Control Variates as a Variance Reduction Technique for Random Projections

Keegan Kang[1] and Giles Hooker[1]

Cornell University, Ithaca NY 14850, USA

Abstract. Control variates are used as a variance reduction technique in Monte Carlo integration, making use of positively correlated variables to bring about a reduction of variance for estimated data. By storing the marginal norms of our data, we can use control variates to reduce the variance of random projection estimates. We demonstrate the use of control variates in estimating the Euclidean distance and inner product between pairs of vectors, and give some insight on our control variate correction. Finally, we demonstrate our variance reduction through experiments on synthetic data and the `arcene`, `colon`, `kos`, `nips` datasets. We hope that our work provides a starting point for other control variate techniques in further random projection applications.

1 Introduction

The random projection technique is used in dimension reduction, where data in high dimensions is projected to a lower dimension using a random matrix R. One of the basic applications of this technique is to estimate the Euclidean distance and inner product between pairs of vectors.

The entries r_{ij} in the random matrix R can either be i.i.d. with mean $\mu = 0$ and second moment $\mu_2 = 1$, or correlated with each other. While it is common to have a random projection matrix R with i.i.d. entries $r_{ij} \sim N(0,1)$, speedups are achieved by having R with binary i.i.d. entries [1], or drawn from a sparse Bernoulli distribution [11]. In the above cases, the entries of the random projection matrix consists of elements $\{-1,0,1\}$, thus matrix multiplication is faster when compared to dense entries in $N(0,1)$.

Further speedups can be achieved by using random matrices with correlated entries, such as matrices constructed by the Lean Walsh Transform [12] to the Fast Johnson Lindenstrauss Transform (FJLT) [2] and the Subsampled Randomized Hadamard Transform (SRHT) [4]. Both these transformations make use of matrix-vector products using the Hadamard matrix, which can be computed recursively.

Consider vectors $\mathbf{x}_i \in \mathbb{R}^p$ mapped to a lower dimensional vector $\tilde{\mathbf{x}}_i \in \mathbb{R}^k$ using a random projection matrix R under the identity $\tilde{\mathbf{x}}^T = \mathbf{x}^T R$. The distance properties of these vectors $\mathbf{x}_i, \mathbf{x}_j$ are preserved in expectations in $\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j$. If we wanted to compute a property of $\mathbf{x}_i, \mathbf{x}_j$ given by some $f(\mathbf{x}_i, \mathbf{x}_j)$, then the goal is to find some function $g(\cdot)$, such that $\mathbb{E}[g(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)] = f(\mathbf{x}_i, \mathbf{x}_j)$. For example, if we want an estimate of the Euclidean distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, using a random projection matrix R with entries i.i.d. from $N(0,1)$, or from $\{-1,1\}$ with equal probability, then $f(a,b) = g(a,b) = \|a-b\|_2$.

Each of these resultant estimates from a chosen random matrix R have probability bounds on accuracy plus bounds on their run time, and it is up to the user to choose a random projection matrix which will suit their purposes.

In this article, we expand upon the conference proceedings *Random Projections with Control Variates* [8] in the following ways

1. We give more insight on why control variates reduce the variance of our random projection estimates.
2. We give more intuition on the control variate correction.
3. We perform more experiments on more datasets.

The control variate approach with random projections can be used with different types of different random projection matrices. This leads to a variance reduction in the estimation of Euclidean distances and inner products between pairs of vectors $\mathbf{x}_i, \mathbf{x}_j$ with a negligible extra cost in speed and storage space. Such measures of distances can be used in clustering [6], [5], classification [15], and set resemblance problems [10].

We structure this article as follows: First, we express our notation differently from the ordinary random projection notation to give intuition on how we can use control variates. Next, we explain the control variate technique of variance reduction and show control variates achieve variance reduction. We next look at related work which inspired our method, before introducing the control variate corrections for Euclidean distances and inner products. Lastly, we demonstrate our method on both synthetic and experimental data and show that we can use a control variate approach together with any random projection method to gain variance reduction in our estimates.

1.1 Notation and Intuition

With classical random projections, we denote $R \in \mathbb{R}^{p \times k}$ to be a random projection matrix. We let $X \in \mathbb{R}^{n \times p}$ to be our data matrix, where each row $\mathbf{x}_i^T \in \mathbb{R}^p$ is a p dimensional observation. In most textbooks, the random projection equation is given by

$$V = \frac{1}{\sqrt{k}} XR \tag{1}$$

However, we will use

$$V = XR \tag{2}$$

without the scaling factor. The motivation is to see each element $v_{ij} \in V$ as a random variable drawn from some probability distribution.

Consider the random matrix R written as

$$R = [\mathbf{r}_1 \mid \mathbf{r}_2 \mid \dots \mid \mathbf{r}_k] \tag{3}$$

where each $\mathbf{r}_i$ is a column vector with i.i.d. entries. Then for a *fixed* row $\mathbf{x}_i^T$, the elements $\{v_{ij}^2\}_{j=1}^k$ are drawn from the same probability distribution with mean $\|\mathbf{x}_i\|_2^2$.

By the Law of Large Numbers, we would expect that as k increases, the mean of the observations $\{v_{ij}^2\}_{j=1}^k$ would converge to the true value of $\|\mathbf{x}_i\|_2^2$.

Similarly, we have the means of $\{(v_{is}-v_{js})^2\}_{s=1}^k$ and $\{(v_{is}v_{js})\}_{s=1}^k$ converging to $\|\mathbf{x}_i-\mathbf{x}_j\|_2^2$ and $\langle \mathbf{x}_i, \mathbf{x}_j\rangle$ respectively.

For other random projection matrices R where the entries come from a different distribution, we can also find equivalent expressions of the form $\{f(\mathbf{v}_{is}, \mathbf{v}_{js})\}_{s=1}^k$ where the mean of these observations converge to either the squared Euclidean distance or inner product.

1.2 Probability Bounds on Random Projection Estimates

We give the form of the probability bounds on random projection estimates in order to show how control variates give us a tighter bound.

Suppose we look at a single row $\mathbf{v_i} \in V$, and let $S_k^{\text{norm}} = \sum_{s=1}^k v_{is}^2$. By finding expressions of the form $f_1(\varepsilon, k_1), f_2(\varepsilon, k_2)$ where

$$\mathbb{P}\left[\frac{S_k^{\text{norm}}}{k} \geq (1+\varepsilon)\|\mathbf{x}\|_2^2\right] \leq f_1(\varepsilon, k_1) \tag{4}$$

$$\mathbb{P}\left[\frac{S_k^{\text{norm}}}{k} \leq (1-\varepsilon)\|\mathbf{x}\|_2^2\right] \leq f_2(\varepsilon, k_2) \tag{5}$$

we can then place bounds on how far our estimate of the norm is relative to our actual value since we have

$$\mathbb{P}\left[(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \frac{S_k^{\text{norm}}}{k} \leq (1+\varepsilon)\|\mathbf{x}\|_2^2\right] \leq 1 - f_1(\varepsilon, k_1) - f_2(\varepsilon, k_2) \tag{6}$$

