Project number IST-25582

CGL

Computational Geometric Learning

Matrix-valued Iterative Random Projections

STREP

Information Society Technologies

Period covered:	November 1, 2012–October 31, 2013
Date of preparation:	November 11, 2013
Date of revision:	November 13, 2013
Start date of project:	November 1, 2010
Duration:	3 years
Project coordinator name:	Joachim Giesen (FSU)
Project coordinator organisation:	Friedrich-Schiller-Universität Jena Jena, Germany

[*]The project CG Learning acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 255827.

[†]Institute of Theoretical Computer Science, ETH Zürich, stich@inf.ethz.ch

[§]Courant Institute of Mathematical Sciences, and Center for Genomics and Systems Biology, New York University, cm192@nyu.edu

[‡]Institute of Theoretical Computer Science, ETH Zürich, gaertner@inf.ethz.ch

Matrix-valued Iterative Random Projections*
Technical Report

Sebastian U. Stich[†] Christian L. Müller[§] Bernd Gärtner[‡]

November 13, 2013

Abstract

We analyze a matrix valued stochastic process with discrete time steps. The process originates from a randomized zeroth-order Hessian approximation scheme recently introduced by Leventhal and Lewis (D. Leventhal and A. S. Lewis., Optimization 60(3), 329–245, 2011). We do not only show convergence rates for summary statistics like trace and Frobenius norm but also study the expected convergence of the full eigenvalue spectrum.

1 Introduction

Consider the following random walk in $\mathbb{R}^n$. The random walk starts at an arbitrary $\mathbf{x}_0 \in \mathbb{R}^n$, and at each step k a uniformly random direction $\mathbf{u}_k \in S^{n-1}$ is chosen and x_k is projected on the hyperplane orthogonal to $\mathbf{u}_k$. In formulas, the process can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \langle \mathbf{x}_k, \mathbf{u}_k \rangle \, \mathbf{u}_k \, . \tag{1}$$

Due to the fact that all the steps are projections, it is intuitively clear that the distance to the origin only decreases in each step and thus $\|x_k\|_2$ converges to zero.

In [9] we study this process in a general Hilbert space. We consider also some variations and applications. Especially, we exemplify that this process arises in several algorithms in the field of derivative-free optimization.

In this report we are going to study one of those applications. Namely we consider a random process that defined on the space of symmetric matrices analogously to (1). For $X_k \in \mathbb{R}^{n \times n}$ symmetric, a step of the random walk is defined as:

$$X_{k+1} := X_k - (X_k \bullet U_k) U_k \, , \tag{2}$$

where $U_k = \mathbf{u}_k \mathbf{u}_k^T$ for a uniform random direction $\mathbf{u}_k \in S^{n-1}$. The bullet notation $A \bullet B = \mathrm{Tr}[A^T B]$ denotes the standard scalar product for symmetric matrices A, B. For a given initial iterate $X_0 \in \mathbb{R}^{n \times n}$ symmetric, we call the sequence $\{X_k\}_{k \geq 0}$ of matrices obtained by this random walk the *Matrix Random Projection Scheme (MRP) starting at* X_0.

This process was introduced in a recent paper by Leventhal and Lewis [6] in the context of randomized optimization. They showed that the expected Frobenius norm $E[X_k]$ converges exponentially fast to zero.

In the first part of this paper, we will investigate the MRP in a pure theoretical context. In the subsequent sections of this introduction we summarize a few properties of the MRP and motivate our further analysis. In Section 2 we present the main results. Especially, we do not only derive exact expressions for the $\mathbb{E}[\|X_k\|]$ which was already previously know, but also the expected values of $\mathbb{E}[X_k]$, $\mathbb{E}[\mathrm{Tr}[X_k]]$ and $\mathbb{E}[\mathrm{Tr}[X_k]^2]$.

In Section 3, we make the link the application side. We present the same application as studied in the original paper by Leventhal and Lewis [6]. Namely, we show hot MRP can be used to estimate a unknown positive definite matrix H by performing only linear measurements on H. In the context of derivative-free optimization, this matrix H could for instance describe a quadratic model of an unknown objective function. A optimization scheme that estimates H and uses this information for further optimization can be regarded as a zeroth-order analogue of the well-known Newton method in convex optimization. We do not cover this application in full detail, but refer the interested reader also to the numerical studies performed in [10, 11].

The appendices A and B contain several small calculations and proofs that are not essential to the ideas developed in this paper. Appendix C contains all details and calculations needed to proof the main theorem of Section 2.

1.1 First observations

The matrices U_k in (2) follow a special distribution. Let $\mathbf{v} \sim \mathcal{N}(0, I_n)$ a normal vector. Then $V = \mathbf{v}\mathbf{v}^T$ defines a probability distribution on the space of symmetric positive definite $n \times n$ matrices. This distribution is known as *Wishart distribution with one degree of freedom* in the literature. Let $\mathbf{u} \sim S^{n-1}$ a uniform random unit vector. (Here we abuse the notation slightly, denoting by S^{n-1} the set of all unit vectors in $\mathbb{R}^n$ *and* the uniform distribution over this set.) Then the matrix $U = \mathbf{u}\mathbf{u}^T$ follows the same distribution as the normalized $V / \|V\|_F$. Due to lack of a common name for the restriction of the Wishart distribution to matrices of unit norm, we will just refer to this distribution as *norm-constrained Wishart distribution* in the remainder of this report.

We remark that there is a one-to-one correspondence between $n \times n$ rank one matrices and unit vectors in $\mathbb{R}^n$. Thus we will refer to the Wishart distribution equivalently either in terms of the matrices U, or in terms of the (uniformly distributed) unit vectors $\mathbf{u}$.

Let us state three simple observations concerning two consecutive elements of a MRP. The first two are due to Lewis and Leventhal [6].

Remark 1.1. *Let $U_k \in \mathbb{R}^{n\times n}$ with $\|U_k\|_F = 1$, let $X_k, X_{k+1} \in \mathbb{R}^{n\times n}$ satisfying* (2) *and let $P \in \mathbb{R}^{n\times n}$ unitary, i.e. $PP^T = I_n$. Then*

$$\begin{aligned} &(i) && X_{k+1} \text{ and } U_k \text{ are orthogonal,} \\ &(ii) && \|X_{k+1}\|_F \leq \|X_k\|_F \ , \\ &(iii) && PX_{k+1}P^T = PX_kP^T - (PX_kP^T \bullet PU_kP^T)PU_kP^T \ , \end{aligned}$$

i.e. the rotated matrices PX_kP^T and $PX_{k+1}P^T$ satisfy 2 as well. Note that if $U_k = \mathbf{u}\mathbf{u}^T$ is norm-constrained Wishart, then so is $PU_kP^T = (P\mathbf{u})(P\mathbf{u})^T$ by definition.

This remark suggests that one step of the MRP can be considered as a projection of X_k on the randomly chosen hyperplane $\{Y \mid Y \bullet U_k = 0\}$, similar to the analogous process (1) in the space $\mathbb{R}^n$. Consequently, it is clear the distance of the iterates to the origin can not grow between two succeeding steps and that the process is invariant to rotations of the space. Thus the proof of this remark should not come as a surprise and can be found in Appendix A.

1.2 Convergence in Frobenius Norm

Remark 1.1 (ii) in the previous subsection states that the the elements of a MRP are not diverging in terms of the Forbenius norm. It is known already from [6] that the Forbenius norm of the elements of a MRP is converging to zero. This result is summarized in part (ii) of the following Remark 1.2.

