Averaged Least-Mean-Square: Bias-Variance Trade-offs and Optimal Sampling Distributions

Supplementary material

Alexandre Défossez
Département de Mathématiques
École Normale Supérieure
Paris, France
alexandre.defossez@ens.fr

Francis Bach
Département d'Informatique
École Normale Supérieure
Paris, France
francis.bach@inria.fr

We give hereafter the proofs for the different results in the main paper. Unless otherwise specified, references are to the present Appendix. We first give a more thorough definition of the space in which our operators live. We then proceed to a proof of Lemma 1. Finally we detail the computation that allowed us to derive both theorems in the main paper.

1 Linear algebra prerequisites

Throughout our results we will use the following notations and results. These are necessary to provide explicit expressions for the constants in the asymptotic expansions.

For any real vector space V of finite dimension d, let $\mathcal{M}(V)$ be the space of linear operators over V which is isomorphic to the space of d-by-d matrices, with the usual results that composition becomes matrix multiplication. As a consequence we will use the same notation for the space of matrices and the space of endomorphisms.

We denote by $\mathcal{I} = \mathcal{M}(\mathcal{M}(\mathbb{R}^d))$ the space of endomorphisms on the space of matrices over $\mathbb{R}^d$. One can index the rows and columns of a matrix $M \in \mathcal{I}$ by a pair (i, j) where $1 \leq i \leq d$ and $1 \leq j \leq d$. We will often denote by $M_{(i,j),(k,l)}$ an element of this matrix on matrices. In the following we will drop the domain of i, j, k, l, i', j' which is implicitly $\{1, 2, \dots, d\}$. Explicitly, if $A \in \mathcal{M}(\mathbb{R}^d)$ and $M \in \mathcal{I}$, then MA is defined through:

$$\forall (i,j)(MA)_{i,j} = \sum_{i'=1,j'=1}^{d} M_{(i,j),(i',j')} A_{i',j'}$$

We will mostly make no distinction between A as a vector in $\mathcal{M}(\mathbb{R}^d)$ on which elements in $\mathcal{I}$ can operate and A as a matrix in $\mathcal{M}(\mathbb{R}^d)$. Then MA can be either usual matrix multiplication if $M, A \in \mathcal{M}(\mathbb{R}^d)$ or $M, A \in \mathcal{I}$ or application of M to A if $M \in \mathcal{I}$ and $A \in \mathcal{M}(\mathbb{R}^d)$. However, if $M \in \mathcal{I}$ and $A \in \mathcal{M}(\mathbb{R}^d)$, then AM does not make sense. For $P \in \mathcal{I}$, and any (i, j), we will define $P_{i,j}$ the matrix in $\mathcal{M}(\mathbb{R}^d)$ with coefficient (i', j') given by $P_{(i,j),(i',j')}$.

For any $V \in \mathcal{S}(\mathbb{R}^d)$ (the set of symmetric matrices of size d), we will denote $\|V\|_{\text{op}}$ the operator norm of V or equivalently its eigenvalue with the largest absolute value. For any $M \in \mathcal{I}$ so that $\mathcal{S}(\mathbb{R}^d)$ is stable under M, we will take $\|M\|_{\text{op}}$ the operator norm of M restricted to $\mathcal{S}(\mathbb{R}^d)$, defined with respect to the Frobenius norm on $\mathcal{S}(\mathbb{R}^d)$, that is

$$\|M\|_{\text{op}} = \sup_{V \in \mathcal{S}(\mathbb{R}^d), \|V\|_F = 1} \|MV\|_F .$$

Equivalentely, it is given by the largest absolute value of the eigenvalues of M.