Furthermore, computing these expressions $f_1(\varepsilon, k_1), f_2(\varepsilon, k_2)$ suffices to place probability bounds on our estimate of Euclidean distances and inner products. Similarly, by defining $S_k^{\text{ED}} := \sum_{s=1}^k (v_{is}-v_{js})^2$ and $S_k^{\text{IP}} := \sum_{s=1}^k v_{is}v_{js}$, we can thus write

$$\mathbb{P}\left[(1-\varepsilon)\|\mathbf{x}_i-\mathbf{x}_j\|_2^2 \leq \frac{S_k^{\text{ED}}}{k} \leq (1+\varepsilon)\|\mathbf{x}_i-\mathbf{x}_j\|_2^2\right] \leq 1 - f_1(\varepsilon, k_1) - f_2(\varepsilon, k_2) \tag{7}$$

$$\mathbb{P}\left[(1-\varepsilon)\langle \mathbf{x}_i, \mathbf{x}_j\rangle \leq \frac{S_k^{\text{IP}}}{k} \leq (1+\varepsilon)\langle \mathbf{x}_i, \mathbf{x}_j\rangle\right] \leq 1 - 2f_1(\varepsilon, k_1) - 2f_2(\varepsilon, k_2) \tag{8}$$

To see this, we can replace $v_{is}^2, \mathbf{x}$ in Equations 4 and 5 by $(v_{is}-v_{js})^2, \mathbf{x}_i - \mathbf{x}_j$ and get the bounds in (7).

We can also use the identities

$$\|\mathbf{v}_i - \mathbf{v}_j\|_2^2 = \|\mathbf{v}_i\|_2^2 + \|\mathbf{v}_j\|_2^2 - 2\langle \mathbf{v}_i, \mathbf{v}_j\rangle \tag{9}$$

$$\|\mathbf{v}_i + \mathbf{v}_j\|_2^2 = \|\mathbf{v}_i\|_2^2 + \|\mathbf{v}_j\|_2^2 + 2\langle \mathbf{v}_i, \mathbf{v}_j\rangle \tag{10}$$

and the union bound to show (8), by expressing $\langle \mathbf{v}_i, \mathbf{v}_j\rangle$ in terms of $\langle \mathbf{x}_i, \mathbf{x}_j\rangle$.

When r_{ij} are i.i.d. $N(0,1)$, we get $f_1(\varepsilon, k_1) = f_2(\varepsilon, k_2) = \exp\left\{-\frac{k(\varepsilon^2-\varepsilon^3)}{4}\right\}$ [18]. The bounds for other random projection matrices can be found in [2, 4, 18].

These probability bounds are usually computed by looking at the respective second moments, and using a Chernoff bound approach.

1.3 Control Variates

Having introduced the notion of each v_{ij} as a random variable, we now look at control variates. Control variates are a technique in Monte Carlo simulation using random variables for variance reduction. A more thorough explanation can be found in Ross, 2006.

The method of control variates assumes we use the same random inputs to estimate $\mathbb{E}[A] = \mu_A$, for which we know B with $\mathbb{E}[B] = \mu_B$. We call B our control variate. Then to estimate $\mathbb{E}[A] = \mu_A$ from some distribution A, we can instead compute the expectation of

$$\mathbb{E}[A + c(B - \mu_B)] \quad = \quad \mathbb{E}[A] + c\mathbb{E}[B - \mu_B] \quad = \quad \mu_A \tag{11}$$

which is an unbiased estimator of μ_A for some constant c, which is our control variate correction. This value of c which minimizes the variance is given by

$$\hat{c} = -\frac{\mathrm{Cov}(A,B)}{\mathrm{Var}(B)} \tag{12}$$

and thus we write

$$\mathrm{Var}[A + c(B - \mu_B)] = \mathrm{Var}(A) - \frac{(\mathrm{Cov}(A,B))^2}{\mathrm{Var}(B)} \tag{13}$$

Suppose we look at

$$S_k^{\text{norm}} = \sum_{j=1}^{n} v_{ij}^2 \tag{14}$$

We can think of A being the probability distribution of the v_{is}, with the mean $\mu_A = \|\mathbf{x}_i\|_2^2$ being our target. If we found some probability distribution B of the form $f(v_{ij})$ and known mean μ', then we must have

$$S_k^{\text{cvnorm}} = \sum_{j=1}^{n} v_{ij}^2 + c(f(v_{ij}) - \mu') \tag{15}$$

The expected value of S_k^{cvnorm} is still $\|\mathbf{x}\|_2^2$, but the second moment of S_k^{cvnorm} has to be lower (or no worse) than S_k^{norm}. To see this, recall that the variance of a distribution is given by

$$\mathrm{Var}[A] = \mathbb{E}[A^2] - (\mathbb{E}[A])^2 \tag{16}$$

where $\mathbb{E}[A^2]$ is the second moment. Therefore, we can take the variance of our expression as a proxy for the second moment, and by (13), have our second moment of S_k^{cvnorm} to be no greater than S_k^{norm}, since the term $\frac{(\mathrm{Cov}(A,B))^2}{\mathrm{Var}(B)}$ is always positive.

Therefore, we need to find some distribution B where the variables b_i are correlated with v_{ij} to get good variance reduction. If they were independent, then the numerator in our term $\frac{(\mathrm{Cov}(A,B))^2}{\mathrm{Var}(B)}$ becomes zero, and we do not get any variance reduction at all.

To find such a distribution B, we necessarily need to fulfill two conditions.

Condition 1: Since each realization v_{ij} is the sum of p random variables $r_{1j},\dots,r_{pj}$, we need to have y_i constructed from these same random variables *and* also correlated with each $x_{i1},\dots,x_{ip}$ in order to get a variance reduction.

Condition 2: We need to know the actual value of μ_B, the mean of B.

This seems like a chicken and egg problem since any μ_B that is related to both $x_{i.}$, $r_{.j}$ would be of some form of either the Euclidean distance or the inner product, both of which we want to estimate in the first place. We solve this problem by considering an expression that relates both the Euclidean distance and the inner product simultaneously.

1.4 Related Work

We draw inspiration from the works of Li, Hastie, and Church [9–11]. In these papers, Li *et al* expressed the tuple (v_{is}, v_{js}) coming from a bivariate normal when the entries of the random projection matrix R is i.i.d. $N(0,1)$.

More formally, given the matrix $V = XR$ where each $r_{ij} \sim N(0,1)$, then for any two rows $\mathbf{v}_i, \mathbf{v}_j$ of V we have the tuple

$$\begin{pmatrix} v_{is} \\ v_{js} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} m_i & a \\ a & m_j \end{pmatrix}\right) \tag{17}$$

where m_i, m_j denote the norms $\|\mathbf{x}_i\|_2^2$, $\|\mathbf{x}_j\|_2^2$ respectively, and a denotes the inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$.

Li et al showed that if marginal information such as the actual norms $\|\mathbf{x}_i\|_2^2$, $\|\mathbf{x}_j\|_2^2$ were precomputed and stored, then it is possible to get a more accurate estimate of the inner product a using an asymptotic maximum likelihood estimator.

