Random Projections with Control Variates

Keegan Kang and Giles Hooker
Department of Statistical Science, Cornell University, Ithaca 14850, New York, U.S.A.
{*tk528, gjh27*}*@cornell.edu*

Keywords: Control Variates, Random Projections.

Abstract: Random projections are used to estimate parameters of interest in large scale data sets by projecting data into a lower dimensional space. Some parameters of interest between pairs of vectors are the Euclidean distance and the inner product, while parameters of interest for the whole data set could be its singular values or singular vectors. We show how we can borrow an idea from Monte Carlo integration by using control variates to reduce the variance of the estimates of Euclidean distances and inner products by storing marginal information of our data set. We demonstrate this variance reduction through experiments on synthetic data as well as the colon and kos datasets. We hope that this inspires future work which incorporates control variates in further random projection applications.

1 INTRODUCTION

Random projection is one of the methods used in dimension reduction, in which data in high dimensions is projected to a lower dimension using a random matrix R. The entries r_{ij} in the matrix R can either be i.i.d. with mean $\mu = 0$ and second moment $\mu_2 = 1$, or correlated with each other. Some examples of random projection matrices with i.i.d. entries are those with binary entries (Achlioptas, 2003), or sparse random projections (Li et al., 2006b). Random matrices with correlated entries range from those constructed by the Lean Walsh Transform (Liberty et al., 2008) to the Fast Johnson Lindenstrauss Transform (FJLT) (Ailon and Chazelle, 2009) and the Subsampled Randomized Hadamard Transform (SRHT) (Boutsidis and Gittens, 2012).

We can think of 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 $\tilde{\mathbf{x}}^T = \mathbf{x}^T R$. 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 $\mathbb{E}[g(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)] = f(\mathbf{x}_i, \mathbf{x}_j)$. If we want the Euclidean distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, then $f(a,b) = g(a,b) = \|a-b\|_2$.

The methods used in construction and application of the random projection matrix R to the vectors $\mathbf{x}_i$s have tradeoffs. Very sparse random projections, FJLT, and the SRHT are fast methods with a tradeoff in accuracy. The former uses extremely sparse R for quick matrix multiplication (optimal R has about $\frac{\sqrt{p}-1}{\sqrt{p}}$ zero entries), and the latter two uses the recursive property of the Hadamard matrix for quick matrix vector multiplication. Dense R with entries generated from the Normal or the Rademacher distribution gives more accurate estimates and desparsifies data but at a cost of speed.

The 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 paper, we propose a method *Random Projections with Control Variates* (RPCV), which is used in conjunction with the above types of different random projection matrices. Our approach 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. These measures of distances are commonly used in clustering (Fern and Brodley, 2003), (Boutsidis et al., 2010), classification (Paul et al., 2012), and set resemblance problems (Li et al., 2006a).

The paper is structured as follows: We first express our notation differently from the ordinary random projection notation to give intuition on how we can use control variates. We then briefly discuss control variates, before describing RPCV. Lastly, we demonstrate RPCV on both synthetic and experimental data and show that we can use RPCV together with any random projection method to gain variance reduc-

Kang, K. and Hooker, G.
Random Projections with Control Variates.
DOI: 10.5220/0006188801380147
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 138-147
ISBN: 978-989-758-222-6

tion 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. The random projection equation is then given by

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

However, we will use

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

without the scaling factor. Consider the random matrix R written as

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

where each $\mathbf{r}_i$ is a column vector with i.i.d. entries. Then for a *fixed* row $\mathbf{x}_i^T$, we have that for all j, $v_{ij} = \mathbf{x}_i^T \mathbf{r}_j$ is a random variable from the same distribution. Here, we focus on each v_{ij} as a single element, rather than seeing $v_{i1}, \ldots, v_{ik}$ comprising the row vector $\mathbf{v}_{i.}^T$.

1.2 Control Variates

Given the notion of each v_{ij} as a random variable, we introduce control variates. Control variates are a technique in Monte Carlo simulation using random variables for variance reduction. A more thorough explanation 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

$$\mathbb{E}[A + c(B - \mu_B)] \;=\; \mathbb{E}[A] + c\mathbb{E}[B - \mu_B] \;=\; \mu_A \tag{4}$$

which is an unbiased estimator of μ_A for some constant c. This value of c which minimizes the variance is given by

$$\hat{c} = -\frac{\mathrm{Cov}(A,B)}{\mathrm{Var}(B)} \tag{5}$$

and thus we write

$$\mathrm{Var}[A + c(B - \mu_B)] = \mathrm{Var}(A) - \frac{(\mathrm{Cov}(A,B))^2}{\mathrm{Var}(B)} \tag{6}$$

In our random projection scenario for a fixed i, we can think of a random variable from A as some v_{ij}, where

$$\mathbb{E}[v_{i.}] = \frac{1}{k}\sum_{m=1}^{k} v_{im} = \frac{1}{k}\sum_{m=1}^{k}\left(\sum_{n=1}^{p} x_{i,n} r_{nm}\right) \tag{7}$$

under the law of large numbers.

