Expected Value of Matrix Quadratic Forms with Wishart distributed Random Matrices

Melinda Hagedorn, Heinrich Heine University, Düsseldorf, Germany

Dec. 13, 2022

All data generated or analysed in this article are available in [4].

Abstract

To explore the limits of a stochastic gradient method, it may be useful to consider an example consisting of an infinite number of quadratic functions. In this context, it is appropriate to determine the expected value and the covariance matrix of the stochastic noise, i.e. the difference of the true gradient and the approximated gradient generated from a finite sample. When specifying the covariance matrix, the expected value of a quadratic form QBQ is needed, where Q is a Wishart distributed random matrix and B is an arbitrary fixed symmetric matrix. After deriving an expression for $\mathbb{E}(QBQ)$ and considering some special cases, a numerical example is used to show how these results can support the comparison of two stochastic methods.

Key words: Wishart distribution, quadratic form, expected value, second momentum, stochastic gradient method, averaging

1. Outline

The Wishart distribution is a generalization of the χ^2 distribution. According to [6] and [7] the Wishart distribution plays a prominent role in estimating the covariance matrix in context of multivariate statistics. Therefore, it is not surprising that this important distribution is subject of current research. For instance, [10] considers quadratic forms $Y^{\mathrm{T}}CY$ with non-negative definite matrix C and normally distributed random matrix Y and investigates what are the necessary and sufficient conditions for $Y^{\mathrm{T}}CY$ to be Wishart distributed. Based on this, [9] examines in the special case of Y with expected value zero under which conditions $Y^{\mathrm{T}}CY$ is central Wishart distributed. Furthermore, in [11] the dispersion matrix of $\mathrm{vec}(Y^{\mathrm{T}}AY)$ is derived, where A is an arbitrary nonrandom matrix.

In this paper we are interested in a different kind of quadratic form: For a Wishart distributed Q and a symmetric matrix B we derive an expression for the expected value of QBQ. For $B = I_n$ this is the second momentum of the Wishart distribution and thus part of the

examination of the momenta of the Wishart distribution in [1].

In [8] a different and more general formula for the expected value of $XAX^{\mathrm{T}} \otimes XBX^{\mathrm{T}}$ was already derived, where $X \sim N_{n,k}(\mu, \Sigma, \Psi)$ with symmetric positive definite matrices Σ and Ψ. While the formulation in [8] is mathematically equivalent to the one derived in this paper for the special case considered here, the actual computation of the expected value is quite different - using a linear system based on Kronecker products in [8] and a lower dimensional variant below.

This paper is structured as follows. First, in chapter 3., we motivate in the context of the stochastic gradient method why an expression for the expected value $\mathbb{E}(QBQ)$ is needed. In the 4. chapter we recall two important, well-known properties of Wishart distributed random matrices. With this preparation, we are then able to present a theorem with a general expression for $\mathbb{E}(QBQ)$ in the 5. chapter, prove the assertion, and derive more compact expressions under stronger assumptions. Also the connection to the result of [8] is worked out in more detail. Finally, we show in chapter 6. that the approximated value for increasing sample size approaches the theoretical value from the previous chapter, which illustrates the statement of the theorem, and use the theorem to compare the ordinary stochastic gradient method with a variant that uses averaging.

An application of the expected value $\mathbb{E}(QBQ)$ derived here is the minimization of a random convex quadratic function. While convex quadratic problems in some form are the simplest nontrivial problems, they are complex enough to reproduce local dynamics of more difficult smooth problems. They arise in practical applications in the form of large scale systems of linear equations and least squares problems. Studying the performance of a method on convex quadratic problems is a fundamental preparation to extend the method to more general problems (see [3] and [2]).

2. Notation

In this paper the all-one and the all-zero vectors and matrices are denoted by

$$0_n := (0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^n, \ 0_{n \times n} := 0_n 0_n^{\mathrm{T}} \in \mathbb{R}^{n \times n},$$
$$\mathbb{1}_n := (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n, \ \mathbb{1}_{n \times n} := \mathbb{1}_n \mathbb{1}_n^{\mathrm{T}} \in \mathbb{R}^{n \times n}$$

and the identity matrix by I_n with dimension $n \in \mathbb{N}$. The Hadamard product of two matrices X and Y of the same dimension is defined componentwise as $(X \circ Y)_{ij} := X_{ij} Y_{ij}$. Let $X \otimes Y$ be the Kronecker product of two arbitrary matrices X and Y. A matrix M has rank $\mathrm{rk}\,(M)$, determinant $\det(M)$ and trace $\mathrm{tr}\,(M) := \sum_{i=1}^n M_{i,i}$. If a matrix M is positive definite, we write $M \succ 0$. The vector with the diagonal elements of a quadratic matrix M is denoted by $\mathrm{diag}(M)$ and for a vector $x \in \mathbb{R}^n$ the expression $\mathrm{Diag}(x)$ symbolizes the $n \times n$ diagonal matrix with the entries of x on its diagonal. The vector $\mathrm{vec}(M)$ is obtained by stacking the columns

of M on top of one another. The inverse function of vec is mat := vec^{-1}. Furthermore, S_+^n denotes the set of all symmetric, positive definite matrices, i. e.

$$S_+^n := \{M \in \mathbb{R}^{n \times n} \mid M = M^{\mathrm{T}}, \ M \succ 0\}. \tag{1}$$

The expected value and the covariance matrix of a random vector X are denoted by $\mathbb{E}(X)$ and $\mathrm{Cov}(X)$ whenever they exist.

