and Figure 5).

Table 1: Information about the datasets

Dataset	# Train	# Query	# Dim
MNIST	10,000	10,000	780
RCV1	10,000	10,000	47,236
YoutubeAudio	10,000	11,930	2,000
YoutubeDescription	10,000	11,743	12,183,626

These four datasets are downloaded from either the UCI repository or the LIBSVM website. When a dataset contains significantly more than 10,000 training samples, we only use a random sample of it. The datasets represent a wide range of application scenarios and data types. See Figure 5 for the frequencies of all pairwise ρ values.

For each data point in the query set, we estimate its similarity with every data point in the training set, using random projections. The goal is to return training data points with which the estimated similarities are larger than a prespecified threshold ρ_0. For each query point, we rank all the

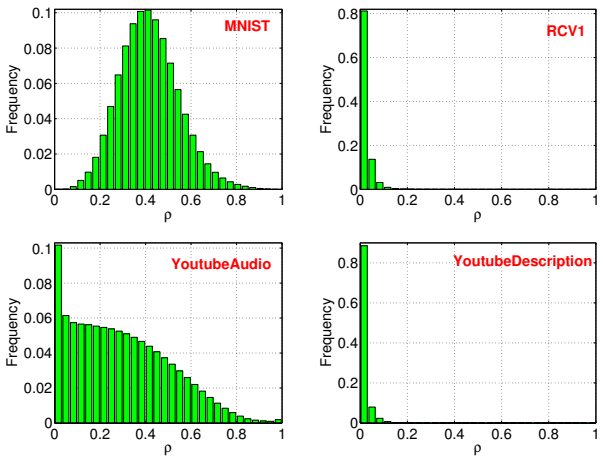

Figure 5: Histograms of all pairwise ρ values.

(estimated) similarities and return top-L points. We can then compute the precision and recall

$$Precision = \frac{\#\text{ retrieved points with true similarities } \geq \rho_0}{L},$$

$$Recall = \frac{\#\text{ retrieved points with true similarities } \geq \rho_0}{\#\text{total points with true similarities } \geq \rho_0}$$

We report the averaged precision-recall values over all query data points. By varying L from 1 to the number of training data points, we obtain a precision-recall curve.

Figure 6 presents the results for the RCV1 datasets, for $\rho_0 \in \{0.9, 0.8, 0.6\}$, and for $k \in \{50, 100\}$. In the first row (i.e., $\rho_0 = 0.9$), we can see that $\hat{\rho}_{s,n}$ is substantially more accurate than both $\hat{\rho}_1$ and $\hat{\rho}_{g,n}$. Since this case represents the high-similarity region, as expected, $\hat{\rho}_{g,n}$ performs poorly. For smaller ρ_0 values, $\hat{\rho}_{g,n}$ performs substantially better, also as expected. Figure 7, Figure 8, and Figure 9 present the results for the other three datasets. The trends are pretty much similar to what we observe in Figure 6.

Conclusion

The method of sign-sign (1-bit) random projections has been a standard tool in practice. In many scenarios such as near-neighbor search and near-neighbor classification, we can store signs of the projected data and discard the original high-dimensional data. As a new data point arrives, one can generate its projected vector and use it (without taking signs) to estimate the similarity. We study various estimators for sign-full random projections and compare their variances with the theoretical limit. Nevertheless, a combination of two estimators ($\hat{\rho}_{g,n}$ and $\hat{\rho}_s$) can almost achieve the theoretical bound. For applications which only allows a single estimator, the proposed $\hat{\rho}_{s,n}$ is the overall best estimator.

References

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. Hoboken, New Jersey: John Wiley & Sons, third edition.