Intuitively, we then need to find some distribution B where the variables b_i are correlated with v_{ij} to get good variance reduction. To do this, B necessarily needs to fulfill two conditions.

Condition 1: Since each realization v_{ij} is the sum of p random variables $r_{1j}, r_{2j}, \ldots, r_{pj}$, we need to have y_i constructed from these same random variables *and* also correlated with each $x_{i1}, \ldots, x_{ip}$ 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.

1.3 Related Work

We draw inspiration from the works of Li and Church, 2007, Li et al., 2006a, and Li et al., 2006b. In these papers, marginal information such as margin counts or margin norms from data is pre-computed and stored. This extra information is then used with asymptotic maximum likelihood estimators to estimate parameters of interests.

We also store marginal information from our matrix X, but instead use this information to determine a control variate, rather than a maximum likelihood estimator. We compute and store all the n norms $\|\mathbf{x_i}\|^2$ from our X. Computing all these norms are cheap as they are of order $O(np)$, and can be done when reading in the data at the same time.

Furthermore if the data is normalized (normalizing is also of order $O(np)$ which we usually take for granted), we get the norms $\|\mathbf{x}_i\|_2^2 = 1$ for free.

1.4 Our Contributions

We propose *Random Projections with Control Variates* in this paper which reduces the variance of the estimates of the Euclidean distances and the inner product between pairs of vectors for a choice of random projection matrix R. In particular

- We describe the process of RPCV, which keeps to the same order of runtime as the particular random projection matrix we use RPCV with.
- 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.

- We demonstrate empirically that RPCV works well with current random projection methods on synthetically generated data and the `colon` and `kos` datasets.

2 PROCESS OF RPCV

We describe and illustrate the process of RPCV 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 $X\mathbf{r}$. As an illustrative example in the case where $p = 2$, we would have

$$V = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = X\mathbf{r} \quad (8)$$

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 RPCV 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 2.1. *Let one realization of $A = (v_1 - v_2)^2$, which is our Euclidean distance in expectation. Let one realization of B to be $(v_1 - v_2)^2 + 2v_1v_2 = v_1^2 + v_2^2$ with mean $\mu_B = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|_2^2$. 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

$$\begin{aligned} &\mathbb{E}[(v_1 - v_2)^2] + 2\mathbb{E}[v_1v_2] \\ &= \|\mathbf{x}_1 - \mathbf{x}_2\|^2 + 2\langle \mathbf{x}_1\mathbf{x}_2 \rangle \quad (9) \\ &= \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 - 2\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + 2\langle \mathbf{x}_1, \mathbf{x}_2 \rangle \quad (10) \\ &= \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 \quad (11) \end{aligned}$$

□

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

Lemma 2.1. 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 under this set up for Euclidean distances in Theorem 2.1, we have

$$\begin{aligned} \mathbb{E}[A^2] = \; & \mu_4 \sum_{j=1}^{p} (x_{1j} - x_{2j})^4 \\ & + 6 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} - x_{2u})^2 (x_{1v} - x_{2v})^2 \quad (12) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[B^2] = \; & \mu_4 \sum_{j=1}^{p} (x_{1j}^4 + x_{2j}^4) + 6 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u}^2 x_{1v}^2 + x_{2u}^2 x_{2v}^2) \\ & + 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u}x_{1v}x_{2u}x_{2v}) + \mu_4 \sum_{j=1}^{p} x_{1j}^2 x_{2j}^2 \\ & + \sum_{i \neq j}^{p} x_{1i}^2 y_{2j}^p \quad (13) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[AB] = \; & 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} - x_{2u})(x_{1v} - x_{2v})(x_{1u}x_{1v} + x_{2u}x_{2v}) \\ & + \mu_4 \sum_{j=1}^{p} (x_{1j} - x_{2j})^2 (x_{1j}^2 + x_{2j}^2) \\ & + \sum_{i \neq j} (x_{1i} - x_{2i})^2 (x_{1i}^2 + x_{2j}^2) \quad (14) \end{aligned}$$

Proof. We repeatedly apply Lemma 4.1 in the Appendix. □

Thus, by following Lemma 2.1, we are able to derive expressions for the optimal control variate correction c in our procedure as follows.