3. Motivation

3.1 Stochastic Gradient Method

In order to find the minimum of a function $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) := \frac{1}{m} \sum_{i=1}^m f_i(x)$, i. e. the root of ∇f, the gradient descent can be used whose iterates $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$ are generated with step length $\gamma_k \geq 0$ starting at a point x^0. If m is very large, the calculation of the exact gradient $\nabla f(x^k)$ is computationally expensive. To avoid computing the full gradient $\nabla f(x^k)$ at each iteration, it can be approximated. Assuming an i. i. d. chosen batch S_k from the uniform distribution of $\{1, \ldots, m\}$, the expected value of $\nabla_{S_k} f(x^k) := \frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x^k)$ is

$$
\begin{aligned}
E(\nabla_{S_k} f(x^k)) &= E\left(\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x^k)\right) = \frac{1}{|S_k|} \sum_{i \in S_k} E(\nabla f_i(x^k)) \\
&= \frac{1}{|S_k|} \sum_{i \in S_k} \sum_{j=1}^m \nabla f_j(x^k) \cdot \mathrm{P}(\nabla f_j(x^k) = \nabla f_i(x^k)) = \frac{1}{|S_k|} \sum_{i \in S_k} \frac{1}{m} \sum_{j=1}^m \nabla f_j(x^k) \\
&= \frac{1}{|S_k|} \cdot |S_k| \cdot \nabla f(x^k) = \nabla f(x^k).
\end{aligned}
$$

This is the motivation for using the approximation $\nabla_{S_k} f(x^k)$ instead of $\nabla f(x^k)$, i. e. the stochastic gradient (descent) method (SGD) is given by the iterates

$$x^{k+1} = x^k - \gamma_k \nabla_{S_k} f(x^k). \tag{2}$$

There exist numerous modifications of the stochastic gradient method. The following example can be useful to examine the limits of a SGD method or to compare two variants of SGD.

3.2 Random quadratic functions

As in [5], we assume a fixed matrix $A \in \mathbb{R}^{n \times n}$ with $\det(A) \neq 0$ and $n \in \mathbb{N}$. At each iteration $\ell \in \{1, \ldots, m\}$ we draw random vectors r^ℓ and b^ℓ independently from the n-variate normal distribution with expected value 0_n and covariance matrix $\Sigma = \Sigma^{\mathrm{T}} \succ 0$. Briefly, this can be written as $r^\ell, b^\ell \sim N_n(0_n, \Sigma)$. With $a^\ell := A r^\ell$ we are able to define the functions

$$f_\ell : \ \mathbb{R}^n \to \mathbb{R}, \ f_\ell(x) := \tfrac{1}{2}((a^\ell)^{\mathrm{T}} x)^2 + (b^\ell)^{\mathrm{T}} x. \tag{3}$$

Since a^ℓ has the expected value $\mathbb{E}(a^\ell) = \mathbb{E}(Ar^\ell) = A\mathbb{E}(r^\ell) = 0_n$ and the covariance matrix $\mathrm{Cov}(a^\ell) = \mathrm{Cov}(Ar^\ell) = A\mathrm{Cov}(r^\ell)A^\mathrm{T} = A\Sigma A^\mathrm{T}$, it holds $a^\ell \sim N_n(0_n, A\Sigma A^\mathrm{T})$ and the second momentum is given by $\mathbb{E}(a^\ell(a^\ell)^\mathrm{T}) = \mathrm{Cov}(a^\ell) + \mathbb{E}(a^\ell)\mathbb{E}(a^\ell)^\mathrm{T} = \mathrm{Cov}(a^\ell) = A\Sigma A^\mathrm{T}$. Because f_ℓ are quadratic functions of r^ℓ and b^ℓ with expected value

$$\mathbb{E}(f_\ell(x)) = \tfrac{1}{2}\mathbb{E}(x^\mathrm{T}a^\ell(a^\ell)^\mathrm{T}x) + x^\mathrm{T}\mathbb{E}(b^\ell) = \tfrac{1}{2}x^\mathrm{T}\mathbb{E}(a^\ell(a^\ell)^\mathrm{T})x = \tfrac{1}{2}x^T A\Sigma A^\mathrm{T}x =: f(x)$$

and due to the existence of the fourth momenta of r^ℓ and b^ℓ, for a given x the variances of $f_\ell(x)$ are bounded and almost surely it exists

$$\lim_{m\to\infty} \sum_{\ell=1}^{m} f_\ell(x) = f(x). \tag{4}$$

The stochastic gradient method uses the approximation $\nabla f_\ell(x)$ instead of the full gradient $\nabla f(x)$. Therefore it is reasonable to examine the noise $\boldsymbol{\xi}^\ell$ defined by

$$\boldsymbol{\xi}^\ell := \nabla f_\ell(x) - \nabla f(x) = a^\ell(a^\ell)^\mathrm{T}x + b^\ell - A\Sigma A^\mathrm{T}x \tag{5}$$

with expected value

$$\mathbb{E}(\boldsymbol{\xi}^\ell) = \mathbb{E}(a^\ell(a^\ell)^\mathrm{T})x + \mathbb{E}(b^\ell) - A\Sigma A^\mathrm{T}x = A\Sigma A^\mathrm{T}x + 0_n - A\Sigma A^\mathrm{T}x = 0_n.$$

Using the independence of a^ℓ and b^ℓ and defining $B := A^\mathrm{T}xx^\mathrm{T}A$, the covariance matrix of $\boldsymbol{\xi}^\ell$ can be written as