Remark 1.2. *Let $X_0 \in \mathbb{R}^{n\times n}$ symmetric, $\{U_k\}_{k\geq 0}$ a sequence of independent $n \times n$ norm-constrained*

Wishart distributed random matrices and $\{X_k\}_{k\geq 0}$ the corresponding MRP starting from X_0. Then

$$\text{(i)}\quad \mathbb{E}\left[\|X_{k+1}\|_F^2 \mid X_k\right] = \|X_k\|_F^2 - \frac{2\|X_k\|_F^2 + \mathrm{Tr}[X_k]^2}{n(n+2)}\,,$$

$$\text{(ii)}\quad \mathbb{E}[\|X_k\|_F^2] \leq \left(1 - \frac{2}{n(n+2)}\right)^k \|X_0\|_F^2\,,$$

$$\text{(iii)}\quad \|X_k\|_F^2 \leq \|X_0\|_F^2 \cdot \exp\left[-\sum_{i=0}^{k-1} \frac{(X_i \bullet U_i)^2}{\|X_i\|_F^2}\right]\,.$$

The proof of this remark can again be found in Appendix A. We would like to point out a few consequences:

Note that by part (iii), $\|X_k\|_F$ is *completely* determined by the sum $S_k := \sum_{i=0}^{k-1} \frac{(X_i \bullet U_i)^2}{\|X_i\|_F^2}$. The authors believe that this result could actually be useful in order to prove concentration bounds on the convergence behavior of the MRP. By establishing concentration bounds on the random variable S_k one could probably obtain bounds of better quality. However, we could not exploit this yet in this report.

Part (i) of the above remark describes the expected *one-step-progress*. The last term in part (i) can be bounded as follows:

$$\frac{2}{n(n+2)}\|X_k\|_F^2 \leq \frac{2\|X_k\|_F^2 + \mathrm{Tr}[X_k]^2}{n(n+2)} \leq \frac{1}{n}\|X_k\|_F^2\,, \tag{3}$$

where the first inequality is trivial and the second follows by Cauchy-Schwarz: $\|X_k\|_F^2 \leq \mathrm{Tr}[X_k]^2 \leq n\|X_k\|_F^2$. Both, the upper and lower bound, are tight in general but they are different by a factor of approximately n.

1.3 Motivation

Remark 1.2 (ii) summarizes the only previously known result about MPR. It comprises an upper bound on the expected convergence of the Forbenius norm. The upper bound on the expected progress after k steps (part (ii) of Remark 1.2) is obtained by repeatedly applying the lower bound of Equation (3). As the lower bound differs by a factor of n from the upper bound, we might suspect that the bound in part (ii) of Remark 1.2 is far off the truth and the true convergence rate could be slightly better.

For instance, if X_0 is a diagonal matrix, we would expect (from part (i) of Remark 1.2) that the expected one-step-progress in the very first few steps is very large (much larger than $\frac{\|X_0\|_F^2}{n^2}$), but after sufficiently many steps, the matrices X_k behave like (almost) "random" matrices independent of the initial matrix X_0, and the result of part (ii) is almost tight.

In the remainder of this report, we derive $\mathbb{E}[\|X_k\|_F^2]$ *exactly* and show that the above intuition is indeed correct.

2 Convergence of MRP

In this section we state an prove the main result of this paper. It will turn out handy, to fix the following parameters (depending on the dimension $n \geq 1$) for the remainder of this section:

$$\alpha = \frac{1}{n}\,, \qquad \alpha' = \frac{1}{n+1}\,, \qquad \beta = \frac{1}{n(n+2)}\,, \qquad \gamma = \frac{1}{n(n+2)(n+4)}\,, \qquad \delta = \frac{1}{n(n+4)}\,,$$

$$\lambda_1 = \frac{2n^2 + 2n - 5 - \omega}{2n(n+2)}\,, \qquad \lambda_2 = \frac{2n^2 + 2n - 5 + \omega}{2n(n+2)}\,,$$

with $\omega = \sqrt{4n^2 + 4n - 7}$. We recognize the term β which already appeared in Remark 1.2.

2.1 The main theorem

Theorem 2.1. *Let $X_0 \in \mathbb{R}^{n\times n}$ symmetric, $n \geq 1$, and $\{X_k\}_{k\geq 0}$ a MRP starting from X_0. Then*

$$\text{(i)} \quad \mathbb{E}\left[\operatorname{Tr}[X_k]\right] = (1-\alpha)^k \operatorname{Tr}[X_0]\,,$$

$$\text{(ii)} \quad \mathbb{E}\left[X_k\right] = (1-\alpha)^k \alpha \operatorname{Tr}[X_0] I_n + (1-2\beta)^k \left(X_0 - \alpha \operatorname{Tr}[X_0] I_n\right)\,,$$

$$\text{(iii)} \quad \mathbb{E}\left[\|X_k\|_F^2\right] = \left(\lambda_1^k + \lambda_2^k\right) \frac{\|X_0\|_F^2}{2} + \left(\lambda_2^k - \lambda_1^k\right) \cdot \left(\frac{(2n+1)\|X_0\|_F^2}{2\omega} - \frac{\operatorname{Tr}[X_0]^2}{\omega}\right)\,,$$

$$\text{(iv)} \quad \mathbb{E}\left[\operatorname{Tr}[X_k]^2\right] = \left(\lambda_1^k + \lambda_2^k\right) \frac{\operatorname{Tr}[X_0]^2}{2} + \left(\lambda_2^k - \lambda_1^k\right) \cdot \left(\frac{2\|X_0\|_F^2}{\omega} - \frac{(2n+1)\operatorname{Tr}[X_0]^2}{2\omega}\right)\,.$$

If in addition $n \geq 5$:

$$\text{(v)} \quad \mathbb{E}\left[X_k^2\right] \preceq \alpha \lambda_1^k \|X_0\|_F^2 I_n + \lambda_2^k \frac{4}{n-4} \|X_0\|_F^2 I_n + (1-3\delta)^k (\operatorname{Tr}[X_0]^2 I_n + X_0^2)\,.$$

Note that X_k is symmetric, so $X_k^2 = X_k X_k^T$. The notation $X \preceq Y$ for symmetric matrices X, Y means $\mathbf{z} X \mathbf{z}^T \leq \mathbf{z} Y \mathbf{z}^T$ for all $\mathbf{z} \in \mathbb{R}^n$. Note that the assumption $n \geq 5$ is not intrinsic to get an upper bound on $\mathbb{E}[X_k^2]$, but just allowed us to make some simplifications for presentation purposes.

Being now able to calculate $\mathbb{E}[X_k^2]$ instead of only $E[\|X_k\|_F^2]$ as in Remark 1.2, we can now extend our upper bound on $E[\|X_k\|_F^2]$ to products $E[\|AX_k\|_F^2]$, for A positive semidefinite.

Theorem 2.2. *Let X_0, $\{X_k\}_{k\geq 0}$ as in Theorem 2.1, $n \geq 5$, $\rho' = 1 + \frac{1}{n}$, and let $A \in \mathbb{R}^{n\times n}$ positive semidefinite. Then for $k \geq (n+2)\ln n$ it holds:*

$$\begin{aligned} \|AX_k\|_F^2 &\leq \left((1-2\beta)^k \frac{4\rho'}{n-4} + (1-3\delta)^k\right) \|X_0\|_F^2 \|A\|_F^2 + (1-3\delta)^k \operatorname{Tr}[X_0]^2 \|A\|_F^2 \\ &\leq (1-2\beta)^k \frac{10}{n-4} \|X_0\|_F^2 \|A\|_F^2 + (1-3\delta)^k \operatorname{Tr}[X_0]^2 \|A\|_F^2\,. \end{aligned}$$

2.2 Estimates and simplifications

Due to the definition of the factors λ_1, λ_2, the statement of Theorem 2.1 might be a little bit obscure at first sight. By establishing upper and lower bounds on these parameters, we can simplify statements (iii) and (iv) slightly. Let

$$\bar{\lambda}_1 = 1 - 2\alpha'\,, \qquad \underline{\lambda}_1 = 1 - 2\alpha\,, \qquad \bar{\lambda}_2 = 1 - 2\beta\,, \qquad \underline{\lambda}_2 = 1 - 2.5\beta\,.$$

By Lemma C.7 it hols $\underline{\lambda}_i \leq \lambda_i \leq \bar{\lambda}_i$ for $i = 1, 2$. Now the following corollary follows directly form Theorem 2.1.