Theorem 2.2. *The optimal value c is given by*

$$c = \frac{Cov(A,B)}{Var[B]} \quad (15)$$

where we have

$$Cov(A,B) = \mathbb{E}[AB - A\mu_B - B\mu_A + \mu_A\mu_B] \quad (16)$$

and

$$Var[B] = \mathbb{E}[B^2] - (\mathbb{E}[B])^2 \quad (17)$$

They expand to

$$\begin{aligned} Cov(A,B) = \; & 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (x_{1u} - x_{2u})(x_{1v} - x_{2v})(x_{1u}x_{1v} + x_{2u}x_{2v}) \\ & + (\mu_4 - 1) \sum_{j=1}^{p} (x_{1j} - x_{2j})^2 (x_{1j}^2 + x_{2j}^2) \quad (18) \end{aligned}$$

and

$$Var[B] = (\mu_4 - 1)\sum_{j=1}^{p}(x_{1j}^4 + x_{2j}^4) + 4\sum_{u=1}^{p-1}\sum_{v=u+1}^{p}(x_{1u}^2 x_{1v}^2 + x_{2u}^2 x_{2v}^2) + 4\sum_{u=1}^{p-1}\sum_{v=u+1}^{p} x_{1u}x_{1v}x_{2u}x_{2v} + (\mu_4 - 2)\sum_{j=1}^{p} x_{1j}^2 x_{2j}^2 - \sum_{i\neq j} x_{1i}^2 x_{2j}^2 \quad (19)$$

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

Theorem 2.3. *The first and second moments are*

$$\mathbb{E}[A + c(B - \mu_B)] = \mathbb{E}[A] + c\mathbb{E}[B - \mu_B] = 0 \quad (20)$$

and

$$\mathbb{E}[(A + c(B - \mu_B))^2] = \mathbb{E}[A^2 + 2cAB - 2c\mu_B A + c^2 B^2 - 2c^2\mu_B B + c^2\mu_B^2] \quad (21)$$

where we substitute in the values of $\mathbb{E}[A^2], \mathbb{E}[AB], \mathbb{E}[B^2]$ *from Lemma 2.1.*

2.1.1 Motivation for c in Euclidean Distance

To give some motivation for the meaning of c, we simplify the general case and consider what the ratio tells us when we have normalized vectors, i.e. $\|\mathbf{x}_i\|_2^2 = 1$ and when $\mu_4 = 1$ (eg, where we generate $r_{ij} \sim \{\pm 1\}$ with equal probability). In this case, we have

$$\mathrm{Cov}(A,B) = 4\sum_{u=1}^{p-1}\sum_{v=u+1}^{p}(x_{1u} - x_{2u})(x_{1v} - x_{2v}) \times (x_{1u}x_{1v} + x_{2u}x_{2v}) \quad (22)$$

$$\mathrm{Var}[B] = 4\sum_{u=1}^{p-1}\sum_{v=u+1}^{p}(x_{1u}^2 x_{1v}^2 + x_{2u}^2 x_{2v}^2) + 4\sum_{u=1}^{p-1}\sum_{v=u+1}^{p} x_{1u}x_{1v}x_{2u}x_{2v} - \sum_{j=1}^{p} x_{1j}^2 x_{2j}^2 - \sum_{i\neq j} x_{1i}^2 x_{2j}^2 \quad (23)$$

Consider the expansion of $\left(\sum_{i=1}^{p} u_i\right)\left(\sum_{i=1}^{p} v_i\right)$, and consider $u_i v_i$ as diagonal terms, and $u_i v_j,\ i \neq j$ as off diagonal terms for some expressions u_i, v_i.

Then c can be seen as some weighted ratio of the sum of off-diagonal terms of the "Euclidean distance vector" $u_i := (x_{1i} - x_{2i})$ weighted by off diagonal terms of x_{1i}, x_{2i} to the sum of off diagonal terms of the norms.

Intuitively, this implies that if the Euclidean distance between two vectors is high, then we would get greater variance reduction (c is large).

2.2 RPCV 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 2.4. *Let one realization of* $A = v_1 v_2$*, which is our inner product in expectation. Let one realization of* B *to be* $(v_1 - v_2)^2 + 2v_1 v_2 = v_1^2 + v_2^2$ *with mean* $\mu_B = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|_2^2$*. 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 2.5. *The optimal value of* c *is given by*

$$c = \frac{Cov(A,B)}{Var[B]} \quad (24)$$

where

$$Cov(A,B) = \mathbb{E}[AB - A\mu_B - B\mu_A + \mu_A\mu_B] = (\mu_4 - 1)\sum_{j=1}^{p} x_{1j}x_{2j}(x_{1j}^2 + x_{2j}^2) + \sum_{i\neq j} x_{1i}x_{2j}(x_{1i}x_{1j} + x_{2i}x_{2j}) \quad (25)$$

and the value of $Var[B]$ *taken from the result in Theorem 2.2.*

However, we should not just stop there at our estimate of the inner product using RPCV. Li et al., 2006a describes a more accurate estimator for the inner product using the marginal information $\|\mathbf{x}_1\|^2$ and $\|\mathbf{x}_2\|^2$, where the estimate of the inner product is the root of the equation

$$f(a) = a^3 - a^2\langle \mathbf{v}_1, \mathbf{v}_2\rangle + a(-\|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2\|\mathbf{x}_1\|^2\|\mathbf{v}_2\|^2 + \|\mathbf{x}_2\|^2\|\mathbf{v}\|_1^2) - \|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2\langle \mathbf{v}_1, \mathbf{v}_2\rangle \quad (26)$$

Since we stored and used $\|\mathbf{x}_1\|^2$ and $\|\mathbf{x}_2\|^2$ in order to get better estimates of the Euclidean distance and the inner product, we should use Li's method to get a better estimate of our inner product, by using RPCV's estimated value of $\langle \mathbf{v}_1, \mathbf{v}_2\rangle$ and $\|\mathbf{v}_i\|^2$ in the cubic equation instead.