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}^\ell) &= \mathbb{E}(\boldsymbol{\xi}^\ell(\boldsymbol{\xi}^\ell)^\mathrm{T}) - \mathbb{E}(\boldsymbol{\xi}^\ell)\mathbb{E}(\boldsymbol{\xi}^\ell)^\mathrm{T} = \mathbb{E}(\boldsymbol{\xi}^\ell(\boldsymbol{\xi}^\ell)^\mathrm{T}) \\
&= \mathbb{E}((a^\ell(a^\ell)^\mathrm{T}x + b^\ell - A\Sigma A^\mathrm{T}x)(a^\ell(a^\ell)^\mathrm{T}x + b^\ell - A\Sigma A^\mathrm{T}x)^\mathrm{T}) \\
&= \mathbb{E}(a^\ell(a^\ell)^\mathrm{T}xx^\mathrm{T}a^\ell(a^\ell)^\mathrm{T}) + \mathbb{E}(a^\ell(a^\ell)^\mathrm{T}x(b^\ell)^\mathrm{T}) - \mathbb{E}(a^\ell(a^\ell)^\mathrm{T}xx^\mathrm{T}A\Sigma A^\mathrm{T}) \\
&\quad + \mathbb{E}(b^\ell x^\mathrm{T}a^\ell(a^\ell)^\mathrm{T}) + \mathbb{E}(b^\ell(b^\ell)^\mathrm{T}) - \mathbb{E}(b^\ell x^\mathrm{T}A\Sigma A^\mathrm{T}) \\
&\quad - \mathbb{E}(A\Sigma A^\mathrm{T}xx^\mathrm{T}a^\ell(a^\ell)^\mathrm{T}) - \mathbb{E}(A\Sigma A^\mathrm{T}x(b^\ell)^\mathrm{T}) + \mathbb{E}(A\Sigma A^\mathrm{T}xx^\mathrm{T}A\Sigma A^\mathrm{T}) \\
&= \mathbb{E}(Ar^\ell(r^\ell)^\mathrm{T}A^\mathrm{T}xx^\mathrm{T}Ar^\ell(r^\ell)^\mathrm{T}A^\mathrm{T}) + \mathbb{E}(a^\ell(a^\ell)^\mathrm{T}x)\mathbb{E}(b^\ell)^\mathrm{T} - \mathbb{E}(a^\ell(a^\ell)^\mathrm{T})xx^\mathrm{T}A\Sigma A^\mathrm{T} \\
&\quad + \mathbb{E}(b^\ell)\mathbb{E}(x^\mathrm{T}a^\ell(a^\ell)^\mathrm{T}) + \big[\mathrm{Cov}(b^\ell) + \mathbb{E}(b^\ell)\mathbb{E}(b^\ell)^\mathrm{T}\big] - \mathbb{E}(b^\ell)x^\mathrm{T}A\Sigma A^\mathrm{T} \\
&\quad - A\Sigma A^\mathrm{T}xx^\mathrm{T}\mathbb{E}(a^\ell(a^\ell)^\mathrm{T}) - A\Sigma A^\mathrm{T}x\mathbb{E}(b^\ell)^\mathrm{T} + A\Sigma A^\mathrm{T}xx^\mathrm{T}A\Sigma A^\mathrm{T} \\
&= A\mathbb{E}(r^\ell(r^\ell)^\mathrm{T}Br^\ell(r^\ell)^\mathrm{T})A^\mathrm{T} + 0_{n\times n} - A\Sigma A^\mathrm{T}xx^\mathrm{T}A\Sigma A^\mathrm{T} + 0_{n\times n} + \Sigma - 0_{n\times n} \\
&\quad - A\Sigma A^\mathrm{T}xx^\mathrm{T}A\Sigma A^\mathrm{T} - 0_{n\times n} + A\Sigma A^\mathrm{T}xx^\mathrm{T}A\Sigma A^\mathrm{T} \\
&= A\mathbb{E}(r^\ell(r^\ell)^\mathrm{T}Br^\ell(r^\ell)^\mathrm{T})A^\mathrm{T} + \Sigma - A\Sigma B\Sigma A^\mathrm{T}. \tag{6}
\end{aligned}$$

This motivates us to determine expected values of the form $\mathbb{E}(r^\ell(r^\ell)^\mathrm{T}Br^\ell(r^\ell)^\mathrm{T})$ with random vectors $r^\ell \sim N_n(0_n, \Sigma)$ and a symmetric matrix B.

4. Introduction

In the context of multivariate statistics, the following definition is of great importance:

Definition 1 *For $k \in \mathbb{N}$ independent and identically distributed (i. i. d.) random vectors $r^\ell \sim N_n(0_n, \Sigma)$ with covariance matrix $\Sigma = \Sigma^T \succ 0$ the random matrix $Q := \sum_{\ell=1}^{k} r^\ell (r^\ell)^T$ is called n-variate Wishart distributed with scale matrix Σ and k degrees of freedom. We write $Q \sim W_n(\Sigma, k)$.*

Remark 1 *Alternatively $Q \sim W_n(\Sigma, k)$ can be defined by $Q := RR^T$, where R is a random $n \times k$ matrix, which columns r^ℓ are independent and identically $N_n(0_n, \Sigma)$ distributed. This coincides exactly with Definition 1.*

The following result is well known:

Lemma 1 *The expected value of $Q \sim W_n(\Sigma, k)$ is*

$$\mathbb{E}(Q) = k\Sigma.$$

In the proof of the main result, the following important property of Wishart distributed random matrices will also be needed. For the sake of completeness it is proved here.