To do this, Li et al computed the log-likelihood function after observing k such draws $\{v_{is}, v_{js}\}_{s=1}^k$ which is given by

$$l(a) \propto -\frac{1}{2}\log(m_1 m_2 - a^2) - \frac{1}{2}\frac{1}{m_1 m_2 - a^2}\sum_{s=1}^{k}(v_{is}^2 m_j - 2 v_{is} v_{js} a + v_{js}^2 m_i) \tag{18}$$

and found the value of $\hat{a}$ which maximizes this function via root finding techniques.

Li et al also showed that the above result also held asymptotically when the entries of the random matrix R do not come from $N(0,1)$. If they were i.i.d. from the Sparse Bernoulli distribution, then under the Central Limit Theorem, the tuple (v_{is}, v_{js}) also converges to the bivariate normal.

We will use these results below.

1.5 Our Contributions

We propose using control variates in this article to reduce the variance of the estimates of the Euclidean distances and the inner products between pairs of vectors for a choice of random projection matrix R. In particular

1. We describe the process of the control variate approach, which has the same time complexity to a non control variate approach.
2. We give the first and second moments of $A + c(B - \mu_B)$ for matrices R with i.i.d. entries, which can then be used to bound the errors in our estimates.
3. We demonstrate empirically that our control variate approach works well with current random projection methods on synthetically generated data and the `arcene`, `colon`, `kos`, `nips` datasets.

2 Process Of Using Control Variates

We describe and illustrate the process of using control variates in this section.

Without loss of generality, suppose we had $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$. Consider $\mathbf{v}$ given by $X\mathbf{r}$. For the case $p = 2$, we would have

$$V = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = X\mathbf{r} \tag{19}$$

for one column of R. We do matrix multiplication $X\mathbf{r}$ and get v_1, v_2.

In the next two sections, we will give the control variate to estimate the Euclidean distance and the inner product. We will also give the respective optimal control variate correction c, and the respective first and second moments of the expression $A + c(B - \mu_B)$. This allows us to compute a more accurate estimate for the Euclidean distance and the inner product, as well as place probability bounds on the errors of our estimates.

2.1 Control Variate for the Euclidean distance

Suppose we computed V as above. The following theorem shows us how to estimate the Euclidean distance with our control variate.

Theorem 1. *Let one realization of $A = (v_1 - v_2)^2$, which is an estimate of our Euclidean distance. Let one realization of B to be $(v_1 - v_2)^2 + 2v_1v_2 = v_1^2 + v_2^2$ with expected value $\mu_B = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|_2^2$. The Euclidean distance (in expectation) between these two vectors is given by $\mathbb{E}[A + c(B - \mu_B)]$, and we can compute $c := -Cov(A,B)/Var(B)$ from our matrix V directly, using the empirical covariance $Cov(A,B)$ and empirical variance $Var(B)$.*

Proof. We have

$$\begin{aligned} &\mathbb{E}[(v_1 - v_2)^2] + 2\mathbb{E}[v_1v_2] \\ &= \|\mathbf{x}_1 - \mathbf{x}_2\|^2 + 2\langle \mathbf{x}_1 \mathbf{x}_2 \rangle && (20) \\ &= \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 - 2\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + 2\langle \mathbf{x}_1, \mathbf{x}_2 \rangle && (21) \\ &= \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 && (22) \end{aligned}$$

We derive the following lemma to help us compute the first and second moments required.

Lemma 1. Suppose we assume that our matrix R has i.i.d. entries, where each r_{ij} has mean $\mu = 0$, second moment $\mu_2 = 1$, and fourth moment μ_4. Then

$$\mathbb{E}[A^2] = \quad \mu_4 \sum_{j=1}^{p} (x_{1j} - x_{2j})^4 + 6 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} - x_{2u})^2 (x_{1v} - x_{2v})^2 \tag{23}$$

$$\begin{aligned}\mathbb{E}[B^2] &= \mu_4 \sum_{j=1}^{p} (x_{1j}^4 + x_{2j}^4) + 6 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u}^2 x_{1v}^2 + x_{2u}^2 x_{2v}^2) \\ &\quad + 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} x_{1v} x_{2u} x_{2v}) + \mu_4 \sum_{j=1}^{p} x_{1j}^2 x_{2j}^2 + \sum_{i \neq j}^{p} x_{1i}^2 y_{2j}^p \qquad (24) \\ \mathbb{E}[AB] &= 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} - x_{2u})(x_{1v} - x_{2v})(x_{1u} x_{1v} + x_{2u} x_{2v}) \\ &\quad + \mu_4 \sum_{j=1}^{p} (x_{1j} - x_{2j})^2 (x_{1j}^2 + x_{2j}^2) + \sum_{i \neq j} (x_{1i} - x_{2i})^2 (x_{1i}^2 + x_{2j}^2) \qquad (25)\end{aligned}$$

Proof. We repeatedly apply Lemma 2 in the Appendix.

Thus, by following Lemma 1, we are able to derive expressions for the optimal control variate correction c in our procedure as follows.

Theorem 2. *The optimal value c is given by*

$$c = -\frac{Cov(A,B)}{Var[B]} \tag{26}$$

where we have

$$Cov(A,B) = \mathbb{E}[AB - A\mu_B - B\mu_A + \mu_A \mu_B] \tag{27}$$

and

$$Var[B] = \mathbb{E}[B^2] - (\mathbb{E}[B])^2 \tag{28}$$

They expand to

$$\begin{aligned}Cov(A,B) &= 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} - x_{2u})(x_{1v} - x_{2v})(x_{1u} x_{1v} + x_{2u} x_{2v}) \\ &\quad + (\mu_4 - 1) \sum_{j=1}^{p} (x_{1j} - x_{2j})^2 (x_{1j}^2 + x_{2j}^2) \qquad (29)\end{aligned}$$

and

$$\begin{aligned}Var[B] &= (\mu_4 - 1) \sum_{j=1}^{p} (x_{1j}^4 + x_{2j}^4) + 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u}^2 x_{1v}^2 \\ &\quad + x_{2u}^2 x_{2v}^2) + 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} x_{1u} x_{1v} x_{2u} x_{2v} + (\mu_4 - 2) \sum_{j=1}^{p} x_{1j}^2 x_{2j}^2 - \sum_{i \neq j} x_{1i}^2 x_{2j}^2 \qquad (30)\end{aligned}$$

We are also able to derive the first and second moments of $A + c(B - \mu_B)$ for Euclidean distances.

Theorem 3. *The first and second moments are*

$$\mathbb{E}[A + c(B - \mu_B)] = \mathbb{E}[A] + c\mathbb{E}[B - \mu_B] = 0 \tag{31}$$

and

$$\mathbb{E}[(A + c(B - \mu_B))^2] = \mathbb{E}[A^2 + 2cAB - 2c\mu_B A + c^2B^2 - 2c^2\mu_B B + c^2\mu_B^2] \tag{32}$$

where we substitute in the values of $\mathbb{E}[A^2], \mathbb{E}[AB], \mathbb{E}[B^2]$ *from Lemma 1.*

These first and second moments could be used to get tighter (and exact) bounds of the form in (4) and (5) by using a Chernoff bound type strategy.