In practice, since the control variate method gives results with similar accuracy to Li's method for inner products, one could use our control variate method for Euclidean distances to complement Li's method for inner products, as both methods make use of storing the norms of each observation.

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 $[(\mathbf{0})_{p/2},(\mathbf{1})_{p/2}]$

2.2.1 Motivation for c in Inner Product

To give some motivation for this meaning of c, we again simplify the general case and consider what the ratio tells us when we have normalized vectors and when $\mu_4 = 1$. In this case, we have

$$\mathrm{Cov}(A,B) = \sum_{u\neq v} x_{1u}x_{2v}(x_{1u}x_{1v} + x_{2u}x_{2v}) \tag{27}$$

Compared to what we have seen for Euclidean distances (recall that the denominator $\mathrm{Var}[B]$ is unchanged), the magnitude of c for inner products is comparatively smaller compared to c for Euclidean distances (expand both $\mathrm{Cov}(A,B)$ for the Euclidean distance (Equation 22), and $\mathrm{Cov}(A,B)$ (Equation 27) for the inner product and compare terms). We would then expect the variance reduction for inner product to not be as substantial as the variance reduction for the Euclidean distance.

2.3 Motivation for Computing First and Second Moments

The probability bounds of the errors in our estimate (where entries of R are i.i.d.) are of the form

$$\mathbb{P}[\|\mathbf{v}\| \leq (1-\varepsilon)\|\mathbf{x}\|] \leq f_1(k,\varepsilon) \tag{28}$$
$$\mathbb{P}[\|\mathbf{v}\| \geq (1+\varepsilon)\|\mathbf{x}\|] \leq f_2(k,\varepsilon) \tag{29}$$

where k is the number of columns of the random projection matrix. The Markov inequality is used to bound $\|\mathbf{v}\|$ by the first and second moments together with the Taylor's expansion. A full description of these results can be found in Vempala, 2004.

If we construct R with i.i.d. $r_{ij} \sim N(0,1)$, or $r_{ij} \sim \{\pm 1\}$, then we can easily find similar probability bounds for the Euclidean distance by setting $\|\mathbf{v}\| = \|\mathbf{v}_1 - \mathbf{v}_2\|$. In the RPCV case, each element $v_{1i} - v_{2i}$ in $\|\mathbf{v}_1 - \mathbf{v}_2\|$ now corresponds to

$$(v_{1i}-v_{2i})^2 + c(\mathbf{v}_{1i}^2 + \mathbf{v}_{2i}^2 - \|\mathbf{x}\|^2 - \|\mathbf{x}_2\|^2) \tag{30}$$

and thus we need to find probability bounds for this expression.

For the inner product, we note that

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \frac{1}{4}\left(\|\mathbf{v}_i + \mathbf{v}_j\|^2 - \|\mathbf{v}_i - \mathbf{v}_j\|^2\right) \tag{31}$$
$$= \frac{1}{4}\left(\|\mathbf{v}_i - \mathbf{v}_j\|^2 - \|\mathbf{v}_i + \mathbf{v}_j\|^2\right) \tag{32}$$

and by rearranging expressions (31) and (32) together with (28) and (29), we get

$$\mathbb{P}[\mathbf{v}_1^T\mathbf{v}_2 \leq (1-\varepsilon)\mathbf{x}_1^T\mathbf{x}_2] \leq f_1(k,\varepsilon) + f_2(k,\varepsilon) \tag{33}$$
$$\mathbb{P}[\mathbf{v}_1^T\mathbf{v}_2 \geq (1+\varepsilon)\mathbf{x}_1^T\mathbf{x}_2] \leq f_1(k,\varepsilon) + f_2(k,\varepsilon) \tag{34}$$

Thus, computing the first and second moments for the expression $A + c(B - \mu_B)$ for $A = (v_1 - v_2)^2$, $B = \|\mathbf{v}_1\|_2^2 + \|\mathbf{v}_2\|_2^2$ for Euclidean distances suffices, provided we compute $\hat{c}$ for the Euclidean distance, and $\tilde{c}$ for the inner product. We necessarily need to substitute the value of $\hat{c}$ (or $\tilde{c}$) to get the first and second moments of $A + c(B - \mu_B)$ for the Euclidean distance (or inner product).

For R constructed with r_{ij} from other distributions, computing these bounds are a bit more involved.

2.4 Overall Computational Time

We need to compute the empirical covariance between all pairs A and B as well as the variance of B, which takes an additional $O(k)$ time. Since the vectors we need to compute this covariance are the elements of V, we do not need to do further computation to get them. Furthermore, computing the covariance takes the same order of time as finding the Euclidean distance (or inner product) between the vectors $\mathbf{v}_i$, $\mathbf{v}_j$.