Lemma 2 *Consider $Q \sim W_n(\Sigma, k)$ with $\Sigma \in S_+^n$, $C \in \mathbb{R}^{n \times m}$ and $rk(C) = m \in \mathbb{N}$. Then*

$$C^T Q C \sim W_m(C^T \Sigma C, k).$$

Proof. Since we assume $Q \sim W_n(\Sigma, k)$, there exist k independent and identically distributed $r^\ell \sim N_n(0_n, \Sigma)$ so that $Q = \sum_{\ell=1}^{k} r^\ell (r^\ell)^T$. Consider $C^T r^\ell$ with the expected value $\mathbb{E}(C^T r^\ell) = C^T \mathbb{E}(r^\ell) = 0_m$ and positive definite $\mathrm{Cov}(C^T r^\ell) = C^T \mathrm{Cov}(r^\ell)C = C^T \Sigma C$ due to the full rank of C, thus $C^T r^\ell \sim N_m(0_m, C^T \Sigma C)$. Then we are able to conclude that $C^T Q C = C^T \left(\sum_{\ell=1}^{k} r^\ell (r^\ell)^T \right) C = \sum_{\ell=1}^{k} (C^T r^\ell)(C^T r^\ell)^T \sim W_m(C^T \Sigma C, k)$. $\qquad\square$

5. Determination of $\mathbb{E}(QBQ)$

The following theorem is the main result of this paper:

Theorem 1 *For $n \in \mathbb{N}$ we consider the n-variate Wishart distributed random matrix Q with the symmetric, positive definite scale matrix Σ and $k \in \mathbb{N}$ degrees of freedom, i. e. $Q \sim W_n(\Sigma, k)$. Furthermore, we deal with a fixed matrix $B = B^T \in \mathbb{R}^{n \times n}$. The expected value of the quadratic form QBQ is*

$$\mathbb{E}(QBQ) = k \cdot tr(B\Sigma)\Sigma + (k^2 + k)\Sigma B \Sigma.$$

Proof. Let the matrix $\Sigma = \Sigma^{\mathrm{T}} \succ 0$ be factorized as $\Sigma = UDU^{\mathrm{T}}$ with diagonal matrix D and orthogonal matrix U, i. e. $UU^{\mathrm{T}} = I_n$. Since $Q \sim W_n(\Sigma, k)$, there exist i. i. d. random vectors $r^{\ell} \sim N_n(0_n, \Sigma)$ so that $Q = \sum_{\ell=1}^{k} r^{\ell}(r^{\ell})^{\mathrm{T}} = \sum_{\ell=1}^{k} Q_{\ell}$ with $Q_{\ell} := r^{\ell}(r^{\ell})^{\mathrm{T}} \sim W_n(\Sigma, 1)$. We note that Lemma 1 gives us $\mathbb{E}(Q_{\ell}) = 1 \cdot \Sigma = \Sigma$. Now we define the unitary transformations

$$\tilde{B} := U^{\mathrm{T}} B U, \quad \tilde{r}^{\ell} := U^{\mathrm{T}} r^{\ell} \quad \text{and} \quad \tilde{Q}_{\ell} := U^{\mathrm{T}} Q_{\ell} U.$$

Since $\tilde{r}^{\ell} \sim N_n(0_n, D)$, its components $\tilde{r}_i^{\ell} \sim N(0, D_{i,i})$ have the momenta $\mathbb{E}(\tilde{r}_i^{\ell}) = 0$, $\mathbb{E}((\tilde{r}_i^{\ell})^2) = D_{i,i}$ and $\mathbb{E}((\tilde{r}_i^{\ell})^4) = 3D_{i,i}^2$, and due to Lemma 2 we find

$$\tilde{Q}_{\ell} = U^{\mathrm{T}} Q_{\ell} U = U^{\mathrm{T}} r^{\ell}(r^{\ell})^{\mathrm{T}} U = U^{\mathrm{T}} r^{\ell}(U^{\mathrm{T}} r^{\ell})^{\mathrm{T}} = \tilde{r}^{\ell}(\tilde{r}^{\ell})^{\mathrm{T}} \sim W_n(D, 1).$$

The expected value of $\tilde{Q}_{\ell}\tilde{B}\tilde{Q}_{\ell}$ is componentwise given by

$$\mathbb{E}(\tilde{Q}_{\ell}\tilde{B}\tilde{Q}_{\ell})_{i,j} = \mathbb{E}(e_i^{\mathrm{T}}\tilde{r}^{\ell}(\tilde{r}^{\ell})^{\mathrm{T}}\tilde{B}\tilde{r}^{\ell}(\tilde{r}^{\ell})^{\mathrm{T}}e_j) = \mathbb{E}(\tilde{r}_i^{\ell}\tilde{r}_j^{\ell}(\tilde{r}^{\ell})^{\mathrm{T}}\tilde{B}\tilde{r}^{\ell}) = \mathbb{E}\left(\tilde{r}_i^{\ell}\tilde{r}_j^{\ell}\sum_{p,q=1}^{n}\tilde{B}_{p,q}\tilde{r}_p^{\ell}\tilde{r}_q^{\ell}\right)$$

$$= \sum_{p,q=1}^{n}\tilde{B}_{p,q}\mathbb{E}(\tilde{r}_i^{\ell}\tilde{r}_j^{\ell}\tilde{r}_p^{\ell}\tilde{r}_q^{\ell})$$

$$= \sum_{p,q=1}^{n}\tilde{B}_{p,q}\begin{cases} 3D_{i,i}^2 & \text{if } i = j = p = q \\ D_{i,i}D_{p,p} & \text{if } i = j \neq p = q \\ D_{i,i}D_{j,j} & \text{if } i \neq j \text{ and } ((i = p, j = q) \text{ or } (i = q, j = p)) \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} 3D_{i,i}^2\tilde{B}_{i,i} + D_{i,i}\sum_{p=1,p\neq i}^{n}\tilde{B}_{p,p}D_{p,p} & \text{if } i = j \\ 2D_{i,i}D_{j,j}\tilde{B}_{i,j} & \text{if } i \neq j \end{cases}$$