Finally, we will look more precisely at three elements of $\mathcal{I}$. For any given $A \in \mathcal{M}(\mathbb{R}^d)$, one can define A_L (resp. A_R) so that A_L is the matrix in $\mathcal{I}$ representing left multiplication (resp. right multiplication) by A. The coefficient of A_L and A_R are given by

$$\begin{aligned}\forall (i,j),(k,l), \quad (A_L)_{(i,j),(k,l)} &= \delta_{j,l} A_{i,k} \\ \forall (i,j),(k,l), \quad (A_R)_{(i,j),(k,l)} &= \delta_{i,k} A_{j,l}.\end{aligned}$$

If A is symmetric, then A_L and A_H are both symmetric operators, and $A_L + A_H$ is stable on the subspace of symmetric matrices, that we have denoted $\mathcal{S}(\mathbb{R}^d)$.

Let X be a random variable in $\mathbb{R}^d$, we consider the linear operator M on $\mathcal{M}(\mathbb{R}^d)$ defined by,

$$\forall A \in \mathcal{M}(\mathbb{R}^d), \quad MA = \mathbb{E}\left[(X^T A X) X X^T\right],$$

then, the coefficients of the associated matrix are given by

$$\forall (i,j,k,l), \quad M_{(i,j),(k,l)} = \mathbb{E}\left[X^{(i)} X^{(j)} X^{(k)} X^{(l)}\right],$$

where $X^{(i)}$ denote the i-th component of the vector X. The matrix M is clearly symmetric. One can also prove that it is stable on $\mathcal{S}(\mathbb{R}^d)$.

We then define $T = H_L + H_R - \gamma M$ with H_L, H_R and M as defined above for the random variable X defined in our setup. It is immediately stable over $\mathcal{S}(\mathbb{R}^d)$. We will denote μ_T the smallest eigenvalue of T.

2 Proof of Lemma 1

Here is a more complete version of Lemma 1 in the paper.

Lemma 1. *Using the notations and assumptions of Section (1.1) of the original paper, define $\gamma_{\max}$ as the supremum of $\gamma > 0$ such that*

$$\forall A \in \mathcal{S}(\mathbb{R}^d), \quad 2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathbb{E}\left[(X^T A X)^2\right] > 0. \tag{2.1}$$

If $0 < \gamma < \gamma_{\max}$ then T is positive definite and $\rho < 1$. More precisely, in dimension $d \geqslant 2$, we have

$$\begin{cases} \rho \leq 1 - 2\gamma \left(1 - \dfrac{\gamma}{\gamma_{\max}}\right) \mu & \text{if } 1 > \frac{\gamma}{\gamma_{\max}} \geq \frac{1}{2} \\ \rho \leq 1 - \gamma\mu & \text{otherwise.} \end{cases} \tag{2.2}$$

In dimension $d = 1$, we have

$$\rho \leq \max\left(\left|1 - \gamma\mu\right|, 1 - 2\gamma\left(1 - \frac{\gamma}{\gamma_{\max}}\right)\mu\right).$$

Otherwise, if $\gamma > \gamma_{\max}$, then $\rho > 1$.

In order to prove it, we will first need some preliminary results.

2.1 Some Lemmas

Lemma 2. *Let $A \in \mathcal{S}(\mathbb{R}^d)$ be any symmetric matrix, then*

$$\forall x \in \mathbb{R}^d, (x^T A x)^2 \leq \mathrm{Tr}\left((x^T x) A x x^T A\right).$$

Proof. Using Cauchy-Schwarz inequality, one has

$$\begin{aligned}(x^T A x)^2 = [x^T (Ax)]^2 &\leq (Ax)^T (Ax)(x^T x) \\ &= x^T A A x (x^T x) = \mathrm{Tr}\left((x^T x) A x x^T A\right).\end{aligned}$$

□

The following lemma is the proof of equation (2.4) in the original paper.