If we want a more accurate estimate of the inner product using Li's method, we can either use a root finding method to find a where $f(a) = 0$, or use the cubic formula to get the root(s) of a degree 3 polynomial. The time for these methods are are bounded above by some constant number of operations.

3 OUR EXPERIMENTS

Throughout our experiments, we use five different types of random projection matrices as shown in Table 1. 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.

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. 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 look at the plots of the ratio ρ defined by

$$\rho = \frac{\text{Variance using control variate with } R_i}{\text{Variance using baseline with } R_i} \quad (35)$$

in Figure 1 for the Euclidean distance. ρ is a measure of the reduction in variance using RPCV with the matrix R_i rather than just using R_i alone. For this ratio, a fraction less than 1 means RPCV 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 RPCV converge quickly to around the same ratio. However, when data is heavy tailed, the choice of random projection matrix R_i with RPCV 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.

We next look at the estimates of the inner product. In our experiments, we use Li et al., 2006a'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 2.

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 RPCV for Euclidean distances, and Li's method for inner products.

3.2 Estimating the Euclidean distance of vectors with real data sets

We now demonstrate RPCV on two datasets, the `colon` dataset from Alon et al., 1999 and the `kos` dataset from Lichman, 2013.

The `colon` dataset is an example of a dense dataset consisting of 62 gene expression levels with 2000 features, and thus we have $\mathbf{x}_i \in \mathbb{R}^{2000}$, $1 \le i \le 62$.

The `kos` dataset is an example of a sparse dataset consisting of 3430 documents and 6906 words from the KOS blog entries, and thus we have $\mathbf{x}_i \in \mathbb{R}^{3430}$, $1 \le i \le 6906$.

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 RPCV of the pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$ which give the 20th, 30th, ..., 90th percentile of Euclidean distances.

We pick a pair in the 50th percentile for both the `colon` and `kos` datasets (Figure 3 and Figure 4), and show that for every different R_i, the bias quickly converges to zero, and 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 the variance reduction for pairs from the 20th to 90th percentile of Euclidean distances from both datasets for R_1 (where $r_{ij} \sim N(0,1)$). This is shown in Figure 5. We omit plots of the biases, as well as plots of ρ varying for different random matrices R_2 to R_5 since the variance reduction follows a similar trend.

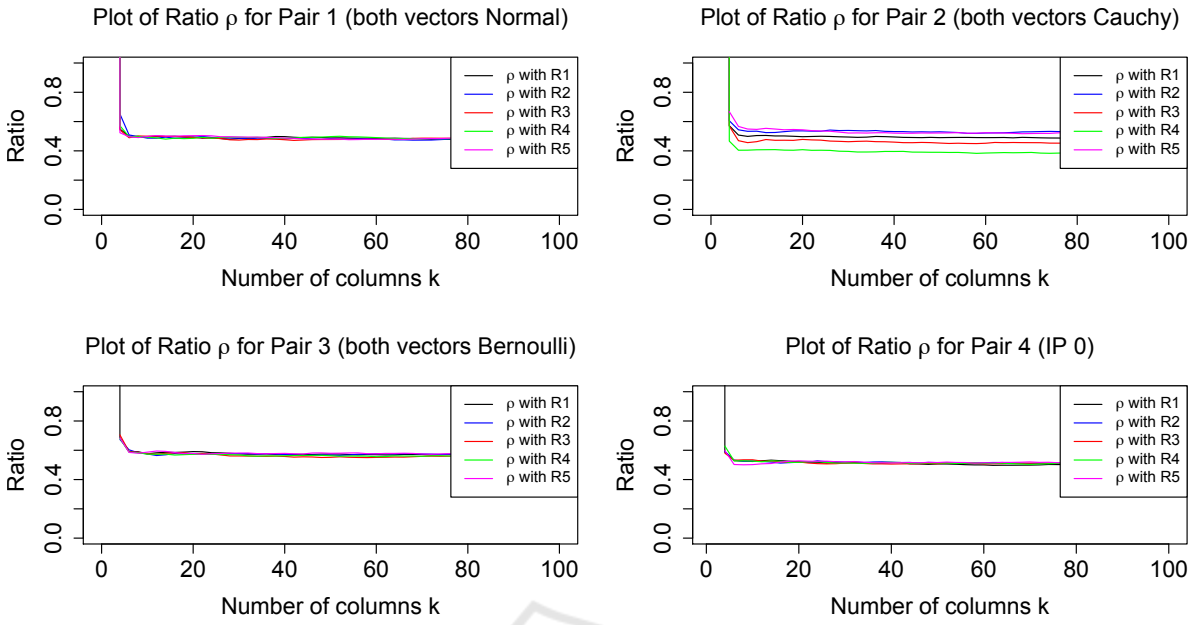

Figure 1: Plots of ρ for Euclidean Distances against number of columns in R_i for each pair of vectors.