Corollary 2.3. *Let $X_0 \in \mathbb{R}^{n\times n}$ symmetric, $n \geq 4$, and $\{X_k\}_{k\geq 0}$ a MRP starting from X_0 and parameter $\rho := 1 + \frac{1}{4n}$. Then*

$$\text{(i)} \quad \mathbb{E}[\|X_k\|_F^2] \leq \bar{\lambda}_2^k \left(\rho \|X_0\|_F^2 - \frac{\operatorname{Tr}[X_0]^2}{2n+1}\right) + \bar{\lambda}_1^k \frac{\operatorname{Tr}[X_0]^2}{2n}\,,$$

$$\text{(ii)} \quad \mathbb{E}[\|X_k\|_F^2] \geq \underline{\lambda}_2^k \left(\|X_0\|_F^2 - \frac{\operatorname{Tr}[X_0]^2}{2n}\right) - \bar{\lambda}_1^k \frac{\|X_0\|_F^2}{4n}\,,$$

$$\text{(iii)} \quad E[\operatorname{Tr}[X_k]^2] \leq \bar{\lambda}_1^k \left(\rho \operatorname{Tr}[X_0]^2 - \frac{2\|X_0\|_F^2}{2n+1}\right) + \bar{\lambda}_2^k \frac{\|X_0\|_F^2}{n}\,,$$

$$\text{(iv)} \quad E[\operatorname{Tr}[X_k]^2] \geq \underline{\lambda}_1^k \left(\operatorname{Tr}[X_0]^2 - \frac{2\|X_0\|_F^2}{2n+1}\right) - \bar{\lambda}_2^k \frac{\operatorname{Tr}[X_0]^2}{4n}\,.$$

Proof. Although being a consequence of Theorem 2.1, the estimates are detailed in Corollary C.5 in the appendix for the factors λ_1, λ_2. It remains to apply the upper and lower bounds to these factors, and noting that the four factors in the brackets in (i)-(iv) are positive (due to $\|X_0\|_F^2 \leq \mathrm{Tr}[X_0]^2 \leq n \|X_0\|_F^2$ by Cauchy-Schwarz). $\square$

The above statement gives still quite accurate descriptions of $\mathbb{E}[\|X_k\|_F^2]$ and $\mathbb{E}[\mathrm{Tr}[X_k]]$. However, we note that for k large enough, only the terms containing the λ_2 factors will significantly contribute to these bounds, as the λ_1 factors converge to zero much faster for $(k \to \infty)$. Thus if we are only interested in approximate asymptotic bounds, we can simply neglect these factors. Alternatively, we can derive a precise mathematical statement. For the following statement, we determine k' such that (approximately) $\lambda_1^{k'} \leq \frac{1}{n}\lambda_2^{k'}$, an the λ_2-terms safely "swallow" all the remaining λ_1-terms.

Corollary 2.4. *Let $X_0 \in \mathbb{R}^{n\times n}$ symmetric, $n \geq 5$, and $\{X_k\}_{k\geq 0}$ a MRP starting from X_0 and parameters $\rho' := 1 + \frac{1}{n}$, $\rho'' := 1 - \frac{1}{n}$. Then for $k \geq k' := (n+2)\ln n$, it holds:*

$$\textit{(i)} \quad \mathbb{E}[\|X_k\|_F^2] \leq (1-2\beta)^k \left(\rho' \|X_0\|_F^2 - \frac{\mathrm{Tr}[X_0]^2}{2n+1}\right),$$

$$\textit{(ii)} \quad \mathbb{E}[\|X_k\|_F^2] \geq \underline{\lambda}_2^k \left(\rho'' \|X_0\|_F^2 - \frac{\mathrm{Tr}[X_0]^2}{2n}\right),$$

$$\textit{(iii)} \quad \mathbb{E}\left[X_k^2\right] \preceq (1-2\beta)^k \frac{4\rho'}{n-4} \|X_0\|_F^2 I_n + (1-3\delta)^k (\mathrm{Tr}[X_0]^2 I_n + X_0^2).$$

Before skipping to the proof of this statement, let us remind the reader that we already have an upper bound on $\mathbb{E}[\|X_k\|_F^2]$, namely the one from Remark 1.2 from Section 1.2. Depending on the application, it might be easier to simply use this old bound.

Proof of Corollary 2.4. First we note that claim (ii) holds for all $k \geq 0$, as $\bar{\lambda}_1 \leq \underline{\lambda}_2$. Now we proceed to show (iii). First we observe by elementary calculation:

$$\frac{1-\alpha}{1-2\beta} \leq \left(1 - \frac{1}{n+2}\right).$$

Thus

$$\left(\frac{1-\alpha}{1-2\beta}\right)^k \leq \left(1 - \frac{1}{n+2}\right)^k \leq e^{-\frac{k}{n+2}} \leq \frac{1}{n},$$

where the last inequality follows by the choice of $k \geq k'$. Therefore $(1-\alpha)^k \leq \frac{1}{n}(1-2\beta)^k$ for $k \geq k'$, and (iii) follows by straightforward calculations.

It remains to show (i). With (i) form Corollary 2.3 we estimate

$$\mathbb{E}[\|X_k\|_F^2] \leq \left(\rho\bar{\lambda}_2^k + \frac{1}{2}\bar{\lambda}_1^k\right) \|X_0\|_F^2 - \bar{\lambda}_2^k \frac{\mathrm{Tr}[X_0]^2}{2n+1}. \tag{4}$$

We have to show that the term in the brackets can be upper bounded by $\rho'\bar{\lambda}_2^k$, for all $k \geq k'$. We observe:

$$\frac{\bar{\lambda}_1}{1-2\beta} \leq \left(1 - \frac{2}{n+2}\right).$$

Therefore, by our choice of k', it holds $\bar{\lambda_1}^k \leq \frac{1}{n^2}(1-2\beta)^k$ for $k \geq k'$ and the term in the brackets in (4) can indeed be upper bounded by $\rho'(1-2\beta)^k$. This shows the claim. $\square$

2.3 Proof of the main theorem

It remains to proof Theorem 2.1. We will proceed as follows. By conditioning on a fixed iterate X_k, we can derive exact formulas for the expectation of the considered function depending for X_{k+1}, $\mathbb{E}[X_{k+1} \mid X_k]$, say. We arrive at a recursive description of the considered expectations. However, the descriptions are typically not independent. By applying a well-known technique from linear algebra, we can decouple the recurrences.

Proof of Theorem 2.1. By Remark 1.1 (iii) from Section 1.1 we can without loss of generality assume that X_0 is a diagonal matrix.

Let us consider claims (i) and (ii). By definition of the MRP (2) and Lemma B.1 from the appendix, we calculate for $\mathbf{u} \sim S^{n-1}$:

$$\mathbb{E}[X_{k+1} \mid X_k] = \mathbb{E}_{\mathbf{u}}[X_k - (\mathbf{u}^T X_k \mathbf{u})\mathbf{u}\mathbf{u}^T] = (1-2\beta)X_k - \beta \mathrm{Tr}[X_k] I_n \,.$$

Consequently, $E[\mathrm{Tr}[X_{k+1}] \mid X_k] = \mathrm{Tr}[E[X_{k+1} \mid X_k] = (1-\alpha)\mathrm{Tr}[X_k]$. Note that if X_k was a diagonal matrix, then also $E[X_{k+1}]$ will be diagonal.

What we now have to do formally, is to condition on X_{k-1} and calculate the expectations again. By the tower property of conditional expectations, $\mathbb{E}[E[X_{k+1} \mid X_k] \mid X_{k-1}] = \mathbb{E}[X_{k+1} \mid X_{k-1}]$. Repeating this procedure for X_{k-2} up to X_0, we finally obtain a description of $E[X_{k+1} \mid X_0] = \mathbb{E}[X_{k+1}]$. We observe that all intermediate expressions only depend linearly on X_0 and $\mathrm{Tr}[X_0]$. By the fact that X_0 was diagonal, $E[X_{k+1}]$ will again be a diagonal matrix. Therefore we can identify each $E[X_k]$ with a vector in $\mathbb{R}^n$ by considering only the entries on the diagonal.