Lemma 3. *Let $H \in \mathcal{S}(\mathbb{R}^d)$ be a positive semi-definite matrix. If $\gamma > 0$ is so that*

$$\forall A \in \mathcal{S}(\mathbb{R}^d),\ 2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathbb{E}\left[(X^T A X)^2\right] > 0$$

then

$$\gamma < \frac{2}{\mathrm{Tr}\,(H)}.$$

Proof. Let $A \in \mathcal{S}(\mathbb{R}^d)$, $2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathbb{E}\left[(X^T A X)^2\right] > 0$ implies that with Jensen's inequality,

$$\begin{aligned} 2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathrm{Tr}\,(AH)^2 &= 2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathrm{Tr}\left(A \mathbb{E}\left[X X^T\right]\right)^2 \\ &= 2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathbb{E}\left[X^T A X\right]^2 \\ &> 0. \end{aligned}$$

Then, let $(u_i)_i \in \mathbb{R}^{d \times d}$ an orthogonal basis that diagonalizes H and λ_i the eigenvalues associated with each eigenvector. Then, taking $A = \sum_i u_i u_i^T$, we get

$$\begin{aligned} 2\mathrm{Tr}\left(A^T H A\right) - \gamma \mathrm{Tr}\,(AH)^2 &= 2\mathrm{Tr}\left(\sum_{i,j} u_i u_i^T H u_j u_j^T\right) - \gamma \left(\sum_i u_i^T H u_i\right)^2 \\ &= 2\left(\sum_i \lambda_i\right) - \gamma \left(\sum_i \lambda_i\right)^2 \geq 0, \end{aligned}$$

so that

$$\gamma < \frac{2}{\sum_i \lambda_i} = \frac{2}{\mathrm{Tr}\,(H)}.$$

□

Lemma 4. *Let $\gamma > 0$, we can define $T = H_L + H_R - \gamma M$ as in Section 1. If $\gamma < \frac{2}{\mathrm{Tr}(H)}$, then*

$$I - \gamma T \succ -I.$$

and if we are in dimension 1,

$$I - \gamma T \succeq 0$$

Proof. The Lemma is equivalent to $\forall A \in \mathcal{S}(\mathbb{R}^d), A \neq 0 \Rightarrow \langle A, (2I - \gamma T)A\rangle > 0$.

If we are in dimension $d = 1$, then we have $I - \gamma T = 1 - 2\gamma h + \gamma m^2$ where $h = \mathbb{E}\left[X^2\right]$ and $m = \mathbb{E}\left[X^4\right] \geq h^2$ so that $I - \gamma T \geq (1 - \gamma h)^2 \geq 0$.

Let now assume we are in dimension two or more. Let $A \in \mathcal{S}(\mathbb{R}^d)$ with $A \neq 0$. Let $P \in \mathbb{R}^{d \times d}$ be an orthogonal matrix such that $PHP^{-1} = D$ where D is diagonal with eigenvalues ordered in decreasing order, with $\lambda_i = D_{i,i}$ and $\lambda_1 = L$. We will denote $U = PAP^{-1} = PAP^T$.

$$\begin{aligned} \langle A, (2I - \gamma T)A\rangle &= \mathrm{Tr}\left(A^T (2I - \gamma T)A\right) \\ &= 2\mathrm{Tr}\left(A^T A\right) - 2\gamma \mathrm{Tr}\left(A^T H A\right) + \gamma^2 \mathbb{E}\left[\left(X^T A X\right)^2\right] \\ &\geq 2\mathrm{Tr}\left(A^T A\right) - 2\gamma \mathrm{Tr}\left(A^T H A\right) + \gamma^2 \mathbb{E}\left[\left(X^T A X\right)\right]^2 \\ &= 2\mathrm{Tr}\left(A^T A\right) - 2\gamma \mathrm{Tr}\left(A^T H A\right) + \gamma^2 \mathrm{Tr}\,(AH)^2 \\ &= 2\mathrm{Tr}\left(U^T U\right) - 2\gamma \mathrm{Tr}\left(U^T D U\right) + \gamma^2 \mathrm{Tr}\,(UD)^2 \\ &= \sum_{i,j=1}^{d} 2U_{i,j}^2 - 2\gamma U_{i,j}^2 \lambda_i + \gamma^2 U_{i,i} U_{j,j} \lambda_i \lambda_j \\ &= \left(\sum_{i \neq j} 2U_{i,j}^2 (2 - \gamma(\lambda_i + \lambda_j))\right) + \sum_{i=1}^{d} 2U_{i,i}^2 - 2\gamma U_{i,i}^2 \lambda_i + \gamma^2 \left(\sum_{i=1}^{d} U_{i,i} \lambda_i\right)^2. \end{aligned}$$