2.2 Control Variate for the inner product

Suppose we computed V as above. The following theorem shows us how to estimate the inner product with our control variate.

Theorem 4. *Let one realization of* $A = v_1v_2$*, which is an estimate of our inner product. Let one realization of* B *to be* $(v_1 - v_2)^2 + 2v_1v_2 = v_1^2 + v_2^2$ *with expected value* $\mu_B = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|_2^2$*. The inner product between these two vectors is given by* $\mathbb{E}[A + c(B - \mu_B)]$*, and we can compute* $c := -Cov(A,B)/Var(B)$ *from our matrix* V *directly, using the empirical covariance* $Cov(A,B)$ *and empirical variance* $Var(B)$*.*

The optimal control variate c in this procedure is given by the next theorem.

Theorem 5. *The optimal value of* c *is given by*

$$c = -\frac{Cov(A,B)}{Var[B]} \tag{33}$$

where

$$\begin{aligned} Cov(A,B) &= \mathbb{E}[AB - A\mu_B - B\mu_A + \mu_A\mu_B] \\ &= (\mu_4 - 1)\sum_{j=1}^{p} x_{1j}x_{2j}(x_{1j}^2 + x_{2j}^2) + \sum_{i \neq j} x_{1i}x_{2j}(x_{1i}x_{1j} + x_{2i}x_{2j}) \end{aligned} \tag{34}$$

and the value of $Var[B]$ *taken from the result in Theorem 2.*

2.3 The optimal control variate correction c

While we have computed an expression c in terms of the first and second moments of our distributions, they are not at all intuitive from first sight. Therefore, we consider what the optimal value of c would be if the random matrix R had i.i.d. entries $r_{ij} \sim N(0,1)$. Thus, we take a second look at the bivariate normal distribution in (17).

Theorem 6. *For $r_{ij} \sim N(0,1)$, and $V = XR$, the optimal control variate correction c_{ED} for the Euclidean distance is given by*

$$c_{ED} = -\frac{(m_i - a)^2 + (m_j - a)^2}{(m_i^2 + m_j^2 + 2a^2)} \tag{35}$$

We use m_i, m_j to denote the norms $\|\mathbf{x}_i\|_2^2, \|\mathbf{x}_j\|_2^2$ respectively, and a to be $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ as in (17).

Proof. We can write the control variate correction c for the Euclidean distance as

$$c_{\mathrm{ED}} = -\frac{\mathrm{Cov}(v_i^2 + v_j^2, v_i^2 + v_j^2 - 2v_iv_j)}{\mathrm{Var}\left(v_i^2 + v_j^2\right)} \tag{36}$$

$$= -\frac{\mathrm{Cov}(\mathbf{v}_i^T\mathbf{v}_j, \mathbf{v}_i{}^T H\mathbf{v}_j)}{\mathrm{Var}\left(\mathbf{v}_i^T\mathbf{v}_j\right)} \tag{37}$$

where $H = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. Next, expanding the numerator gives

$$\mathrm{Cov}(\mathbf{v}_i^T\mathbf{v}_j, \mathbf{v}_i{}^T H\mathbf{v}_j) = \mathbb{E}[\mathbf{v}_i^T\mathbf{v}_j\mathbf{v}_i^T H\mathbf{v}_j] - \mathbb{E}\left[\mathbf{v}_i^T\mathbf{v}_j\right]\mathbb{E}\left[\mathbf{v}_i^T H\mathbf{v}_j\right] \tag{38}$$

For $(v_i, v_j) \sim N(0, \Sigma)$, we have the identities

$$\mathbb{E}\left[\mathbf{v}_i^T\mathbf{v}_j\mathbf{v}_i^T H\mathbf{v}_j\right] = \mathrm{Tr}(\Sigma(H + H^T)\Sigma) + \mathrm{Tr}(\Sigma)\mathrm{Tr}(H\Sigma) \tag{39}$$

$$= 2(m_i - a)^2 + 2(m_j - a)^2 + (m_i + m_j)(m_i + m_j - 2a) \tag{40}$$

$$\mathbb{E}[\mathbf{v}_i^T\mathbf{v}_j] = m_i + m_i \tag{41}$$

$$\mathbb{E}[\mathbf{v}_i^T H\mathbf{v}_j] = m_j + m_j - 2a \tag{42}$$

and therefore, we have

$$\mathrm{Cov}(\mathbf{v}_i^T\mathbf{v}_j, \mathbf{v}_i{}^T H\mathbf{v}_j) = 2(m_i - a)^2 + 2(m_j - a)^2 \tag{43}$$

The denominator expands to be

$$\mathrm{Var}\left(\mathbf{v}_i^T\mathbf{v}_j\right) = \mathrm{Tr}\left(\Sigma(2I)\Sigma\right) \tag{44}$$

$$= 2(m_i^2 + m_j^2 + 2a^2) \tag{45}$$

Simplifying, we get:

$$c_{\mathrm{ED}} = -\frac{(m_i - a)^2 + (m_j - a)^2}{(m_i^2 + m_j^2 + 2a^2)} \tag{46}$$

Theorem 7. *For $r_{ij} \sim N(0,1)$, and $V = XR$, the optimal control variate correction c_{IP} for the inner product is given by*

$$c_{IP} = -\frac{m_i a + m_j a}{m_i^2 + m_j^2 + 2a^2} \tag{47}$$

Proof. Analogous to the proof of Theorem 6, we express

$$\mathrm{Cov}(v_i^2 + v_j^2, v_i v_j) = \mathrm{Cov}\left(\mathbf{v}_i^2 \mathbf{v}_j, \mathbf{v}_i^T H \mathbf{v}_j\right) \tag{48}$$

where $H = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Therefore, we similarly compute

$$\begin{aligned}
\mathbb{E}\left[\mathbf{v}_i^T \mathbf{v}_j \mathbf{v}_i^T H \mathbf{v}_j\right] &= \mathrm{Tr}(\Sigma(H + H^T)\Sigma) + \mathrm{Tr}(\Sigma)\mathrm{Tr}(H\Sigma) && (49)\\
&= 2m_i a + 2m_j a + (m_i + m_j)(a) && (50)\\
\mathbb{E}[\mathbf{v}_i^T \mathbf{v}_j] &= m_i + m_j && (51)\\
\mathbb{E}[\mathbf{v}_i^T H \mathbf{v}_j] &= a && (52)
\end{aligned}$$

which results in

$$c_{\mathrm{IP}} = \frac{m_i a + m_j a}{m_i^2 + m_j^2 + 2a^2} \tag{53}$$

Without loss of generality, we assume that our data is normalized such that $m_i = m_j = 1$. In this case, we can compute the variance reduction for our Euclidean distances and inner products respectively.