By linear algebra, we can now decouple the linear recurrence. This is carried out in detail in Corollary C.2 in the appendix. We arrive at the claimed expressions for $E[X_k]$ and $E[\mathrm{Tr}[X_k]]$.

We apply the same procedure to show (iii)-(v) simultaneously. For $\mathbf{u} \sim S^{n-1}$ and again by Lemma B.1 from Appendix B we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|X_{k+1}\|_F^2 \mid X_k\right] &= \mathbb{E}_{\mathbf{u}}\left[1 - (\mathbf{u}^T X_k \mathbf{u})^2\right] \|X_k\|_F^2 = (1-2\beta)\|X_k\|_F^2 - \beta \mathrm{Tr}[X_k]^2 \,, \\
\mathbb{E}\left[\mathrm{Tr}[X_{k+1}]^2 \mid X_k\right] &= E_{\mathbf{u}}\left[\mathrm{Tr}[X_k]^2 - 2\mathbf{u}^T X_k \mathbf{u} + (\mathbf{u}^T X_k \mathbf{u})^2\right] \\
&= (1-(2n-3)\beta)\mathrm{Tr}[X_k]^2 + 2\beta \|X_k\|_F^2 \,, \\
\mathbb{E}\left[X_{k+1}^2 \mid X_k\right] &= \mathbb{E}_{\mathbf{u}}\left[X_k^2 - (\mathbf{u}^T X_k \mathbf{u})\left(\mathbf{u}\mathbf{u}^T X_k + X_k \mathbf{u}\mathbf{u}^T\right) + (\mathbf{u}^T X_k \mathbf{u})^2 \mathbf{u}\mathbf{u}^T\right] \\
&= (1-4\delta)X_k^2 - 2\delta \mathrm{Tr}[X_k]X_k + \gamma \mathrm{Tr}[X_k]^2 I_n + 2\gamma \|X_k\|_F^2 I_n \,.
\end{aligned}$$

Note that $\mathbb{E}\left[X_{k+1}^2 \mid X_k\right]$ does not linearly depend solely on $\mathrm{Tr}[X_k]^2, \|X_k\|_F^2$, but it depends also on $\mathrm{Tr}[X_k]X_k$. But with a trick we can get back to a linear recurrence. Observe, that for X symmetric $n \times n$ matrix:

$$-2\mathrm{Tr}[X]X = \mathrm{Tr}[X]^2 I_n + X^2 I_n - (Tr[X]I_n + X)^2 \preceq Tr[X]^2 I_n + X^2 I_n \,.$$

Because we are only interested in an upper bound (in the $\preceq$-sense) on $E[X_{k+1}^2 \mid X_k]$, we can estimate:

$$\mathbb{E}\left[X_{k+1}^2 \mid X_k\right] \preceq (1-3\delta)X_k^2 + (\gamma+\delta)Tr[X_k]^2 I_n + 2\gamma \|X_k\|_F^2 I_n \,.$$

Now the proof follows by same technique as described earlier, the technicalities are carried out in Corollary C.5 in the appendix. □

Proof of Theorem 2.2. We observe

$$\mathbb{E}\left[\|AX_{k+1}\|_F^2 \mid X_k\right] = \mathbb{E}\left[\mathrm{Tr}[AX_k X_k^T A^T \mid X_k]\right] = \mathrm{Tr}\left[A\mathbb{E}[X_k X_k^T \mid X_k]A^T\right] \,.$$

Thus the theorem is a direct consequence of Corollary 2.4 (iii). □

2.4 Concentration Bounds

In the previous sections we calculated expectations of random variables, for instance $\mathbb{E}[\|X_k\|_F^2]$. In this section, we will use Markov's Inequality to derive some simple upper bounds that hold with high probability. We did a similar analysis for the random variables $\|X_k\|_F^2$ in [11] and repeat it here shortly.

Lemma 2.5. *Let $X_0 \in \mathbb{R}^{n\times n}$ symmetric, let $\{X_k\}_{k\geq 0}$ a MRP starting from X_0 and let $0 \leq K = n(n+2)\left(\ln(\|X_0\|_F \sqrt{b}c) + a\right)$ for parameters a, b, c. Then the following statement holds with probability at least $1 - \frac{1}{b}$:*

$$\|X_K\|_F^2 \leq e^{-a}c^{-1}\,.$$

Proof. By the choice of K and Remark 1.2 we have

$$\mathbb{E}\left[\|X_K\|_F^2\right] \leq (1-2\beta)^K \|X_0\|_F^2 \leq e^{-2\beta K} \|X_0\|_F^2 \leq e^{-2a}b^{-1}c^{-2}\,. \tag{5}$$

Thus we apply Markov's Inequaitly on the positive random variable $\|X_K\|_F^2$ to prove the lemma:

$$\mathbf{Pr}\left[\|X_K\|_F^2 \leq e^{-2a}/c\right] \leq e^{2a}c^2\mathbb{E}[\|X_K\|_F^2] \leq 1/b\,.$$ □

Now concentrate on $\mathbb{E}[X_k]$. We are dealing with matrix valued random variables. Therefore, let us recall the Operator Matrix Inequality [1, 8]. For X random positive semidefinite matrix, and A a fixed positive definite matrix, we have $\mathbf{Pr}[X \npreceq A] \leq \operatorname{Tr}\left[\mathbb{E}[X]A^{-1}\right]$, similar to the classical version of Markov's Inequality. Here the notation $X \npreceq Y$ for positive semidefinite matrices X, Y means that $Y - X$ is not positive semidefinite.

Lemma 2.6. *Let $X_0 \in \mathbb{R}^{n\times n}$ symmetric, let $\{X_k\}_{k\geq 0}$ a MRP starting from X_0 and let $0 \leq K = n(n+2)\left(\ln(\|X_0\|_F \sqrt{b}c) + a\right)$ for parameters a, b, c. Then each of the following two statements hold with probability at least $1 - \frac{1}{b}$:*

$$\text{(i)}\quad X_K \preceq E[X_K] + e^{-a}c^{-1}I_n\,, \text{ and } E[X_K] - e^{-a}c^{-1}I_n \preceq X_K\,,$$
$$\text{(ii)}\quad X_K \preceq e^{-a}c^{-1}I_n\,, \text{ and } X_K \succeq -e^{-a}c^{-1}I_n\,.$$

The notation $X \preceq Y$ for symmetric matrices X, Y means $\mathbf{z}X\mathbf{z}^T \leq \mathbf{z}Y\mathbf{z}^T$ for all $\mathbf{z} \in \mathbb{R}^n$.

Proof. To show the second claim, we apply the Operator Markov Inequality on the positive semidefinite random matrix $X_K X_K^T = (X_K)^2$:

$$\Pr\left[X_K^2 \npreceq e^{-2a}c^{-2}I_n\right] \leq e^{2a}c^2\mathbb{E}\left[\operatorname{Tr}[X_K^2]\right] = e^{2a}c^2\mathbb{E}\left[\|X_K\|_F^2\right] \leq 1/b\,,$$

where we have again used Equation (5). The Lemma now follows with Lemma A.1 from Appendix A.

To show the first claim, we apply the Operator Markov Inequality on the positive semidefinite random matrix $(X_K - E[X_K])^2$, and the Lemma follows because of

$$\begin{aligned}\Pr\left[(X_K - E[X_K])^2 \npreceq e^{-2a}c^{-2}I_n\right] &\leq e^{2a}c^2\left(\mathbb{E}\left[\operatorname{Tr}[X_K^2]\right] - \operatorname{Tr}[E[X_K]^2]\right)\\ &\leq e^{2a}c^2\mathbb{E}\left[\|X_K\|_F^2\right]\,,\end{aligned}$$

like in the first part of the proof. □

2.5 Numerical Demonstration

In order to verify the results of Theorem 2.1 numerically, we perform a small experiment. We will compare our bounds for the expected Frobenius norm of MRP iterates with the empirical results. We consider two different start matrices $X, Y \in \mathbb{R}^{50\times 50}$ in $n = 50$ dimensions, defined as:

$$X := \frac{1}{\sqrt{n}} \cdot I_n\,, \qquad \|X\|_F^2 = 1\,, \qquad \mathrm{Tr}[X]^2 = n$$
$$Y := \frac{1}{\sqrt{n(n-1)}} \cdot (J_n - I_n)\,, \qquad \|X\|_F^2 = 1\,, \qquad \mathrm{Tr}[Y]^2 = 0\,,$$

where J_n denotes the all-1 matrix. For each of these two matrices, we perform 20 independent runs of MRP up to $k = 300$ iterations. The empirical mean is shown in Figure 1.