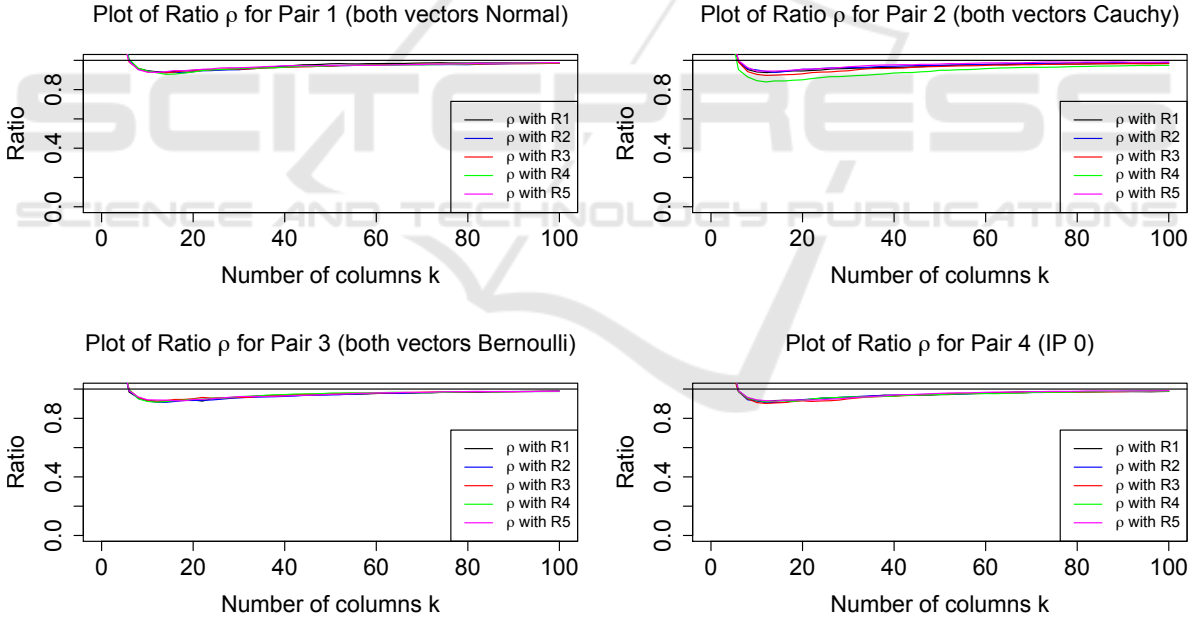

Figure 2: Plots of ρ for inner product against number of columns in R_i for each pair of vectors.

We also see that as the Euclidean distances between vectors increases (percentile increases), we get more variance reduction in our estimates. This increase in variance reduction is strongly seen in our dense `colon` dataset. Furthermore, both datasets show substantial variance reduction regardless of the percentile values.

4 CONCLUSION AND FUTURE WORK

We have presented a new method RPCV 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 differ-

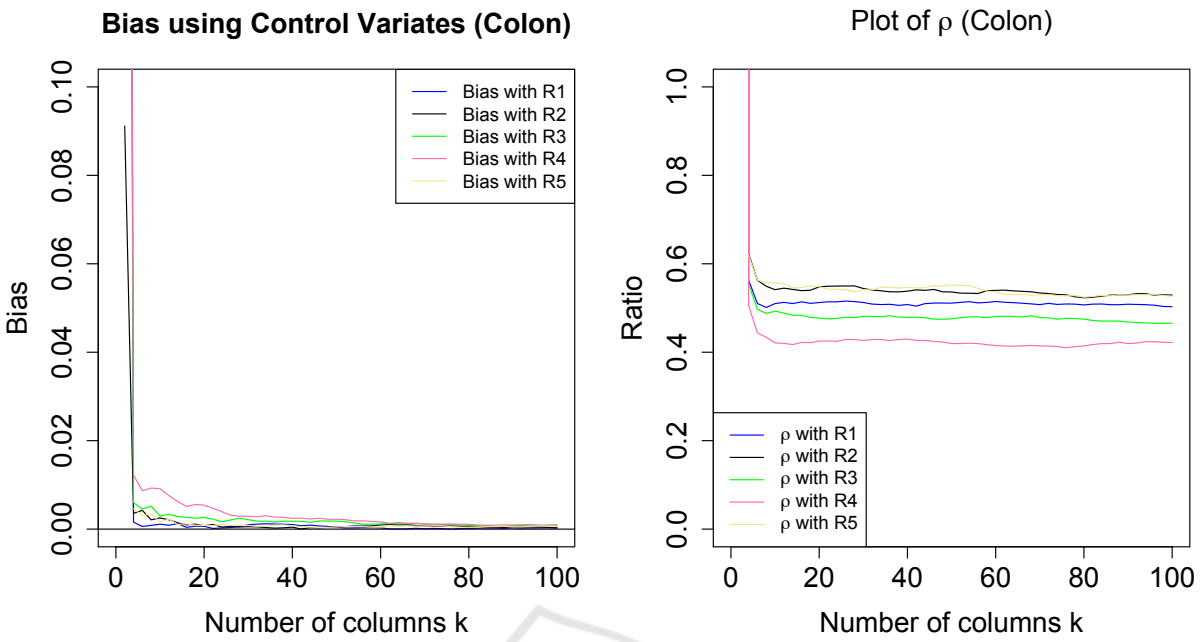

Figure 3: Plots of bias and variance reduction of Euclidean distances at 50th percentile against number of columns in R_i for colon data.