Theorem 8. *Given $r_{ij} \sim N(0,1)$ and $V = XR$, then for any pair $\mathbf{x}_i, \mathbf{x}_j$*

1. *The variance of the estimate of the Euclidean distance between the pair is given by*

$$\sigma_{ED} = 8(1-a)^2 \tag{54}$$

2. *The variance of the estimate of the Euclidean distance with the control variate correction between the pair is given by*

$$\sigma_{EDCV} = 8(1-a)^2 - \frac{4(1-a)^4}{(1+a^2)} \tag{55}$$

3. *The variance of the estimate of the inner product between the pair is given by*

$$\sigma_{IP} = 1 + a^2 \tag{56}$$

4. *The variance of the estimate of the inner product with the control variate correction between the pair is given by*

$$\sigma_{IPCV} = 1 + a^2 - \frac{4a^2}{1+a^2} \tag{57}$$

Proof. This follows from direct substitution of the optimal control variate corrections in Theorems 6 and 7 into (13).

Theorem 8 allows us to analyze the effect of our control variate correction, given in Figure 1.

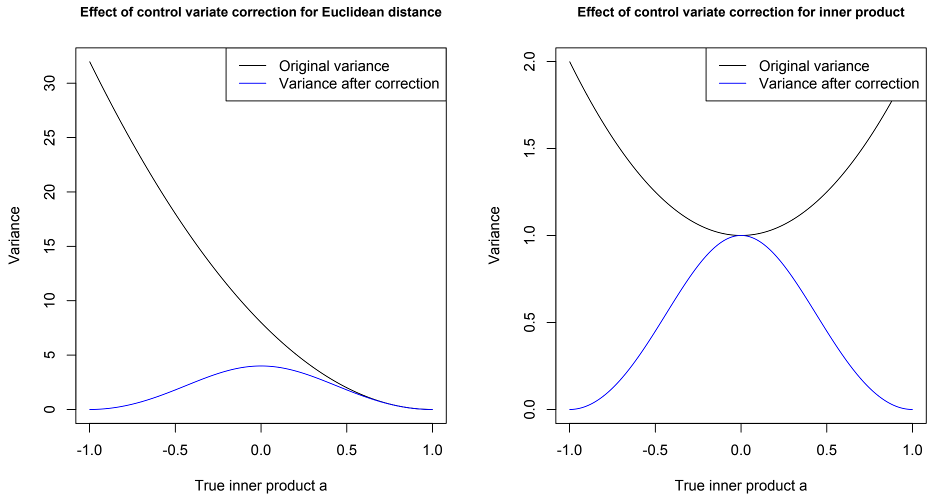

Fig. 1. Effects of control variate correction on estimates

Recall that when observations are normalized, we have the Euclidean distance between any pair of vectors being in the range $[0,2]$, and the inner product between any pair of vectors being in the range $[0,1]$.

We can now analyze the effect of our control variate correction on Euclidean distance. For vectors $\mathbf{x}_i, \mathbf{x}_j$ in the same direction (inner product close to 1), we do not get much variance reduction in our estimate of Euclidean distance. Conversely, if the vectors were in opposite directions, then we would get a reasonable variance reduction in the estimates of their Euclidean distance.

Similarly, we analyze the effect of our control variate correction on the inner product. If the vectors $\mathbf{x}_i, \mathbf{x}_j$ are orthogonal to each other (inner product near 0), then we do not get much variance reduction from our control variate correction. Conversely, if the vectors share the same or opposite directions, then we would get a reasonable variance reduction in the estimates of their inner product.

Theorems 6, 7, and Li ıet al’s result that the tuple (v_{is}, v_{js}) converges to a bivariate normal even if r_{ij} from R do not come from $N(0,1)$ suggests an alternative method of computing the control variate correction.

Instead of computing the control variate correction c_{ED} empirically from our data, we could choose to either compute the vanilla estimate $\hat{a} = v_{is}v_{js}$ for the inner product, or $\hat{a}$ using Li’s method. We can then substitute $\hat{a}$ into the results of Theorem 6 to compute the optimal control variate correction, using the fact that we get convergence to a bivariate normal when the number of observations increase. In fact, since we are storing the marginal norms, we can compute $\hat{a}$ via Li’s method, and use this to compute c_{ED} directly since this does not increase the time complexity.

Similarly, we could compute the control variate correction c_{IP}, but only using the ordinary estimate $\hat{a} = v_{is}v_{js}$ for the inner product.

2.4 Overall computational time

Constructing the matrix $V = X_{n\times p}R_{p\times k}$ takes $O(npk)$ time, and computing the pairwise Euclidean distances (or inner products) takes an additional $O(n^2k)$ of time.

If we want to use control variates, we either need to compute the empirical covariance between all pairs A and B and the variance of B, or compute an estimate of the inner product a to put into our control variate correction. Both these options take an additional $O(nk)$ time for all pairwise computations. We also need to compute and store the norm of each vector, which takes $O(np)$ time.

Thus, the overall computational time is given by $O(npk+n^2k+n(k+p)) = O(npk+n^2k)$.

3 Our Experiments

Throughout our experiments, we use five different types of random projection matrices as shown in Table 1 (also used in [8]). We pick these five types of random projection matrices as they are commonly used random projection matrices.

We use $N(0,1)$ to denote the Normal distribution with mean $\mu = 0$ and $\sigma^2 = 1$. We denote $(\mathbf{1})_p$ to be the length p vector with all entries being 1, and $(\mathbf{0})_p$ to be the length p vector with all entries being 0. We denote the baseline estimates to be the respective estimates given by using the type of random projection matrix R_i.

We run our simulations for 10000 iterations for every experiment.

Table 1. Random projection matrices.

R	Type
R_1	Entries i.i.d. from $N(0,1)$
R_2	Entries i.i.d. from $\{-1,1\}$ with equal probability
R_3	Entries i.i.d. from $\{-\sqrt{p},0,\sqrt{p}\}$ with probabilities $\{\frac{1}{2p},1-\frac{1}{p},\frac{1}{2p}\}$ for $p=5$
R_4	Entries i.i.d. from $\{-\sqrt{p},0,\sqrt{p}\}$ with probabilities $\{\frac{1}{2p},1-\frac{1}{p},\frac{1}{2p}\}$ for $p=10$
R_5	Constructed using the Subsampled Randomized Hadamard Transform (SRHT)

Table 2. Generated data $\mathbf{x}_1$, $\mathbf{x}_2$.

Pairs	$\mathbf{x}_1$	$\mathbf{x}_2$
Pair 1	Entries i.i.d. from $N(0,1)$	Entries i.i.d. from $N(0,1)$
Pair 2	Entries i.i.d. from standard Cauchy	Entries i.i.d. from standard Cauchy
Pair 3	Entries i.i.d. from Bernoulli(0.05)	Entries i.i.d. from Bernoulli(0.05)
Pair 4	Vector $[(\mathbf{1})_{p/2},(\mathbf{0})_{p/2}]$	Vector $[(\mathbf{0})_{p/2},(\mathbf{1})_{p/2}]$

3.1 Generating vectors from synthetic data

We first perform our experiments on a wide range of synthetic data. We look at normalized pairs of vectors $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathbb{R}^{5000}$ generated from the following distributions in Table 2 (also used in [8]). In short, we look at data that can be Normal, heavy tailed (Cauchy), sparse (Bernoulli), and an adversarial scenario where the inner product is zero.

We first compare the relative bias of the estimates of Euclidean distance using our control variate approach against the baseline estimates for each projection R_i for a sanity check. Plots can be seen in Figure 2, and the relative bias goes to zero as expected.