$$= \begin{cases} 2D_{i,i}^2\tilde{B}_{i,i} + D_{i,i}\mathrm{tr}\,(\tilde{B}D) & \text{if } i = j \\ 2D_{i,i}D_{j,j}\tilde{B}_{i,j} & \text{if } i \neq j \end{cases}$$

$$= \left[2(\mathrm{diag}(D)\mathrm{diag}(D)^{\mathrm{T}}) \circ \tilde{B} + \mathrm{tr}\,(\tilde{B}D)D\right]_{i,j}.$$

All in all, we obtain

$$\mathbb{E}(QBQ) = \mathbb{E}\left(\left(\sum_{\ell=1}^{k}Q_{\ell}\right)B\left(\sum_{h=1}^{k}Q_h\right)\right) = \sum_{\ell,h=1}^{k}\mathbb{E}(Q_{\ell}BQ_h)$$

$$= \sum_{\ell=1}^{k}\mathbb{E}(Q_{\ell}BQ_{\ell}) + \sum_{\ell,h=1,\ell\neq h}^{k}\mathbb{E}(Q_{\ell}BQ_h)$$

$$= \sum_{\ell=1}^{k}\mathbb{E}(UU^{\mathrm{T}}Q_{\ell}UU^{\mathrm{T}}BUU^{\mathrm{T}}Q_{\ell}UU^{\mathrm{T}}) + \sum_{\ell,h=1,\ell\neq h}^{k}\mathbb{E}(Q_{\ell})B\mathbb{E}(Q_h)$$

$$= \sum_{\ell=1}^{k}U\mathbb{E}(\tilde{Q}_{\ell}\tilde{B}\tilde{Q}_{\ell})U^{\mathrm{T}} + \sum_{\ell,h=1,\ell\neq h}^{k}\Sigma B\Sigma$$

$$= kU\left[2(\mathrm{diag}(D)\mathrm{diag}(D)^{\mathrm{T}}) \circ \tilde{B} + \mathrm{tr}\,(\tilde{B}D)D\right]U^T + (k^2 - k)\Sigma B\Sigma.$$

With $\tilde{B} = U^{\mathrm{T}}BU$ as defined above one gets

$$
\begin{aligned}
\mathbb{E}(QBQ) &= 2kU(D\tilde{B}D)U^{\mathrm{T}} + k \cdot \mathrm{tr}\,(\tilde{B}D)UDU^{\mathrm{T}} + (k^2 - k)\Sigma B\Sigma \\
&= 2k(UDU^{\mathrm{T}})B(UDU^{\mathrm{T}}) + k \cdot \mathrm{tr}\,(U^{\mathrm{T}}BUD)\Sigma + (k^2 - k)\Sigma B\Sigma \\
&= 2k\Sigma B\Sigma + k \cdot \mathrm{tr}\,(BUDU^{\mathrm{T}})\Sigma + (k^2 - k)\Sigma B\Sigma = k \cdot \mathrm{tr}\,(B\Sigma)\Sigma + (k^2 + k)\Sigma B\Sigma,
\end{aligned}
$$

which corresponds exactly to the assertion. $\qquad\square$

In practice, the following special cases might be of interest. Additionally to the assumptions of Theorem 1 we demand $k = 1$. Then the result of the theorem simplifies to

$$
\mathbb{E}(QBQ) = \mathrm{tr}\,(B\Sigma)\Sigma + 2\Sigma B\Sigma \tag{7}
$$

If we assume in addition that $\Sigma = \sigma^2 I_n$, we get

$$
\mathbb{E}(QBQ) = \sigma^4 \left[2B + \mathrm{tr}\,(B)I_n \right]. \tag{8}
$$

Finally, the assumption $\Sigma = I_n$ leads to the special case analyzed in paper [5]. With $B = I_n$ it follows immediately from Theorem 1:

Corollary 1 *The second momentum of a $W_n(\Sigma, k)$ distributed random matrix Q with scale matrix $\Sigma \in S_+^n$ is given by*

$$
\mathbb{E}(Q^2) = (k^2 + k)\Sigma^2 + tr\,(\Sigma)k\Sigma.
$$

Now we are able to present an expression for the covariance matrix of the noise (6) in which the random variables r^ℓ have been eliminated. Since $r^\ell \sim N_n(0_n, \Sigma)$ we can define the random matrix $Q^\ell := r^\ell(r^\ell)^{\mathrm{T}} \sim W_n(\Sigma, 1)$ and using (7) we conclude

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}^\ell) &= A\mathbb{E}(r^\ell(r^\ell)^{\mathrm{T}}Br^\ell(r^\ell)^{\mathrm{T}})A^{\mathrm{T}} + \Sigma - A\Sigma B\Sigma A^{\mathrm{T}} = A\mathbb{E}(Q^\ell BQ^\ell)A^{\mathrm{T}} + \Sigma - A\Sigma B\Sigma A^{\mathrm{T}} \\
&= AU\left[2\left(\mathrm{diag}(D)(\mathrm{diag}(D))^{\mathrm{T}}\right) \circ (U^{\mathrm{T}}BU) + \mathrm{tr}\,(U^{\mathrm{T}}BUD)D\right]U^{\mathrm{T}}A^{\mathrm{T}} \\
&\quad + \Sigma - A\Sigma B\Sigma A^{\mathrm{T}},
\end{aligned}
$$

where D and U satisfy $\Sigma = UDU^{\mathrm{T}}$ and $B = A^{\mathrm{T}}xx^{\mathrm{T}}A$.
In case of $\Sigma = I_n$ this expression can be simplified to