The first sum immediately defines a definite positive form over the subspace generated by $(U_{i,j})_{i \neq j}$ as $\gamma < \frac{2}{\lambda_i + \lambda_j}$ for all $i \neq j$. The second part also defines a bilinear form over the orthogonal subspace generated by $(U_{i,i})_{1 \leq i \leq d}$. $2I - \gamma T$ is definite positive if and only if those two forms are definite positive. We will introduce $x_i = U_{i,i}$ so that the second form is given by $x^T G x$

where $G = 2I - 2\gamma \text{Diag}(\Lambda) + \gamma^2 \Lambda\Lambda^T$, with $\Lambda = (\lambda_i)_{1\leq i\leq d}$ and $\text{Diag}(\Lambda)$ the diagonal matrix with values from Λ on the diagonal.

We can decompose G as

$$G = \begin{pmatrix} B & \gamma^2\lambda_1 C^T \\ \gamma^2\lambda_1 C & D \end{pmatrix},$$

with $B = 2 - 2\gamma\lambda_1 + \gamma^2\lambda_1$, $C = (\lambda_i)_{2\leq i\leq d}$ and $D = 2I - 2\gamma\text{Diag}(C) + \gamma^2 CC^T$. Using the Schur completement condition for positive definiteness, we have that $G \succ 0$ if and only if $D \succ 0$ and $B - \gamma^4\lambda_1^2 C^T D^{-1} C > 0$. We immediately have that $D \succ 0$ as $I - \gamma\text{Diag}(C) \succ 0$, indeed, for all $d \geq i \geq 2$, we have that $\gamma\lambda_i < 1$.

Let us introduce $E = 2I - 2\gamma\text{Diag}(C)$, then we have

$$D^{-1} = E^{-1} - \frac{\gamma^2}{1 + \gamma^2 C^T E^{-1} C} E^{-1} C C^T E^{-1}.$$

We will assume that $\sum_{i=2}^d \lambda_i < \lambda_1$, otherwise one trivially has that $\gamma\lambda_1 < 1$ and $G \succ 0$. Let us denote

$$\begin{aligned} q &= C^T E^{-1} C \\ &= \sum_{i=2}^d \frac{\lambda_i^2}{2(1-\lambda_i\gamma)} \\ &\leq \frac{(\sum_{i=2}^d \lambda_i)^2}{2(1-\gamma\sum_{i=2}^d \lambda_i)} \\ &= \frac{l^2}{2(1-\gamma l)}, \end{aligned}$$

where $l = \sum_{i=2}^d \lambda_i$. We will take $l = \lambda_1\alpha$ so that $0 < \alpha < 1$. We have

$$\begin{aligned} B - \gamma^4\lambda_1^2 C^T D^{-1} C &= \gamma^2\lambda_1^2 + 2 - 2\lambda_1\gamma - \gamma^4\lambda_1^2 \left(q - \frac{\gamma^2 q^2}{1+\gamma^2 q}\right) \\ &= \frac{\gamma^2\lambda_1^2}{1+\gamma^2 q} - 2\lambda_1\gamma + 2 \\ &\geq \frac{\gamma^2\lambda_1^2}{1+\gamma^2\frac{l^2}{2(1-\gamma l)}} - 2\lambda_1\gamma + 2. \end{aligned}$$

Denoting $y = \gamma\lambda_1$, we get

$$B - \gamma^4\lambda_1^2 C^T D^{-1} C = \frac{2y^2(1-y\alpha)}{2 - 2y\alpha + \alpha^2 y^2} - 2y + 2.$$

Using standard analysis tools, one can show that the last quantity is positive for $0 < y < \frac{2}{1+\alpha}$ and $0 < \alpha < 1$. As a conclusion, G is definite positive and so is $2I - \gamma T$.