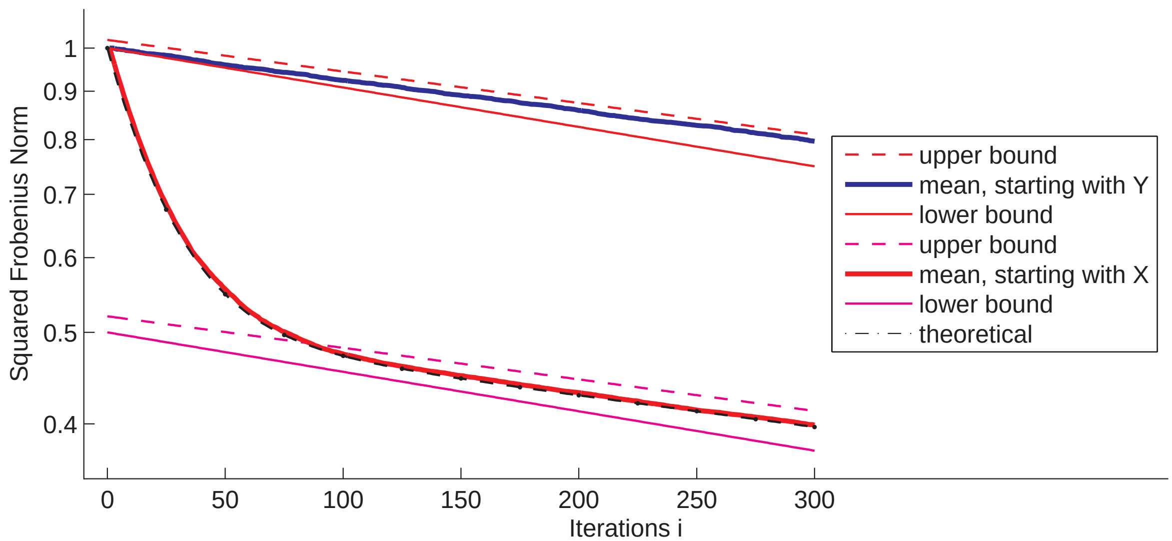

Figure 1: Dependency of MRP on the trace of the starting matrix. Squared Frobenius norm of the MRP matrix versus the number of iterations averaged over 20 independent runs of MRP in dimension $n = 50$. Blue: initial matrix X with zero trace. Red: initial matrix Y with positive trace. The theoretical bounds are shown as dashed an dotted lines. The black curve is the bound from Thm 2.1 (i). See main text for further information.

We observe that the empirical mean of the runs starting from X decrease much more rapidly than the mean for the runs starting from Y, in agreement with Theorem 2.1. In red an magenta we plotted the corresponding (theoretically expected) upper and lower bounds, as derived in Corollary 2.3. We see, that indeed for approximately $k = \frac{1}{2}n\log(n) \approx 100$, the mean of the MRP starting from X is in between the upper and lower bound. In black we also depicted the upper bound on the expectation from the more detailed Theorem 2.1 (i). Also this curve matches almost perfectly the observed mean.

3 An Application

We present an application of MRP. Leventhal and Lewis [6] showed that the MRP scheme can be used to estimate a *unknown* Hessian matrix H of a quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x}$. This estimate of the Hessian can then be used to improve the performance of optimization algorithms that try to minimize f. This concept is well-known in the optimization literature under the term *Variable Metric method*. We refer the interested reader to [6] or [11], where the optimization algorithm is discussed in detail.

The *Randomized Hessian Approximation (RHA)* scheme as introduced in [6] is a random walk on the space of symmetric $n \times n$ matrices, depending on a (fixed) positive definite matrix H. For $Y_0 \in \mathbb{R}^{n\times n}$

symmetric, a step of the random walk is defined as:

$$Y_{k+1} := Y_k - ((Y_k - H) \bullet U_k)U_k \,, \tag{6}$$

where $U_k = \mathbf{u}_k \mathbf{u}_k^T$ for a uniform random direction $\mathbf{u}_k \in S^{n-1}$. We observe, that the error matrices $X_k := Y_k - H$ constitute a MRP, starting form $X_0 = Y_0 - H$. Hence, all our results form Section 2 can also be applied to RHA.

4 Discussion and Conclusions

We have studied the MRP scheme and derived exact expressions for the expected convergence. Our calculations of exact expected values improve upon previous results. We did not study single trajectories $\{X_k\}_{k\geq 0}$ for the scheme, although this seems to be necessary to fully explain the behavior observed in certain applications, for instance RHA, which was introduced in Section 3 and subject to previous studies [6, 9, 10].

The MRP (or equivalently the RHA) scheme can be seen as an estimator of a (given) matrix H. By considering the trace of the RHA, we also obtain an estimator of $\mathrm{Tr}[H]$, and we also have calculated the variance of this estimator. RHA uses only (random) linear measurements of H to estimate all entries of H. The corresponding MRP measures the error of the estimation. A variety of random estimators of a matrix H have been discussed in the literature. These estimators typically focus on the regime where the number of samples $m \ll n$ is much smaller than the dimension n of the matrix. Bekas et al. [3] discuss an estimator of the diagonal of H. Avron and Toledo [2] discuss several classical estimators of the trace $\mathrm{Tr}[H]$. The estimators have the form $\frac{1}{m}\sum_{i=1}^{m} H \bullet \mathbf{r}\mathbf{r}^T$, where the random vectors $\mathbf{r}$ follow some distribution. For instance $\mathbf{r} \sim \mathcal{N}(0, I_n)$.

In contrast to the estimators mentioned above, RHA aims at not only estimating a partial statistics of the matrix (trace or diagonal, but all entries). For RHA the number of needed measurements is of the same order as the number of entries of the matrix H. This aspect has also be discussed in [11].

References

[1] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[2] H. Avron and S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM*, 58(2):8, 2011.

[3] C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Appl. Numer. Math.*, 57(11-12):1214–1229, Nov. 2007.

[4] R. Heijmans. When does the expectation of a ratio equal the ratio of expectations? *Statist. Papers*, 40:107–115, 1999.

[5] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12:134–139, 1918.

[6] D. Leventhal and A. S. Lewis. Randomized Hessian estimation and directional search. *Optimization*, 60(3):329–345, 2011.

[7] J. Michalowicz, J. Nichols, F. Bucholtz, and C. Olson. An isserlis theorem for mixed gaussian variables: Application to the auto-bispectral density. *Journal of Statistical Physics*, 136:89–102, 2009.

[8] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[9] S. U. Stich. Random Pursuit in Hilbert Space. Technical report, ETH Zürich, 2013. Technical Report CGL-TR-xx.

[10] S. U. Stich and C. L. Müller. On spectral invariance of randomized hessian and covariance matrix adaptation schemes. In C. Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, and M. Pavone, editors, *Parallel Problem Solving from Nature - PPSN XII*, volume 7491 of *Lecture Notes in Computer Science*, pages 448–457. Springer Berlin Heidelberg, 2012.

[11] S. U. Stich, C. L. Müller, and B. Gärtner. Variable Metric Random Pursuit. *in preparation for Math. Prog.*, 2012.