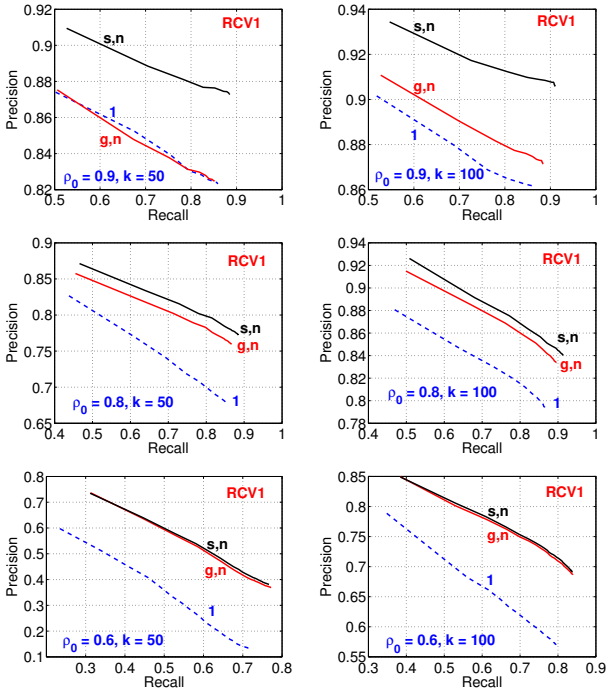

Figure 6: **RCV1**: precision-recall curves for selected ρ_0 and k values, and for three estimators:$\hat{\rho}_{s,n}$, $\hat{\rho}_{g,n}$, $\hat{\rho}_1$.

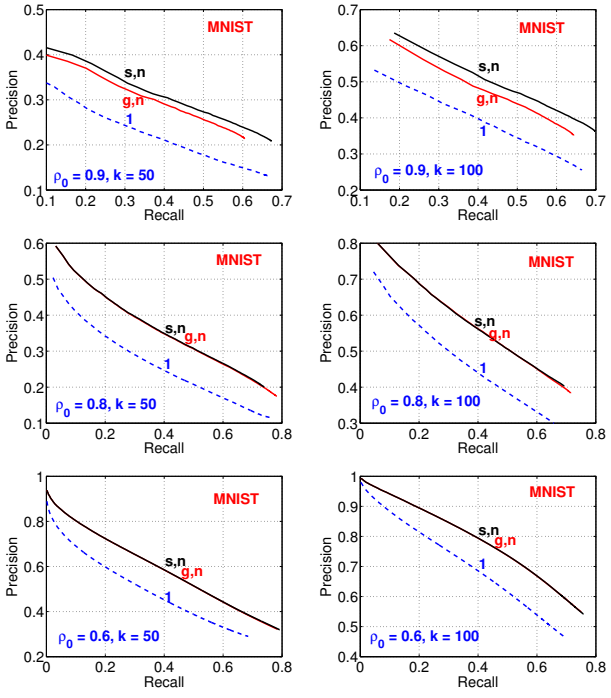

Figure 7: **MNIST**: precision-recall curves for selected ρ_0 and k values, and for three estimators:$\hat{\rho}_{s,n}$, $\hat{\rho}_{g,n}$, $\hat{\rho}_1$.

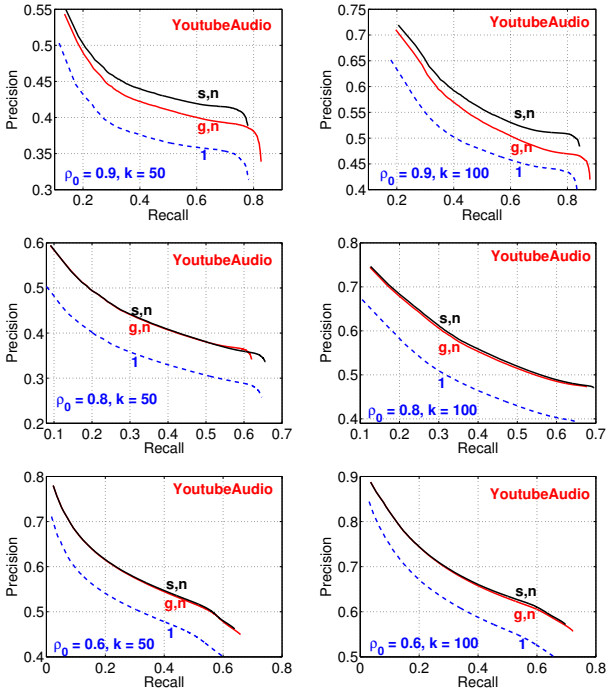

Figure 8: **YoutubeAudio**: precision-recall curves for selected ρ_0 and k values, and three estimators:$\hat{\rho}_{s,n}$, $\hat{\rho}_{g,n}$, $\hat{\rho}_1$.