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}^\ell) &= A\left[2B + \mathrm{tr}\,(B)I_n\right]A^{\mathrm{T}} + I_n - ABA^{\mathrm{T}} = 2ABA^{\mathrm{T}} + \mathrm{tr}\,(B)AA^{\mathrm{T}} + I_n - ABA^{\mathrm{T}} \\
&= AA^{\mathrm{T}}xx^{\mathrm{T}}AA^{\mathrm{T}} + \mathrm{tr}\,(A^{\mathrm{T}}xx^{\mathrm{T}}A)AA^{\mathrm{T}} + I_n = AA^{\mathrm{T}}xx^{\mathrm{T}}AA^{\mathrm{T}} + \|A^{\mathrm{T}}x\|_2^2 AA^{\mathrm{T}} + I_n.
\end{aligned}
$$

As mentioned before, an alternative expression for $\mathbb{E}(QBQ)$ can be derived. If one chooses $A = I_k$, $B = I_k$, $\Psi = I_k$ and $\mu = 0_{k \cdot n}$ in Theorem 2.2.9 (ii) from [8], one obtains for $Q := XX^{\mathrm{T}}$

$$
\mathbb{E}(Q \otimes Q) = k^2 \cdot \Sigma \otimes \Sigma + k \cdot \mathrm{vec}(\Sigma)\mathrm{vec}(\Sigma)^{\mathrm{T}} + k \cdot K_{n,n} \cdot \Sigma \otimes \Sigma, \tag{9}
$$

where $K_{n,n}$ is the commutation matrix consisting of $n \times n$ blocks with $n \times n$ entries each. In the (i,j)-th block the only non-zero element is a "1" in position (j,i). It is a permutation matrix that can be used to describe the relationship between the vectorized forms of a square matrix A and its transpose, since $\text{vec}(A^{\mathrm{T}}) = K_{n,n}\text{vec}(A)$. Using the calculation rule $(A \otimes B)\text{vec}(V) = \text{vec}(BVA^{\mathrm{T}})$ for Kronecker products we get

$$\begin{aligned}
\mathbb{E}(QBQ) &= \text{mat}\left(\mathbb{E}(\text{vec}(QBQ))\right) = \text{mat}(\mathbb{E}((Q \otimes Q)\text{vec}(B))) = \text{mat}(\mathbb{E}(Q \otimes Q)\text{vec}(B)) \\
&= \text{mat}((k^2 \cdot \Sigma \otimes \Sigma + k \cdot \text{vec}(\Sigma)\text{vec}(\Sigma)^{\mathrm{T}} + k \cdot K_{n,n} \cdot \Sigma \otimes \Sigma)\text{vec}(B)) \\
&= \text{mat}(k^2 \cdot \text{vec}(\Sigma B \Sigma) + k \cdot \text{vec}(\Sigma)\text{vec}(\Sigma)^{\mathrm{T}}\text{vec}(B) + k \cdot K_{n,n} \cdot \text{vec}(\Sigma B \Sigma)) \\
&= \text{mat}(\text{vec}(k^2 \Sigma B \Sigma) + k \cdot \text{vec}(\Sigma)\text{sum}(\Sigma \circ B) + \text{vec}(k \Sigma B \Sigma)) \\
&= k^2 \Sigma B \Sigma + k \Sigma \cdot \text{sum}(\Sigma \circ B) + k \Sigma B \Sigma = k \cdot \text{tr}\,(B\Sigma)\Sigma + (k^2 + k)\Sigma B \Sigma,
\end{aligned}$$

where $\text{sum}(M)$ denotes the sum over all entries of a matrix M. Thus, we get the same expression as in Theorem 1.

6. Numerical examples

6.1 Illustrative example

For dimension $n = 10$ we generate randomly the matrices B and Σ with i.i.d. standard normally distributed entries and ensure, that B is symmetric and $\Sigma \in S_+^n$. Let $k = 3$ and $Q \sim W_n(\Sigma, k)$. The aim is to determine $\mathbb{E}(QBQ)$. With the diagonal matrix D and the orthogonal matrix U from an eigendecomposition $\Sigma = UDU^{\mathrm{T}}$ and with Theorem 1 we get on the one hand

$$E_{\text{exact}} := kU\left[2\left(\text{diag}(D)(\text{diag}(D))^{\mathrm{T}}\right) \circ (U^{\mathrm{T}}BU) + \text{tr}\,(U^{\mathrm{T}}BUD)D\right]U^{\mathrm{T}} + (k^2 - k)\Sigma B \Sigma.$$

On the other Hand, the expected value can be approximated with m realizations Q^i as

$$E_{\text{empiric}} := \frac{1}{m}\sum_{i=1}^{m}Q^iBQ^i, \tag{10}$$

because the law of large numbers provides

$$\lim_{m \to \infty}\frac{1}{m}\sum_{i=1}^{m}Q^iBQ^i = \mathbb{E}(QBQ) \quad \text{almost surely.}$$

To get an impression of how fast E_{empiric} approaches the theoretical value E_{exact} for increasing sample size m, we plot the relative error $\|E_{\text{exact}} - E_{\text{empiric}}\|_2/\|E_{\text{exact}}\|_2$ against m. Due to the randomness during the generation of B and Σ and in the realizations of Q, ten independent runs are made. At each run the relative error is calculated for $m \in \{1, 10, 100, 10^3, 10^4, 10^5, 10^6\}$. Thus, to be more precise, the logarithmic plot in figure 1 shows the arithmetic means of the relative errors in dependence of m.