A Miscellaneous

Proof of Remark 1.1. To show (i), we calculate the scalar product

$$X_{k+1} \bullet U_k = X_k \bullet U_k - (X_k \bullet U_k) \underbrace{U_k \bullet U_k}_{=1} = 0\,.$$

Next, (ii) follows by calculating the Frobenius norm of X_{k+1} explicitly:

$$\begin{aligned} \|X_{k+1}\|_F^2 &= X_{k+1} \bullet X_{k+1} = X_k \bullet X_k - 2(X_k \bullet U_k) U_k \bullet X_k + (X_k \bullet U_k)^2 \\ &= \|X_k\|_F^2 - \underbrace{(X_k \bullet U_k)^2}_{\geq 0}\,. \end{aligned} \tag{7}$$

Finally, to show (iii), we verify:

$$\begin{aligned} P X_{k+1} P^T &= P\left(X_k - (X_k \bullet U_k) U_k\right) P^T \\ &= P X_k P^T - (I_n X_k \bullet U_k I_n) P U_k P^T \\ &= P X_k P^T - (P X_k P^T \bullet P U_k P^T) P U_k P^T\,, \end{aligned}$$

where the last line follows from the cyclic-shift property of the trace operator: $\mathrm{Tr}[X_k I_n U_k P^T P] = \mathrm{Tr}[P X_k I_n U_k P^T] = \mathrm{Tr}[P X_k P^T P U_k P^T]$. □

Proof of Remark 1.2. Exactly as in the proof of Remark 1.1 we obtain

$$\mathbb{E}\left[\|X_{k+1}\|_F^2 \mid X_k\right] = \|X_k\|_F^2 - \mathbb{E}\left[(X_k \bullet U_k)^2 \mid X_k\right]\,.$$

Now claim (i) follows with Lemma B.1 (iii) form the appendix. Now we proceed to (ii). By Equation (3) we can bound

$$\mathbb{E}\left[\|X_k\|_F^2 \mid X_{k-1}\right] \leq \left(1 - \frac{2}{n(n+2)}\right) \|X_{k-1}\|_F^2\,. \tag{8}$$

What we now have to do formally, is to condition on X_{k-2} and calculate the expectation of $\mathbb{E}[E[X_k \mid X_{k-1}] \mid X_{k-2}]$. By the tower property of conditional expectations, $\mathbb{E}[E[X_k \mid X_{k-1}] \mid X_{k-2}] = \mathbb{E}[X_k \mid X_{k-2}]$. Repeating this procedure for X_{k-3} up to X_0, we finally obtain $E[X_{k+1} \mid X_0] = \mathbb{E}[X_{k+1}] = \left(1 - \frac{2}{n(n+2)} \|X_k\|_F^2\right)$.

Finally, for part (iii), we deduce form Equation (7):

$$\|X_k\|_F^2 = \left(1 - \frac{(X_{k-1} \bullet U_{k-1})^2}{\|X_{k-1}\|_F^2}\right) \|X_{k-1}\|_F^2\,.$$

With $1 - x \leq e^{-x}$ we obtain the final result. □

Lemma A.1. *Let X be symmetric $n \times n$ matrix and $a > 0$. If $\lambda_{\max}(XX^T) \leq a^2$ then $X \preceq aI_n$, $-X \preceq aI_n$. Especially, $|\lambda_i(X)| \leq a$ for $i = 1, \ldots, n$.*

Here the notation $X \preceq Y$ for symmetric matrices X, Y means $\mathbf{z} X \mathbf{z}^T \leq \mathbf{z} Y \mathbf{z}^T$ for all $\mathbf{z} \in \mathbb{R}^n$.

Proof. If there exists $\mathbf{z} \in \mathbb{R}^n$ with $\|\mathbf{z}\|_2 = 1$ and $\mathbf{z} X \mathbf{z}^T > a$, then $a^2 < \mathbf{z} X \mathbf{z}^T \mathbf{z} X \mathbf{z}^T = \mathbf{z} X X^T \mathbf{z}^T$; a contradiction. The same argument also applies to $(-X)$. The last statement follows from the variational characterization of the eigenvalues. □

B Expectations

Lemma B.1. *Let $\mathbf{v} \sim S^{n-1}$, the uniform distribution on all n-dimensional unit vectors, and let $A \in \mathrm{SYM}_n$ be a symmetric $n \times n$ matrix. Then*

$$\begin{aligned}
&\textit{(i)} \quad \mathbb{E}\left[\mathbf{v}^T A \mathbf{v}\right] = \frac{\mathrm{Tr}[A]}{n}\,, \\
&\textit{(ii)} \quad \mathbb{E}\left[(\mathbf{v}^T A \mathbf{v})\mathbf{v}\mathbf{v}^T\right] = \frac{2A + \mathrm{Tr}[A] I_n}{n(n+2)}\,, \\
&\textit{(iii)} \quad \mathbb{E}\left[(\mathbf{v}^T A \mathbf{v})^2\right] = \frac{2\mathrm{Tr}[A^2] + \mathrm{Tr}[A]^2}{n(n+2)}\,, \\
&\textit{(iv)} \quad \mathbb{E}\left[(\mathbf{v}^T A \mathbf{v})^2 \mathbf{v}\mathbf{v}^T\right] = \frac{8A^2 + 4\mathrm{Tr}[A]A + (\mathrm{Tr}[A]^2 + 2\mathrm{Tr}[A^2]) I_n}{n(n+2)(n+4)}\,.
\end{aligned}$$

The proof of this lemma is in the appendix. Part (i) and (ii) have already been shown by the authors in [11], but will be repeated for the sake of completeness.

In order to prove Lemma B.1 we will first show a similar statement for normal distributed random variables.

Lemma B.2. *Let $\mathbf{u} \sim \mathcal{N}(0, I_n)$, and let $A \in \mathrm{SYM}_n$ be a symmetric $n \times n$ matrix. Then*

$$\begin{aligned}
&\textit{(i)} \quad \mathbb{E}\left[\mathbf{u}^T A \mathbf{u}\right] = \mathrm{Tr}[A]\,, \\
&\textit{(ii)} \quad \mathbb{E}\left[(\mathbf{u}^T A \mathbf{u})\mathbf{u}\mathbf{u}^T\right] = 2A + \mathrm{Tr}[A] I_n\,, \\
&\textit{(iii)} \quad \mathbb{E}\left[(\mathbf{u}^T A \mathbf{u})^2 \mathbf{u}\mathbf{u}^T\right] = 8A^2 + 4\mathrm{Tr}[A]A + (\mathrm{Tr}[A]^2 + 2\mathrm{Tr}[A^2]) I_n\,.
\end{aligned}$$

Proof. This lemma might be well known to many of the readers, as it is a straight forward application of Isserlis' Theorem [5]. This theorem states that if $u_1, \ldots u_{2N}$ are Gaussian random variables, then

$$\mathbb{E}[u_1 u_2 \ldots u_{2N}] = \sum \prod \mathbb{E}[u_i u_j]\,,$$

where notation $\sum\prod$ means summing over all distinct ways of partitioning $u_1, u_2, \ldots u_{2N}$ into pairs (cf. [7]). We can establish the following relations for the components of a random multivariate Gaussian vector $\mathbf{u} \sim \mathcal{N}(0, I_n)$:

$$\mathbb{E}[u_i u_j] = \delta_{ij}\,, \qquad \mathbb{E}[u_i u_j u_k u_l] = \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\,,$$

and the expression for $\mathbb{E}[u_i u_j u_k u_l u_m u_o]$ involves already 15 products of three δ-functions each, following the same pattern.