□

Lemma 5. *Let $\gamma > 0$, we can define $T = H_L + H_R - \gamma M$ which is symmetric and is stable over $\mathcal{S}(\mathbb{R}^d)$.*

If

$$\forall A \in \mathcal{S}(\mathbb{R}^d),\ 2\text{Tr}\left(A^T H A\right) - \gamma\mathbb{E}\left[(X^T A X)^2\right] > 0,$$

or (this second assumption implies the first one)

$$\mathbb{E}\left[XX^T\right] - \gamma\mathbb{E}\left[X^T X X X^T\right] \succ 0,$$

then

- $\|I - \gamma H\|_{op} < 1$,
- $T \succ 0$,
- $\|I - \gamma T\|_{op} < 1$.

Proof. We should first notice that using Lemma 3, we necessarely have

$$\gamma < \frac{2}{\mathrm{Tr}\,(H)}. \tag{2.3}$$

We first need, $I - \gamma H \prec I$ which is always true as long as H is invertible (i.e. H is positive). Then we need $I - \gamma H \succ -I$, or $\gamma H \prec 2I$, which means $\gamma < \frac{2}{L}$ where L is H largest eigenvalue. However this is implied by (2.3).

Now, we need $I - \gamma T \prec I$, i.e., $T \succ 0$ (this will also prove T invertible). This is equivalent to

$$\forall A \in \mathcal{S}(\mathbb{R}^d), A \neq 0 \Rightarrow \langle A, TA\rangle > 0$$

Let us compute this term for $A \in \mathcal{S}(\mathbb{R}^d)$ with $A \neq 0$

$$\begin{aligned}
\langle A, TA\rangle &= \mathrm{Tr}\left(A^T(TA)\right) \\
&= \mathrm{Tr}\left(A^TAH + A^THA - \gamma A^T\mathbb{E}\left[XX^TAXX^T\right]\right) \\
&= 2\mathrm{Tr}\left(A^THA\right) - \mathbb{E}\left[(X^TAX)^2\right] \quad \text{and we can stop here if we have first assumption} \\
&\geq \mathrm{Tr}\left(A^T\left(2H - \gamma\mathbb{E}\left[X^TXXX^T\right]\right)A\right) \quad \text{using Lemma 2}
\end{aligned}$$

A sufficient condition here is that $K = 2H - \gamma\mathbb{E}\left[X^TXXX^T\right] \succ 0$. Indeed, let $I = \mathrm{Ker}(A)^\perp$ be the orthogonal space of the kernel of A, which is stable under A as A is symmetric, so we can define A' the restriction of A to I which is invertible. It is of dimension greater than 1 as A is not 0. K defines on I a bilinear symmetric definite positive application K'. Then, $\mathrm{Tr}\left(A^TKA\right) = \mathrm{Tr}\left(A'^TK'A'\right) > 0$ because $A'^TK'A'$ is also symmetric definite positive.

Finally, we want $I - \gamma T \succ -I$. Using Lemma 4, this is a direct consequence of (2.3). □

2.2 Proof of Lemma 1

Let assume $0 < \gamma < \gamma_{\max}$. Lemma 5 already tells us that our operators have good properties as we have $\rho < 1$ and $T \succ 0$. We will now get a finer result in order to have an explicit bound on ρ depending on γ.

As we will be using different values for γ we will explicitely mark the dependency in γ for T by writing $T(\gamma)$. We will only consider $0 < \gamma < \gamma_{\max}$ so that $T(\gamma)$ is positive. We will denote by $L_{T(\gamma)}$ the largest eigenvalue of $T(\gamma)$ and by $\mu_{T(\gamma)}$ its smallest. We then have

$$\rho_T(\gamma) = \max(1 - \gamma\mu_{T(\gamma)}, \gamma L_{T(\gamma)} - 1).$$

One should also notice that the smallest eigenvalue of $H_L + H_R$ is 2μ and the largest $2L$.