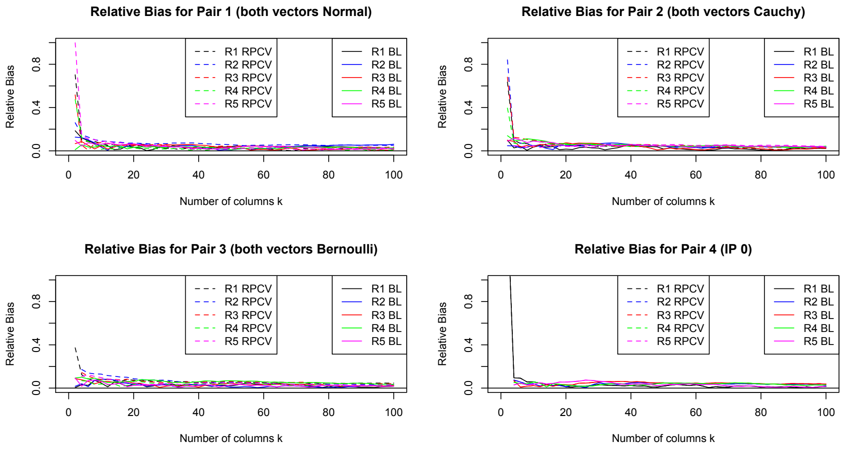

Fig. 2. Plots of relative bias of estimates of Euclidean distances against number of columns in R_i for each pair of vectors.

We then look at the plots of the ratio ρ defined by

$$\rho = \frac{\text{Variance using control variate with } R_i}{\text{Variance using baseline with } R_i} \tag{58}$$

in Figure 3 for the Euclidean distance (also used in [8]). ρ is a measure of the reduction in variance using our control variate approach with the matrix R_i rather than just using R_i alone. For this ratio, a fraction less than 1 means our control variate approach performs better than the baseline.

For all pairs $\mathbf{x}_i, \mathbf{x}_j$ except Cauchy, the reduction of variance of the estimates of the Euclidean distance using different R_is with our control variate approach converge quickly to around the same ratio. However, when data is heavy tailed, the choice of random projection matrix R_i with a control variate approach affects the reduction of variance in the estimates of the Euclidean distance, and sparse matrices R_i have a greater variance reduction for the estimates of the Euclidean distance.

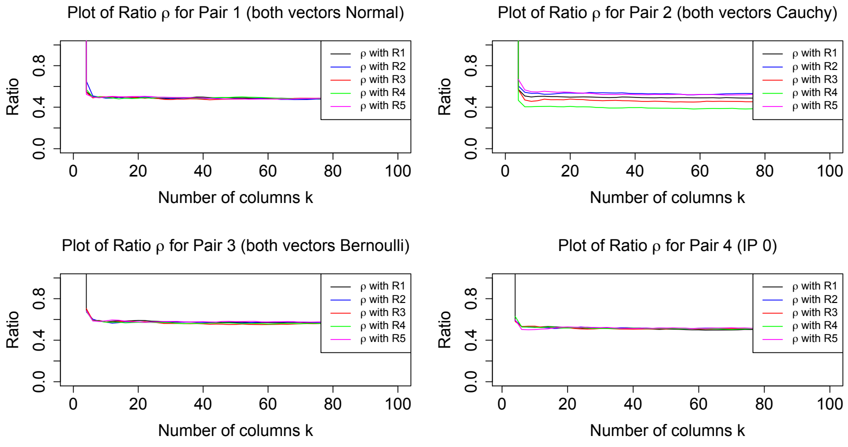

Fig. 3. Plots of ρ for Euclidean distances against number of columns in R_i for each pair of vectors.

We next look at the estimates of the inner product. In our experiments, we use Li *et al*'s method as the baseline for computing the estimates of the inner product. Our rationale for doing this is that both Li's method and our method stores the marginal norms of X, thus we should compare our method with Li's method for a fair comparison. The ratio of variance reduction is shown in Figure 4 (also used in [8]).

As the number of columns k of the random projection matrix R increases, the variance reduction in our estimate of the inner product decreases, but then increases again up to a ratio just below 1. Since Li's method uses an asymptotic maximum likelihood estimate of the inner product, then as the number of columns of R increases, the estimate of the inner product would be more accurate.

Thus, it is reasonable to use a control variate approach for Euclidean distances, and Li's method for inner products.

3.2 Experiments With Real Data

We now demonstrate our control variate approach on four datasets, the `arcene` dataset, `colon` dataset, `kos` dataset, and the `NIPS` dataset. We select these datasets since they have different characteristics (sparse / dense, variance explained / not explained in few principal components). In short

1. the `arcene` dataset [7] is an example of a dense dataset consisting of $n = 900$ observations with $p = 10000$ features. Most of the variance in this dataset is explained by the first 500 eigenvectors.
2. the `colon` dataset [3] is an example of a dense dataset consisting of $n = 62$ gene expression levels with 2000 features. Most of the variance in this dataset is explained by a few eigenvectors.

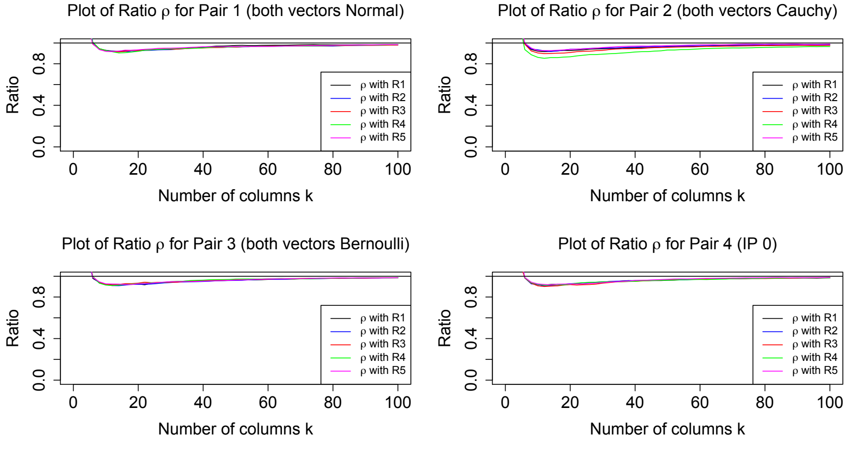

Fig. 4. Plots of ρ for inner product against number of columns in R_i for each pair of vectors.

3. the `kos` dataset [13] is an example of a sparse dataset consisting of $n = 3430$ documents and $p = 6906$ words from the KOS blog entries. Most of the variance in this dataset is explained by about a third of the eigenvectors.
4. the `nips` dataset [16] is an example of a sparse dataset consisting of $n = 5812$ observations (conference papers) in this dataset, and $p = 11463$ words. Most of the variance in this dataset is explained by slightly less than half of the eigenvectors.

We normalize each dataset such that every observation $\|\mathbf{x}_i\|_2^2 = 1$.

For each dataset, we consider the pairwise Euclidean distances of all observations $\{\mathbf{x}_i, \mathbf{x}_j\}, \ \forall \, i \neq j$, and compute the estimates of the Euclidean distance with a control variate approach of the pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$ which give the 20th, 30th, $\ldots$, 90th percentile of Euclidean distances.