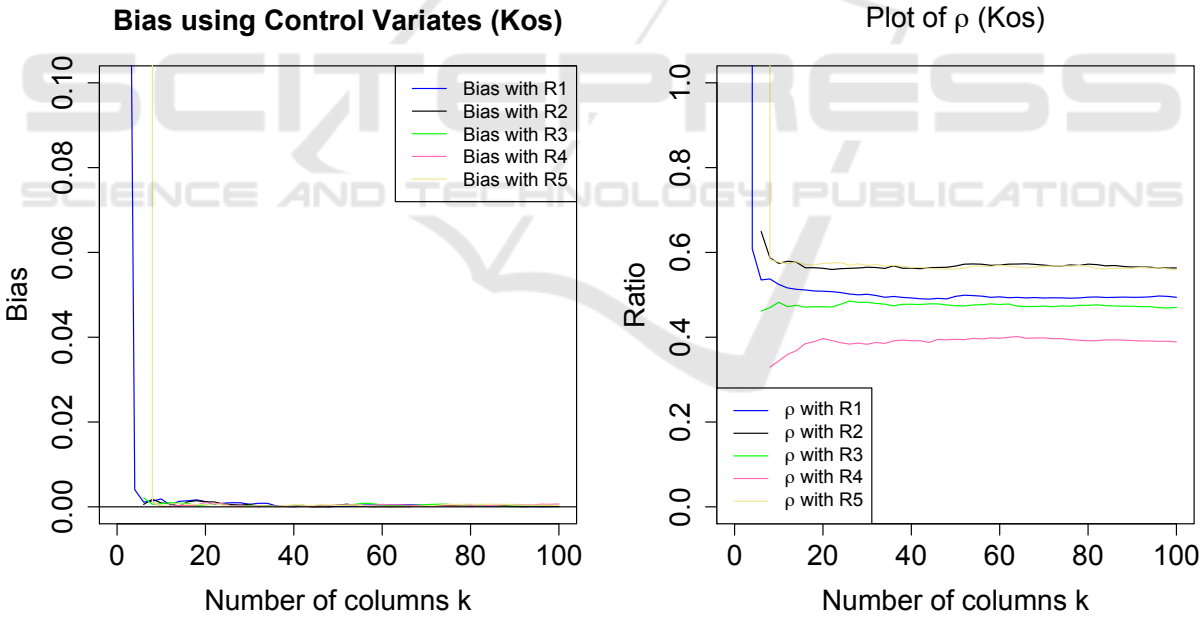

Figure 4: Plots of bias and variance reduction of Euclidean distances at 50th percentile against number of columns in R_i for kos data.

ent 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 RPCV 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