We have $T(\gamma_{\max}) \succeq 0$ using Lemma 5. For any $0 < \gamma < \gamma_{\max}$ we can define $\alpha = \frac{\gamma}{\gamma_{\max}}$. Then we have

$$\begin{aligned}
T(\gamma) &= (1-\alpha)(H_L + H_R) + \alpha(H_L + H_R) - \alpha\gamma_{\max}M \\
&= (1-\alpha)(H_L + H_R) + \alpha T(\gamma_{\max}) \\
&\succeq (1-\alpha)(H_L + H_R) \\
&\succeq 2(1-\alpha)\mu,
\end{aligned}$$

so that $\mu_{T(\gamma)} \geq 2(1-\alpha)\mu$.

Using Lemma 4 we have that $T(\gamma_{\max}) \preceq \frac{2I}{\gamma_{\max}}$ so that we obtain

$$\begin{aligned}
T(\gamma) &= (1-\alpha)(H_L + H_R) + \alpha T(\gamma_{\max}) \\
&\preceq 2(1-\alpha)L + \frac{2\alpha}{\gamma_{\max}}.
\end{aligned}$$

As a consequence if we take

$$\begin{aligned}
a(\gamma) &= 1 - 2\alpha\gamma_{\max}(1-\alpha)\mu \\
b(\gamma) &= 2(1-\alpha)\alpha\gamma_{\max}L + 2\alpha^2 - 1
\end{aligned}$$

we have $\rho_T(\gamma) = \max(a(\gamma), b(\gamma))$. Besides, if we are in dimension $d = 2$ or more,

$$\begin{aligned}
a(\gamma) - b(\gamma) &= 2 - 2\alpha\gamma_{\max}(L+\mu)(1-\alpha) - 2\alpha^2 \\
&\geq 2 - 4\alpha(1-\alpha) - 2\alpha^2 \quad \text{as } \gamma_{\max}(L+\mu) \leq 2 \\
&= 2 + 2\alpha^2 - 4\alpha \\
&\geq 0,
\end{aligned}$$

so that,

$$\rho_T(\gamma) \leq 1 - 2\gamma\mu(1 - \frac{\gamma}{\gamma_{\max}}).$$

In dimension $d = 1$, the same result holds as we have $1 - \gamma T \geq 1 - 2\gamma H + \gamma^2 H^2 \geq 0$ so that $\gamma L_{T(\gamma)} - 1 \leq 0$.

We can now look as ρ_H which is given by $\rho_H = \max(1 - \gamma\mu, \gamma L - 1)$.

Let assume we are in dimension 2 or more, then we have $1 - \gamma\mu \geq \gamma L - 1$ so that $\rho_H = 1 - \gamma\mu$. In dimension 1, we have $\rho_H = |1 - \gamma\mu|$. Comparing ρ_H and ρ_T we obtain the result of this Lemma.

Finally, if $\gamma > \gamma_{\max}$, then T has a negative eigenvalue and so $\rho_T > 1$ and $\rho > 1$.

3 Proof of the theorems

We will first give a more complete version of both theorems.

Theorem 1 (Asymptotic covariance of the bias)**.** *Let $E_0 = \mathbb{E}\left[\eta_0\eta_0^T\right]$ (or just $\eta_0\eta_0^T$ if the starting point is not randomized). If $0 < \gamma < \gamma_{\max}$ and $\forall i \geq 1, \varepsilon_i = 0$, then*

$$\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] = \frac{1}{n^2\gamma^2}\left(H_L^{-1} + H_R^{-1} - \gamma I\right)\left(T^{-1}E_0\right) + O\left(\frac{\rho^n}{n}\right).$$

Theorem 2 (Asymptotic covariance of the variance)**.** *Let $\Sigma_0 = \mathbb{E}\left[\varepsilon^2 XX^T\right]$ and let assume that $\eta_0 = 0$. If $0 < \gamma < \gamma_{\max}$*

$$\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] = \frac{1}{n}(H_L^{-1} + H_R^{-1} - \gamma I)T^{-1}\Sigma_0 - \frac{1}{\gamma n^2}\left(H_L^{-1} + H_R^{-1} - \gamma I\right)(I - \gamma T)T^{-2}\Sigma_0 + O\left(\frac{\rho^n}{n}\right).$$