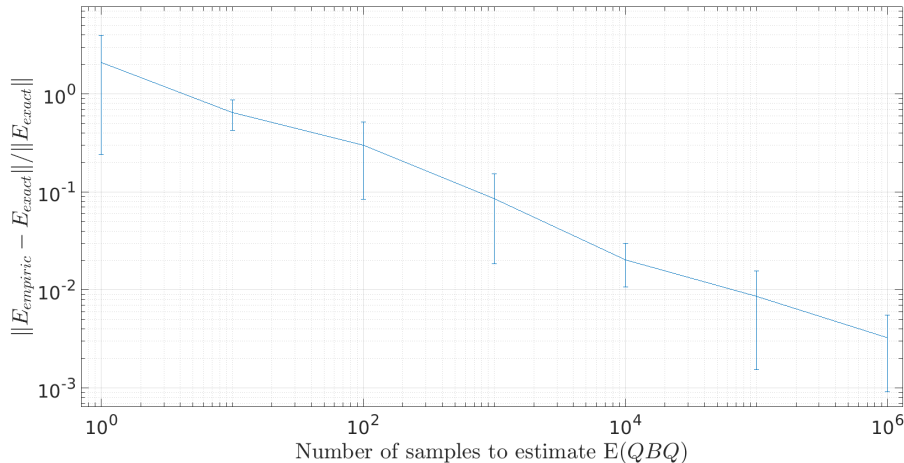The standard deviation is represented by error bars:

Figure 1: Relative error $\|E_{\text{exact}} - E_{\text{empiric}}\|_2 / \|E_{\text{exact}}\|_2$ with respect to the number of samples m used for the approximation E_{empiric}.

For $m = 1$ the mean distance between E_{exact} and E_{empiric} is about $4 \cdot 10^4$ which leads to a relative error of 2. Using 10^6 samples this distance reduces to approximately 90 and the relative error to $3 \cdot 10^{-3}$. The curve in the logarithmic plot is roughly linear decreasing. Two interesting observations arise: For large m the approximation E_{empiric} tends to the result of Theorem 1 and in order to approximate $\mathbb{E}(QBQ)$ adequately by (10) many samples and a lot of time is needed. Thus, the main result is not only theoretically fascinating but also of practical relevance.

6.2 Comparison of two SGD methods

Our actual goal, as mentioned before, is to compare two algorithms that approximate solutions for the problem $\min_x f(x)$ with $f(x) = \lim_{m \to \infty} \sum_{\ell=1}^{m} f_\ell(x)$ from (4). For $n = 10$ dimensions we randomly generate the entries of the matrices Σ and A i.i.d. from the $N(0,1)$ distribution and ensure that $\Sigma = \Sigma^{\mathrm{T}}$ is positive definite and that $A = A^T$ is positive semidefinite with norm $\|A\|_2 = 1$ and condition number $\text{cond}(A) = 5$.

As initial value $x^0 \in \mathbb{R}^n$ we choose the entries randomly from $N(0,1)$ and normalize the vector. Set the number of iterations to $k_{\max} = 10^7$ and let the step length be given by $\gamma \equiv \gamma_k = 10^{-3}$. Let $\{x^1, \ldots, x^{k_{\max}}\}$ be the iterates generated by the SGD method (2). A variation of the SGD method described above is the averaged SGD as analyzed in [12]. Starting with $\bar{x}^0 := x^0$ the iterates of the ASGD can be defined as

$$\bar{x}^k := \frac{1}{k} \sum_{\ell=1}^{k} x^\ell, \tag{11}$$

9

where x^ℓ are the iterates of the ordinary SGD method and $k \in \{1, \ldots, k_{\max}\}$. In each iteration we randomly draw r^k and b^k from $N_n(0_n, \Sigma)$, calculate $a^k = Ar^k$, the gradient of $f_k(x^k)$ defined in (3), the iterates x^k and $\bar{x}^k$ and the noise $\boldsymbol{\xi}^k$ from (5). The necessary condition for a minimum of f at x^* is that the gradient has to vanish, i.e. $\nabla f(x^*) = 0$. By construction the global optimal solution is $x^{\mathrm{opt}} = 0$. Below the two algorithms are compared by creating graphs of $\|\nabla f(x^k)\|_2$ and $\|x^k - x^{\mathrm{opt}}\|_2 = \|x^k\|_2$ in dependence of the number of iterations k, respectively.

Additionally, we are interested in using our insights about the noise $\boldsymbol{\xi}^k$ for this comparison. Since $E_{\mathrm{exact}} \equiv \mathbb{E}(\xi^k) = 0$ is valid independently of k, the approximation $E_{\mathrm{empiric}} := \frac{1}{k} \sum_{\ell=1}^{k} \boldsymbol{\xi}^\ell$ should tend to E_{exact}. This motivates plotting $\|E_{\mathrm{exact}} - E_{\mathrm{empiric}}\|_2$ with respect to the number of samples to estimate $E(\boldsymbol{\xi}^k)$.