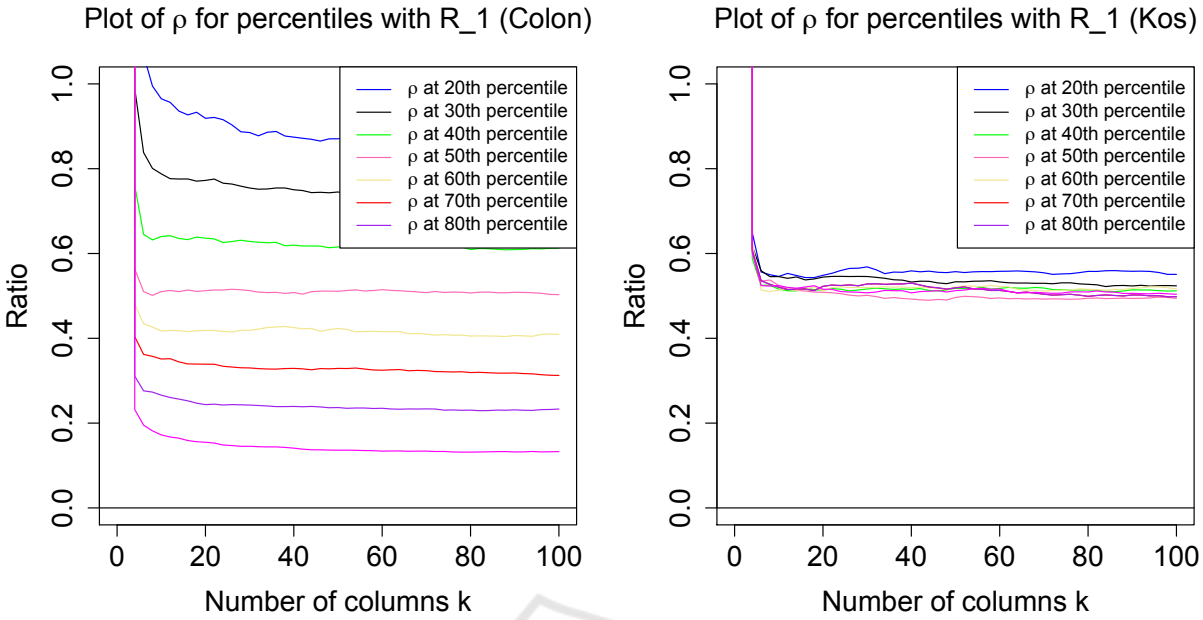

Figure 5: Plots of ρ for percentiles of Euclidean distance for R_1 in both `colon` and `kos` data.

columns increases.

Although RPCV 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.

In fact, RPCV can be seen as a method that nicely complements Li's method since both methods require storing marginal norms. RPCV 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.

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 (Mardia et al., 1979). 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 (Achlioptas, 2003). If we are focused on speed and quick information retrieval, then very sparse random projections (Li et al., 2006b) or random projection matrices formed by the SHRT (Boutsidis and Gittens, 2012) would be more preferable. RPCV allows us to reduce the error in all these estimates.

While we have demonstrated good empirical results in the variance reduction for Euclidean distances for RPCV, we still need an expression for the first and second moments of $A + c(B - \mu_B)$ when the elements in the random projection matrix R are correlated in order to theoretically show that RPCV does achieve this reduction in variance. We are currently working on this.

Finally, we want look forward to extending this method of control variates to other applications of random projections.

REFERENCES

Achlioptas, D. (2003). Database-friendly Random Projections: Johnson-Lindenstrauss with Binary Coins. *J. Comput. Syst. Sci.*, 66(4):671–687.

Ailon, N. and Chazelle, B. (2009). The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM J. Comput.*, 39(1):302–322.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). 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.

Boutsidis, C. and Gittens, A. (2012). Improved matrix algorithms via the subsampled randomized hadamard transform. *CoRR*, abs/1204.0062.

Boutsidis, C., Zouzias, A., and Drineas, P. (2010). Random projections for k-means clustering. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information*

Processing Systems 23, pages 298–306. Curran Associates, Inc.

Fern, X. Z. and Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. pages 186–193.

Li, P. and Church, K. W. (2007). A Sketch Algorithm for Estimating Two-Way and Multi-Way Associations. *Comput. Linguist.*, 33(3):305–354.

Li, P., Hastie, T., and Church, K. W. (2006a). Improving Random Projections Using Marginal Information. In Lugosi, G. and Simon, H.-U., editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 635–649. Springer.

Li, P., Hastie, T. J., and Church, K. W. (2006b). Very Sparse Random Projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 287–296, New York, NY, USA. ACM.

Liberty, E., Ailon, N., and Singer, A. (2008). Dense fast random projections and lean walsh transforms. In Goel, A., Jansen, K., Rolim, J. D. P., and Rubinfeld, R., editors, *APPROX-RANDOM*, volume 5171 of *Lecture Notes in Computer Science*, pages 512–522. Springer.

Lichman, M. (2013). UCI machine learning repository.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Paul, S., Boutsidis, C., Magdon-Ismail, M., and Drineas, P. (2012). Random Projections for Support Vector Machines. *CoRR*, abs/1211.6085.

Ross, S. M. (2006). *Simulation, Fourth Edition*. Academic Press, Inc., Orlando, FL, USA.

Vempala, S. S. (2004). *The Random Projection Method*, volume 65 of *DIMACS series in discrete mathematics and theoretical computer science*. Providence, R.I. American Mathematical Society. Appendice p.101-105.

APPENDIX

While computing first and second moments necessarily require lots of algebra, we use the following lemma for ease of computation.

Lemma 4.1. 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:

$$\mathbb{E}\left[\left(\sum_{i=1}^{p} t_i\right)^2\right] = \sum_{i=1}^{p} a_i^2 = \|\mathbf{a}\|_2^2 \tag{36}$$

$$\mathbb{E}\left[\left(\sum_{i=1}^{p} t_i\right)^4\right] = \mu_4 \sum_{i=1}^{p} a_i^4 + 6 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} a_u^2 a_v^2 \tag{37}$$

$$\mathbb{E}\left[\left(\sum_{i=1}^{p} s_i\right)\left(\sum_{i=1}^{p} t_i\right)\right] = \sum_{i=1}^{p} a_i b_i = \langle \mathbf{a}, \mathbf{b} \rangle \tag{38}$$

$$\mathbb{E}\left[\left(\sum_{i=1}^{p} s_i\right)^2\left(\sum_{i=1}^{p} t_i\right)^2\right] = \sum_{i=1}^{p} a_i^2 b_i^2 + \sum_{i \neq j} a_i^2 b_j^2 + 4 \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} a_u b_u a_v b_v \tag{39}$$

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