We can now proceed to prove the theorem. For part (i), we easily get

$$\mathbb{E}[\mathbf{u}^T A \mathbf{u}] = \sum_{i,j=1}^{n} E[u_i u_j A_{ij}] = \mathrm{Tr}[A^T] = \mathrm{Tr}[A]\,.$$

Similarly, we can compute the ij-th entry of the matrix of (ii):

$$\left(\mathbb{E}[(\mathbf{u}^T A \mathbf{u})\mathbf{u}\mathbf{u}^T]\right)_{ij} = \sum_{k,l=1}^{n} \mathbb{E}[u_i u_j u_k u_l A_{kl}] = A_{ij} + A_{ji} + \delta_{ij} \sum_{k=1}^{n} A_{kk}\,,$$

amounting to $\mathbb{E}\left[(\mathbf{u}^T A\mathbf{u})\mathbf{u}\mathbf{u}^T\right] = A + A^T + \mathrm{Tr}[A]I_n$ as desired. For (iii):

$$\begin{aligned}\left((\mathbf{u}^T A\mathbf{u})^2\mathbf{u}\mathbf{u}^T\right)_{ij} &= \sum_{k,l,m,o=1}^{n} \mathbb{E}[u_i u_j u_k u_l u_m u_o A_{kl} A_{mo}] \\ &= 4\sum_{k=1}^{n} A_{ik}A_{kj} + 4\sum_{k=1}^{n} A_{ik}A_{jk} + 2A_{ij}\sum_{k=1}^{n} A_{kk} + 2A_{ji}\sum_{k=1}^{n} A_{kk} \\ &\quad + \delta_{ij}\sum_{k,l=1}^{n} \left(A_{kl}A_{kl} + A_{kl}A_{lk} + A_{kk}A_{ll}\right) ,\end{aligned}$$

proving claim (iii). □

Lemma B.3. *Let* $\mathbf{u} \sim \mathcal{N}(0, I_n)$*. Then*

$$E[\|\mathbf{u}\|^2] = n\,, \qquad E[\|\mathbf{u}\|^4] = n(n+2)\,, \qquad E[\|\mathbf{u}\|^6] = n(n+2)(n+4)\,.$$

Proof. The proof of this (probably again well known fact) is an immediate application of Isserli's theorem. The first equation follows directly form Lemma B.2 with $A = I_n$, the remaining two identities are due to

$$\begin{aligned}E[\|\mathbf{u}\|^4] &= \sum_{i,j}^{n} \mathbb{E}[u_i^2 u_j^2] = (n-1)n + 3n\,, \\ E[\|\mathbf{u}\|^6] &= \sum_{i,j,k}^{n} \mathbb{E}[u_i^2 u_j^2 u_k^2] = (n-2)(n-1)n + 3\cdot 3(n-1)n + 15n\,.\end{aligned}$$ □

Now we are able to proceed to the proof of Lemma B.1.

Proof of Lemma B.1. First, let us affirm a simple fact. Consider $f\colon \mathbb{R}^n \to R$ with $f(\mathbf{x}) = f(\mathbf{x}/\|\mathbf{x}\|)$. Then – omitting the technical details – $f(\mathbf{x})$ is clearly independent of the norm $\|\mathbf{x}\|$ (or any power $\|\mathbf{x}\|^k$).

We observe, that for $\mathbf{u} \sim \mathcal{N}(0, I_n)$ the random variable $\frac{\mathbf{u}}{\|\mathbf{u}\|}$ follows the S^{n-1} distribution. E.g. the expectation in (i) can be rewritten as

$$\mathbb{E}_{\mathbf{v}}[\mathbf{v}^T A\mathbf{v}] = \mathbb{E}_{\mathbf{u}}\left[\frac{\mathbf{u}^T A\mathbf{u}}{\|\mathbf{u}\|^2}\right] .$$

By the comment at the beginning of this proof, the random variables $\frac{\mathbf{u}^T A\mathbf{u}}{\|\mathbf{u}\|^2}$ and $\|\mathbf{u}\|^2$ are independent. Thus by a (fairly tautological result) of Heijmans [4], the expectation of $\frac{\mathbf{u}^T A\mathbf{u}}{\|\mathbf{u}\|^2}$ equals the ratio of the expectations of $\mathbf{u}^T A\mathbf{u}$ and $\|\mathbf{u}\|^2$. These expectations have been calculated in Lemmas B.2 and B.3.

The proof of part (ii) and (iv) follows by the same method. For (iii) we note that we can reduce it to the case (ii):

$$\mathbb{E}\left[(\mathbf{v}^T A\mathbf{v})^2\right] = \mathbb{E}\left[\mathrm{Tr}\left[A(\mathbf{v}^T A\mathbf{v})\mathbf{v}\mathbf{v}^T\right]\right] .$$ □

C Solving Linear Recurrence Relations

For the reminder of this section, let us fix the following three parameters, depending on the dimension $n \geq 1$:

$$\alpha = \frac{1}{n}\,, \qquad \beta = \frac{1}{n(n+2)}\,, \qquad \gamma = \frac{1}{n(n+2)(n+4)}\,, \qquad \delta = \frac{1}{n(n+4)}\,.$$

C.1 Matrix I

Lemma C.1 (Matrix diagonalization I)**.** *Let $n \geq 1$ and consider the following $(n+1) \times (n+1)$ matrix:*

$$A(n) := \begin{bmatrix} 1-\alpha & \mathbf{0}_n^T \\ \beta \mathbf{1}_n & (1-2\beta)I_n \end{bmatrix}\,,$$

where $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ denote the all-zero and all-one n-dimensional vectors. Then

$$A(n) = \begin{bmatrix} 1 & \mathbf{0}_n^T \\ \frac{1}{n}\mathbf{1}_n & I_n \end{bmatrix} \begin{bmatrix} 1-\alpha & \mathbf{0}_n^T \\ \mathbf{0}_n & (1-2\beta)I_n \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}_n^T \\ -\frac{1}{n}\mathbf{1}_n & I_n \end{bmatrix}\,.$$

Proof. The claim can be verified by calculating the product of the tree matrices. □

Corollary C.2. *Let $x_0 \in \mathbb{R}$, $\mathbf{y}_0 \in \mathbb{R}^n$, $k \geq 0$ and $x_k, \in R$, $\mathbf{y}_k \in R^n$ with $(x_k, \mathbf{y}_k)^T = A(n)^k (x_0, \mathbf{y}_0)^T$. Then*

$$\begin{aligned} x_k &= (1-\alpha)^k x_0\,, \\ \mathbf{y}_k &= (1-\alpha)^k \alpha x_0 \mathbf{1}_n + (1-2\beta)^k(\mathbf{y}_0 - \alpha x_0 \mathbf{1}_n)\,. \end{aligned}$$

C.2 Matrix II

Lemma C.3 (Matrix diagonalization II)**.** *Let $n \geq 1$ and consider the following $(n+2) \times (n+2)$ matrix:*

$$B(n) := \begin{bmatrix} 1-2\beta & -\beta & \mathbf{0}_n^T \\ 2\beta & 1-(2n-3)\beta & \mathbf{0}_n^T \\ 2\gamma\mathbf{1}_n & (n+3)\gamma\mathbf{1}_n & (1-3\delta)I_n \end{bmatrix}\,,$$

where $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ denote the all-zero and all-one n-dimensional vectors. Then

$$B(n) = \begin{bmatrix} \frac{2n+1-\omega}{4\omega} & \frac{2n+1+\omega}{4\omega} & \mathbf{0}_n^T \\ \frac{1}{\omega} & \frac{1}{\omega} & \mathbf{0}_n^T \\ -\epsilon\mathbf{1}_n & -\zeta\mathbf{1}_n & I_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \mathbf{0}_n^T \\ 0 & \lambda_2 & \mathbf{0}_n^T \\ \mathbf{0}_n & \mathbf{0}_n & (1-3\delta)I_n \end{bmatrix} \begin{bmatrix} -2 & \frac{\omega+2n+1}{2} & \mathbf{0}_n^T \\ 2 & \frac{\omega-2n-1}{2} & \mathbf{0}_n^T \\ 2(\zeta-\epsilon)\mathbf{1}_n & \eta\mathbf{1}_n & I_n \end{bmatrix}\,,$$

with $\omega = \sqrt{4n^2+4n-7}$,

$$\lambda_1 = \frac{2n^2+2n-5-\omega}{2n(n+2)}\,, \qquad \lambda_2 = \frac{2n^2+2n-5+\omega}{2n(n+2)}\,,$$

$$\epsilon = \frac{(n+3)\gamma}{1-3\delta-\lambda_1}\,, \qquad \zeta = \frac{2\gamma}{1-3\delta-\lambda_2}\,, \qquad \eta = (\epsilon+\zeta)\frac{\omega}{2} + (\epsilon-\zeta)\frac{2n+1}{2}\,.$$