We first do a quick sanity check in Figure 5. Here, we pick a pair in the 50th percentile for these datasets and show that for every different R_i, the bias quickly converges to zero.

Next, we look at the variance reduction for these pairs in Figure 6 with different types of random projection matrices R_i. We see that the variance reduction for the R_is are around the same range. Since the bias converges to zero, this implies that our control variates work. i.e., we do not get extremely biased estimates with lower variance.

We now look at what happens at different percentile pairs. Since the random projection matrices have a similar pattern in Figure 6, we will only take a look at varying pairs for the random projection matrix R_1.

Figure 7 thus shows the ratio ρ of variance reduction from the 10th percentile to the 90th percentile. Note that for dense datasets (`arcene`, `colon`), we can see a substantial percentage increase in variance reduction as the percentiles increase, but not as much for sparse datasets (`kos`, `NIPS`).

Finally, we take a look at the inner product estimates. We do not get good variance reduction results, when we used Li's method as a baseline.

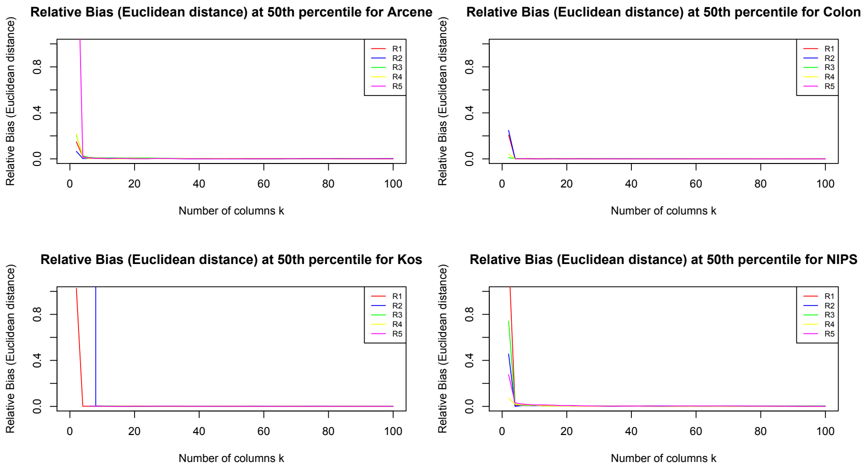

Fig. 5. Plots of relative bias in Euclidean distance for real data

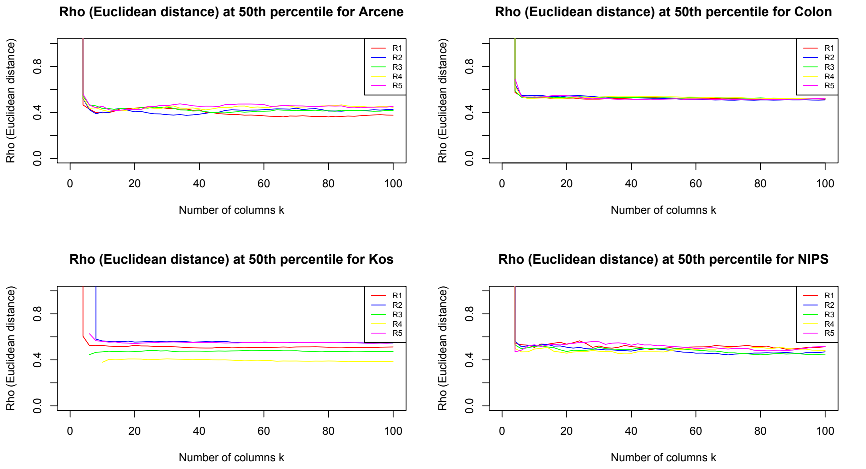

Fig. 6. Plots of ρ for Euclidean distance (varying *R*) for real data

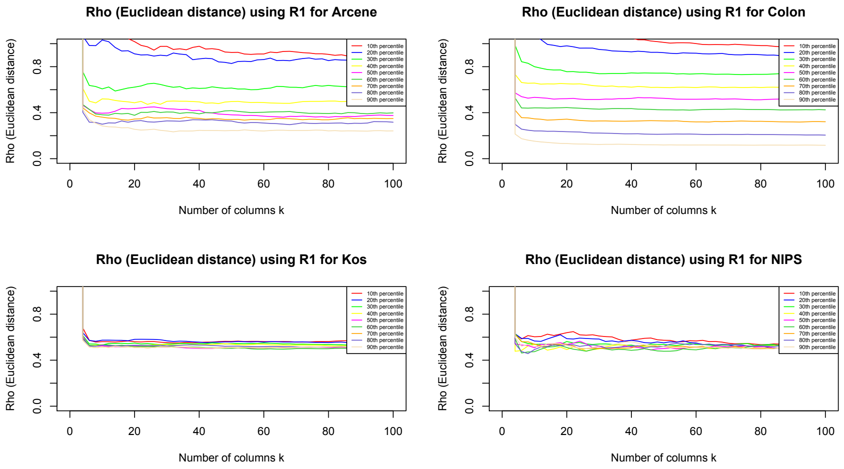

Fig. 7. Plots of ρ for Euclidean distance (varying percentile) for real data

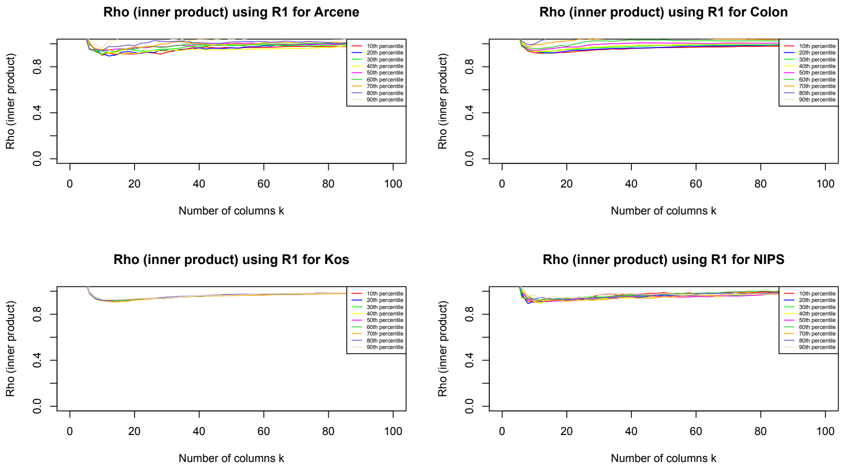

Fig. 8. Plots of ρ for inner product (varying percentile) for real data

Figure 8 shows the plots of ρ with the random projection matrix R_1 for varying percentiles. The same pattern holds for different types of matrices R_2 to R_4, and is similar to what we saw in synthetic data. While there is some small variance reduction, the ratio ρ quickly converges to a value near 1.

This matches what we see in our synthetic data.