On the other hand, $\mathrm{Cov}(\boldsymbol{\xi}^k)$ depends on k. At the optimal solution x^{opt} the covariance matrix of the noise is just the scale matrix Σ. We can use this to investigate, how the covariance matrices of the iterates, that can be calculated exactly using Theorem 1, approach to Σ with increasing number of iterations. Thus we plot $\|\mathrm{Cov}(\boldsymbol{\xi}^k) - \Sigma\|_2$ in dependence of the number of iterations k. This way we obtain the following four graphs:
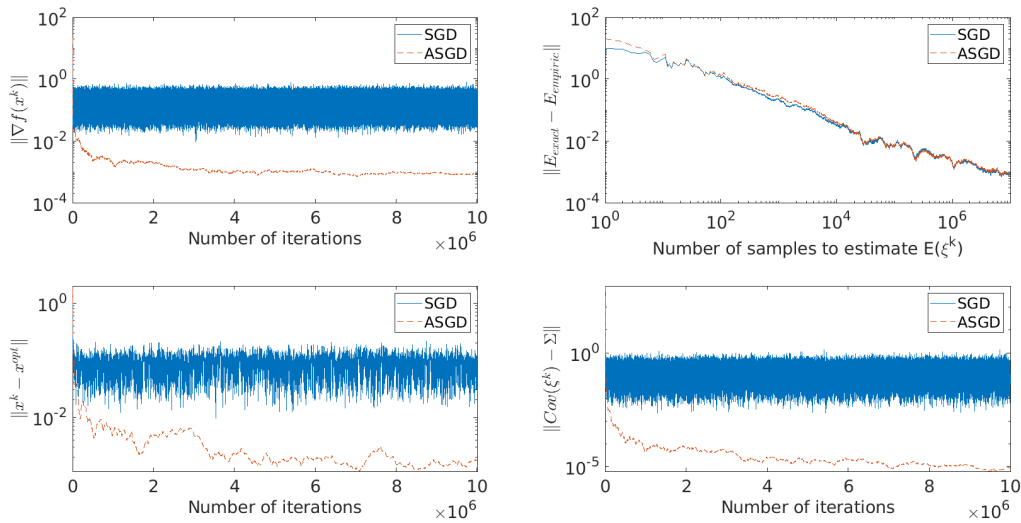

Figure 2: In all four graphs the iterates of the SGD method (blue, solid line) is compared to the iterates of the ASGD method (red, dashed line).

On the left hand side the norm of the gradient and the distance to the optimal solution are shown with respect to the number of iterations k. The ASGD method reaches lower values in both cases. In addition to that statistical fluctuations are much smaller.

At the top on the right hand side there is a plot of the distance of the empiric estimate of the expected value of the noise to the exact expected value in dependence of the number of samples. Both algorithms perform comparably well. Bottom right we have a plot of the distance of the covariance matrix of the noise in the kth iteration to the covariance matrix

at the optimal solution with respect to k. Again, the ASGD method performs better in both counts: by reaching lower values and by fluctuating less.

The advantages of the ASGD are not surprising and consistent with the results of [12]. This serves as a simple example of how two algorithms can be compared using Theorem 1.

7. Conclusion

In Theorem 1 it was proven that the expected value of the quadratic form QBQ with $Q \sim W_n(\Sigma, k)$ and $B = B^{\mathrm{T}}$ can be expressed using k, B, Σ and an eigendecomposition $\Sigma = UDU^{\mathrm{T}}$. Moreover, special cases for certain k, Σ and B were derived from this general formula, for instance the second momentum of a Wishart distributed random matrix Q, i.e. $\mathbb{E}(Q^2)$. A first example demonstrates the validity of the theorem. Beyond that the result is used to compare two stochastic methods.

Acknowledgment

References

[1] BISHOP, Adrian N. ; DEL MORAL, Pierre ; NICLAS, Angèle u. a.: An introduction to Wishart matrix moments. In: *Foundations and Trends® in Machine Learning* 11 (2018), Nr. 2, S. 97–218

[2] GOH, Gabriel: Why Momentum Really Works. In: *Distill* (2017). `http://dx.doi.org/10.23915/distill.00006`. – DOI 10.23915/distill.00006

[3] GONZAGA, Clóvis C ; SCHNEIDER, Ruana M.: On the steepest descent algorithm for quadratic functions. In: *Computational Optimization and Applications* 63 (2016), Nr. 2, S. 523–542

[4] HAGEDORN, Melinda: GitHub repository. In: *https://github.com/MHagedorn/wishart* (2022)

[5] HAGEDORN, Melinda ; JARRE, Florian: Optimized convergence of stochastic gradient descent by weighted averaging. In: *Preprint, https://optimization-online.org/2022/09 /optimized-convergence-of-stochastic-gradient-descent-by-weighted-averaging/* (2022)

[6] HÄRDLE, Wolfgang K. ; HLÁVKA, Zdeněk: *Multivariate statistics: exercises and solutions.* Springer, 2015

[7] KANTI, Mardia ; KENT, JT ; BIBBY, John: *Multivariate analysis.* 1979

[8] Kollo, Tonu ; Von Rosen, Dietrich: *Advanced multivariate statistics with matrices.* Springer, 2005

[9] Masaro, Joe ; Wong, Chi S.: Wishart distributions associated with matrix quadratic forms. In: *Journal of multivariate analysis* 85 (2003), Nr. 1, S. 1–9

[10] Mathew, Thomas ; Nordström, Kenneth: Wishart and chi-square distributions associated with matrix quadratic forms. In: *journal of multivariate analysis* 61 (1997), Nr. 1, S. 129–143

[11] Neudecker, H: On the dispersion matrix of a matrix quadratic form connected with the noncentral Wishart distribution. In: *Linear algebra and its applications* 70 (1985), S. 257–261

[12] Polyak, Boris T. ; Juditsky, Anatoli B.: Acceleration of stochastic approximation by averaging. In: *SIAM journal on control and optimization* 30 (1992), Nr. 4, S. 838–855