Proof. The claim can be verified by calculating the product of the tree matrices. □

Corollary C.4. *Let $n \geq 1$ and consider the following 2×2 matrix:*

$$C(n) := \begin{bmatrix} 1-2\beta & -\beta \\ 2\beta & 1-(2n-3)\beta \end{bmatrix}\,.$$

Then

$$C(n) = \begin{bmatrix} \frac{2n+1-\omega}{4\omega} & \frac{2n+1+\omega}{4\omega} \\ \frac{1}{\omega} & \frac{1}{\omega} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} -2 & \frac{\omega+2n+1}{2} \\ 2 & \frac{\omega-2n-1}{2} \end{bmatrix}\,.$$

Proof. This follows directly of Lemma C.3 by observing that $C(n)$ is the restriction of $B(n)$ on the subspace spanned by the first two basis vectors of $\mathbb{R}^{n+2}$. ☐

Corollary C.5. *Let $x_0, y_0 \in \mathbb{R}$, $\mathbf{z}_0 \in \mathbb{R}^n$, $k \geq 0$ and $x_k, y_k \in R$, $\mathbf{z}_k \in R^n$ with $(x_k, y_k, \mathbf{z}_k)^T = B(n)^k (x_0, y_0, \mathbf{z}_0)^T$. Then*

$$\begin{aligned} x_k &= \left(\lambda_1^k + \lambda_2^k\right) \frac{x_0}{2} + \left(\lambda_2^k - \lambda_1^k\right) \cdot \left(\frac{(2n+1)x_0}{2\omega} - \frac{y_0}{\omega}\right), \\ y_k &= \left(\lambda_1^k + \lambda_2^k\right) \frac{y_0}{2} + \left(\lambda_2^k - \lambda_1^k\right) \cdot \left(\frac{2x_0}{\omega} - \frac{(2n+1)y_0}{2\omega}\right), \\ \mathbf{z}_k &= \epsilon \lambda_1^2 \left(2x_0 - \frac{\omega + 2n + 1}{2} y_0\right) \mathbf{1}_n - \zeta \lambda_2^k \left(2x_0 + \frac{\omega - 2n - 1}{2} y_0\right) \mathbf{1}_n \\ &\quad + (1 - 3\delta)^k \left(2(\zeta - \epsilon) x_0 \mathbf{1}_n + \eta y_0 \mathbf{1}_n + \mathbf{z}_0\right). \end{aligned}$$

Moreover, for $x_0, y_0, \mathbf{z}_0 \geq 0$, $n \geq 4$ and parameter $\rho := 1 + \frac{1}{4n}$, we get the following upper and lower bounds (where the '$\leq$' sign for vectors means that the inequality holds element-wise):

$$x_k \leq \lambda_2^k \left(\rho x_0 - \frac{y_0}{2n+1}\right) + \lambda_1^k \frac{y_0}{2n}, \qquad x_k \geq \lambda_2^k \left(x_0 - \frac{y_0}{2n}\right) - \lambda_1^k \frac{x_0}{4n},$$

$$y_k \leq \lambda_1^k \left(\rho y_0 - \frac{2x_0}{2n+1}\right) + \lambda_2^k \frac{x_0}{n}, \qquad y_k \geq \lambda_1^k \left(y_0 - \frac{2x_0}{2n+1}\right) - \lambda_2^k \frac{y_0}{4n}.$$

$$\mathbf{z}_k \leq (\epsilon \lambda_1^k - \zeta \lambda_2^k) 2x_0 \mathbf{1}_n + (1 - 3\delta)^k \left(\eta y_0 \mathbf{1}_n + \mathbf{z}_0\right).$$

Proof. The equalities all follow from Lemma C.3. To verify the inequalities, we first note $2n \leq \omega \leq (2n+1)$ for $n \geq 2$. Hence,

$$\begin{aligned} \frac{2n+1}{2\omega} + \frac{1}{2} &\leq \frac{2n+1+2n}{4n} = 1 + \frac{1}{4n}, \\ \frac{2n+1}{2\omega} - \frac{1}{2} &\leq \frac{2n+1-2n}{4n} = \frac{1}{4n+1} \leq \frac{1}{4n}. \end{aligned}$$

Now the inequalities for x_k and y_k are easily checked. For the remaining one, we observe that $\epsilon \geq 0$, $\zeta \leq 0$ (for $n \geq 4$) and $\eta \geq 0$ (for $n \geq 4$). ☐

For large n, we can simplify the estimate of $\mathbf{z}_k$ even further.

Corollary C.6. *Let $n \geq 5$. Then*

$$\mathbf{z}_k \leq \left(\frac{\lambda_1^k}{4n} + \frac{\lambda_2^k}{n-4}\right) 4x_0 \mathbf{1}_n + (1 - 3\delta)^k (y_0 \mathbf{1}_n + \mathbf{z}_0).$$

Proof. First we show $\epsilon \leq \frac{1}{2n+1}$. It is easy to check $\omega \leq (2n+1)$ for $n \geq 2$, and therefore $\lambda_1 \leq \frac{2n^2-5}{2n(n+2)}$. It follows

$$(2n+1)\epsilon \leq \frac{2n(n+3)\gamma}{1 - \frac{3}{n(n+4)} - \frac{2n^2-5}{2n(n+2)}} = \frac{4n^5 + 38n^4 + 122n^3 + 148n^2 + 48n}{4n^5 + 39n^4 + 130n^3 + 22n^2 + 64n} \leq 1.$$

Now we show $-\zeta \leq \frac{2}{n-4}$. With subsequent Lemma C.7 we have $\lambda_2 \geq 1 - \frac{5}{2n(n+2)}$ for $n \geq 2$, hence

$$-\zeta \leq \frac{2\gamma}{\frac{3}{n(n+4)} - \frac{5}{2n(n+2)}} = \frac{2}{n-4}.$$

Finally, we show $\eta \leq 1$. It is easy to verify $\epsilon + \zeta \geq 0$. Together with the already established fact $\omega \leq (2n+1)$ for $n \geq 2$, we obtain

$$\eta \leq (\epsilon + \zeta)\frac{2n+1}{2} + (\epsilon - \zeta)\frac{2n+1}{2} = \epsilon(2n+1) \leq 1.$$ ☐

Lemma C.7. *Let λ_1, λ_2 as in Lemma C.1 and $n \geq 2$. Then*

$$\begin{aligned} 1 - \frac{2}{n} &\leq \lambda_1 \leq 1 - \frac{2}{n+1}\,, \\ 1 - \frac{5}{2n(n+2)} &\leq \lambda_2 \leq 1 - \frac{2}{n(n+2)}\,. \end{aligned}$$

Proof. The main inequality we use is again $2n \leq \omega \leq (2n+1)$ for $n \geq 2$. Thus

$$\begin{aligned} \lambda_1 &\leq \frac{2n^2-5}{2n(n+2)} = \frac{(n+1)(2n^2-5)}{2n(n+1)(n+2)} \leq \frac{2(n-1)n(n+2)}{2n(n+1)(n+2)} = 1 - \frac{2}{n+1}\,, \\ \lambda_1 &\geq \frac{n^2-3}{n(n+2)} \geq \frac{(n-2)(n+2)}{n(n+2)} = 1 - \frac{2}{n}\,. \end{aligned}$$

Similarly for the last two inequalities:

$$\begin{aligned} \lambda_2 &\leq \frac{n^2+2n-2}{n(n+2)} = 1 - \frac{2}{n(n+2)}\,, \\ \lambda_2 &\geq \frac{2n^2+4n-5}{2n(n+2)} \geq \frac{n^2+2n-5/2}{n(n+2)} = 1 - \frac{5}{2n(n+2)}\,. \end{aligned}$$

□