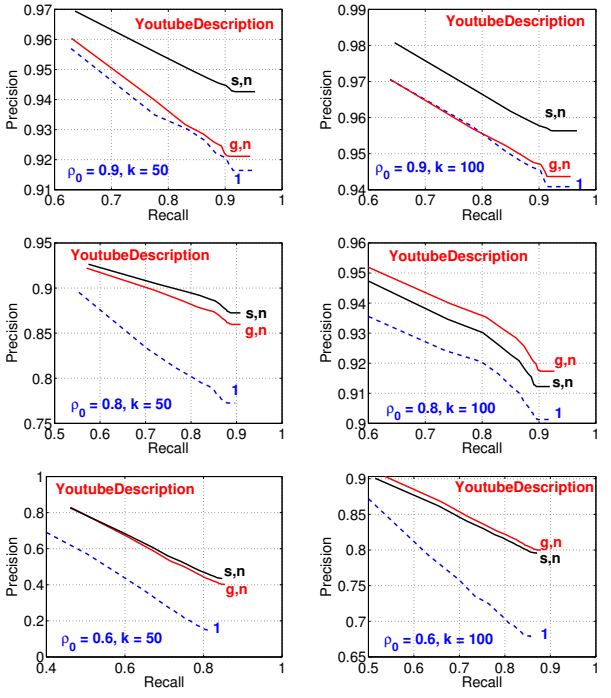

Figure 9: **YoutubeDescription**: precision-recall curves for selected ρ_0 and k values, and three estimators:$\hat{\rho}_{s,n}$, $\hat{\rho}_{g,n}$, $\hat{\rho}_1$.

Broder, A. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences (Sequences)*, 21–29.

Charikar, M. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, 380–388.

Dasgupta, S. 1999. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, 634–644.

Datar, M.; Immorlica, N.; Indyk, P.; and Mirrokni, V. S. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the 20th ACM Symposium on Computational Geometry (SoCG)*, 253–262.

Dong, W.; Charikar, M.; and Li, K. 2008. Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 123–130.

Goemans, M. X., and Williamson, D. P. 1995. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM* 42(6):1115–1145.

Gradshteyn, I. S., and Ryzhik, I. M. 1994. *Table of Integrals, Series, and Products*. New York: Academic Press, fifth edition.

Grimes, C. 2008. Microscale evolution of web pages. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, 1149–1150.

Hajishirzi, H.; Yih, W.-t.; and Kolcz, A. 2010. Adaptive near-duplicate detection via similarity learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, 419–426.

Henzinger, M. R. 2006. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 284–291.

Jégou, H.; Douze, M.; and Schmid, C. 2011. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(1):117–128.

Johnson, W. B., and Lindenstrauss, J. 1984. Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics* 26:189–206.

Lehmann, E. L., and Casella, G. 1998. *Theory of Point Estimation*. New York, NY: Springer, second edition.

Li, P., and König, A. C. 2010. b-bit minwise hashing. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 671–680.

Li, P., and Slawski, M. 2017. Simple strategies for recovering inner products from coarsely quantized random projections. In *Advances in Neural Information Processing Systems (NIPS)*, 4570–4579.

Li, P.; Hastie, T.; and Church, K. W. 2006. Improving random projections using marginal information. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, 635–649.

Li, P.; Samorodnitsky, G.; and Hopcroft, J. E. 2013. Sign cauchy projections and chi-square kernel. In *Advances in Neural Information Processing Systems (NIPS)*, 2571–2579.

Manku, G. S.; Jain, A.; and Sarma, A. D. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, 141–150.

Manzoor, E. A.; Milajerdi, S. M.; and Akoglu, L. 2016. Fast memory-efficient anomaly detection in streaming heterogeneous graphs. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1035–1044.

Papadimitriou, C. H.; Raghavan, P.; Tamaki, H.; and Vempala, S. 1998. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 159–168.

Türe, F.; Elsayed, T.; and Lin, J. J. 2011. No free lunch: brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 943–952.

Vempala, S. 2004. *The Random Projection Method*. Providence, RI: American Mathematical Society.