4 Conclusion and Future Work

We have shown that which works well in conjunction with different random projection matrices to reduce the variance of the estimates of the Euclidean distance and inner products on different types of vectors $\mathbf{x}_i$, $\mathbf{x}_j$. This allows for more accurate estimates of the Euclidean distance. As the Euclidean distance between two vectors increases, we expect greater variance reduction. In essence, we have shown that it is possible to juxtapose statistical variance reduction methods with random projections to give better results.

While a control variate approach gives a variance reduction for the estimates of the inner products, the ratio of variance reduction becomes minimal as the number of columns increases when compared to Li's method. This is not surprising since Li's method for estimating the inner products is an asymptotic maximum likelihood estimator, and is extremely accurate as the number of columns of the random projection matrix increases.

Although a control variate approach requires storing marginal norms and computing the covariance between two p dimensional vectors, the cost of doing so is negligible when compared to matrix multiplication. Furthermore, the computation of marginal norms is unnecessary when the data is already normalized to have norm of 1.

In fact, a control variate approach can be seen as a method that nicely complements Li's method since both methods require storing marginal norms. This approach substantially reduces the errors of the estimates of the Euclidean distance, while Li's method substantially reduces the errors of the estimates of the inner product. The estimate of the inner product given by Li's method can even be used in computing the control variate correction c_{ED}, instead of evaluating the empirical value of c_{ED} directly, which is less costly.

We note that different applications may require a certain type of random projection matrix. Thus if we want to reduce the errors in our estimates, we cannot just switch to a different random projection matrix where the entries allow us to place sharper probability bounds on our errors. If we want data to be invariant under rotations, then a Normal random projection matrix would be best suited [14]. If we wanted to desparsify data, then a random projection matrix with i.i.d. entries from $\{-\sqrt{p}, 0, \sqrt{p}\}$, p small might be preferred [1]. If we are focused on speed and quick information retrieval, then very sparse random projections [11] or random projection matrices formed by the SHRT [4] would be more preferable. A control variate approach allows us to reduce the error in all these estimates.

We believe our work can be extended by looking at a control variate approach for other types of random projection matrices, such as sign random projections.

We further hope that this control variate approach can be adopted to current algorithms using random projections which require the computation of Euclidean distances (or inner products).

Lastly, Figure 1 suggests that we could adopt a multiple control variate approach using several dominant eigenvectors of the data as control variates, since these eigenvectors would point in the general direction of the variance, and we are currently exploring this idea.

References

1. Achlioptas, D.: Database-friendly Random Projections: Johnson-Lindenstrauss with Binary Coins. J. Comput. Syst. Sci. 66(4), 671–687 (Jun 2003), http://dx.doi.org/10.1016/S0022-0000(03)00025-4
2. Ailon, N., Chazelle, B.: The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors. SIAM J. Comput. 39(1), 302–322 (May 2009), http://dx.doi.org/10.1137/060673096
3. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96(12), 6745–6750 (Jun 1999)
4. Boutsidis, C., Gittens, A.: Improved matrix algorithms via the subsampled randomized hadamard transform. CoRR abs/1204.0062 (2012), http://arxiv.org/abs/1204.0062
5. Boutsidis, C., Zouzias, A., Drineas, P.: Random projections for k-means clustering. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23, pp. 298–306. Curran Associates, Inc. (2010), http://papers.nips.cc/paper/3901-random-projections-for-k-means-clustering.pdf
6. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: A cluster ensemble approach. pp. 186–193 (2003)
7. Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 545–552. MIT Press (2005), http://papers.nips.cc/paper/2728-result-analysis-of-the-nips-2003-feature-selection-challenge.pdf
8. Kang, K., Hooker, G.: Random projections with control variates. In: Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,. pp. 138–147. INSTICC, ScitePress (2017)
9. Li, P., Church, K.W.: A Sketch Algorithm for Estimating Two-Way and Multi-Way Associations. Comput. Linguist. 33(3), 305–354 (Sep 2007), http://dx.doi.org/10.1162/coli.2007.33.3.305
10. Li, P., Hastie, T., Church, K.W.: Improving Random Projections Using Marginal Information. In: Lugosi, G., Simon, H.U. (eds.) COLT. Lecture Notes in Computer Science, vol. 4005, pp. 635–649. Springer (2006)
11. Li, P., Hastie, T.J., Church, K.W.: Very Sparse Random Projections. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 287–296. KDD '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1150402.1150436
12. Liberty, E., Ailon, N., Singer, A.: Dense fast random projections and lean walsh transforms. In: Goel, A., Jansen, K., Rolim, J.D.P., Rubinfeld, R. (eds.) APPROX-RANDOM. Lecture Notes in Computer Science, vol. 5171, pp. 512–522. Springer (2008)

13. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
14. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press (1979)
15. Paul, S., Boutsidis, C., Magdon-Ismail, M., Drineas, P.: Random Projections for Support Vector Machines. CoRR abs/1211.6085 (2012), http://arxiv.org/abs/1211.6085
16. Perrone, V., Jenkins, P.A., Spano, D., Teh, Y.W.: Poisson random fields for dynamic feature models (2016), arXiv e-prints: 1611.07460
17. Ross, S.M.: Simulation, Fourth Edition. Academic Press, Inc., Orlando, FL, USA (2006)
18. Vempala, S.S.: The Random Projection Method, DIMACS series in discrete mathematics and theoretical computer science, vol. 65. Providence, R.I. American Mathematical Society (2004), http://opac.inria.fr/record=b1101689, appendice p.101-105

APPENDIX

We use the following lemma for ease of computation of first and second moments.

Lemma 2. Suppose we have a sequence of terms $\{t_i\}_{i=1}^p = \{a_i r_i\}_{i=1}^p$ for $\mathbf{a} = (a_1, a_2, \ldots, a_p)$, $\{s_i\}_{i=1}^p = \{b_i r_i\}_{i=1}^p$ for $\mathbf{b} = (b_1, b_2, \ldots, b_p)$ and r_i i.i.d. random variables with $\mathbb{E}[r_i] = 0$, $\mathbb{E}[r_i^2] = 1$ and finite third, and fourth moments, denoted by μ_3, μ_4 respectively. Then:

$$\mathbb{E}\left[\left(\sum_{i=1}^p t_i\right)^2\right] = \sum_{i=1}^p a_i^2 = \|\mathbf{a}\|_2^2 \tag{59}$$

$$\mathbb{E}\left[\left(\sum_{i=1}^p t_i\right)^4\right] = \mu_4 \sum_{i=1}^p a_i^4 + 6 \sum_{u=1}^{p-1} \sum_{v=u+1}^p a_u^2 a_v^2 \tag{60}$$

$$\mathbb{E}\left[\left(\sum_{i=1}^p s_i\right)\left(\sum_{i=1}^p t_i\right)\right] = \sum_{i=1}^p a_i b_i = \langle \mathbf{a}, \mathbf{b} \rangle \tag{61}$$

$$\mathbb{E}\left[\left(\sum_{i=1}^p s_i\right)^2 \left(\sum_{i=1}^p t_i\right)^2\right] = \sum_{i=1}^p a_i^2 b_i^2 + \sum_{i \neq j} a_i^2 b_j^2 + 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^p a_u b_u a_v b_v \tag{62}$$

The motivation for this lemma is that we do expansion of terms of the above four forms to prove our theorems.