3.1 Complete expression of the covariance matrix

Let us recall that we have the update rule

$$\eta_i = (I - \gamma X_i X_i^T)\eta_i + \gamma\varepsilon_i X_i. \tag{3.1}$$

We can then introduce the following matrices

$$M_{k,j} = \left(\prod_{i=k+1}^{j}\left(I - \gamma X_i X_i^T\right)\right)^T \in \mathbb{R}^{d\times d},$$

and by iterating over (3.1) we obtain,

$$\eta_n = \gamma\sum_{k=1}^{n} M_{k,n}X_k\varepsilon_k + M_{0,n}\eta_0.$$

We have

$$\begin{aligned}\bar{\eta}_n &= \frac{\gamma}{n}\sum_{j=0}^{n-1}\sum_{k=1}^{j} M_{k,j}X_k\varepsilon_k + \frac{1}{n}\sum_{j=0}^{n-1} M_{0,j}\eta_0 \\ &= \frac{\gamma}{n}\sum_{k=1}^{n-1}\left(\sum_{j=k}^{n-1} M_{k,j}\right)X_k\varepsilon_k + \frac{1}{n}\sum_{j=0}^{n-1} M_{0,j}\eta_0.\end{aligned}$$

One can already see the decomposition between the variance and bias term, one depending only on η_0 and the other on ε.

If we assume that ε_k is independent of X_k, then we can immediately see that when computing $\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right]$, cross-terms between bias and variance will be zero as they will contain only one ε_k. If that is not true, then extra cross-terms will appear and there is no longer a simple bias/variance decomposition. Let us look at one of the cross terms,

$$\frac{\gamma}{n^2}\mathbb{E}\left[M_{k,j}X_k\varepsilon_k\eta_0^T M_{0,p}\right].$$

If $p < k$, then one can immediately notice that $X_k\varepsilon_k$ will be independant from the rest so that the term will be 0, as it is always true that $\mathbb{E}[\varepsilon X] = 0$. If not, X_k will also appear in $M_{0,p}$ as a factor $I - \gamma X_k X_k^T$ so that the term can be expressed as $G(\mathbb{E}\left[X_k\varepsilon_k\eta_0^T X_k X_k^T\right])$ where G is a linear

operator obtained using the independance of the other X_i and ε_i for $i \neq k$. As a consequence, we can recover a simple decomposition as soon as

$$\forall\, 1 \leq i, j, k \leq d, \mathbb{E}\left[X^{(i)}X^{(j)}X^{(k)}\varepsilon\right] = 0,$$

where $X^{(i)}$ is the i-th component of X.

In any case, because of Minkowski's inequality as noted in Bach and Moulines (2013), we always have that

$$f_n^{\text{total}} - f^* \leq 2(f_n^{\text{bias}} - f^*) + 2(f_n^{\text{variance}} - f^*),$$

so that we are never too far from the true error when assuming X and ε independant.

3.2 Proof for the bias term

First, let us assume that $\varepsilon_k = 0$ a.s. Then we have

$$\bar{\eta}_n = \frac{1}{n}\sum_{j=0}^{n-1} M_{0,j}\eta_0,$$

and

$$\begin{aligned}
\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] =& \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,j}^T\right] \\
=& \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T + \sum_{j=i+1}^{n-1} M_{0,i}\eta_0\eta_0^T M_{0,i}^T M_{i,j}^T + \sum_{j=0}^{i-1} M_{j,i}M_{0,j}\eta_0\eta_0^T M_{0,j}^T\right]\right) \\
=& \frac{1}{n^2}\sum_{i=0}^{n-1}\Bigg(\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right] + \sum_{j=i+1}^{n-1}\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right](I-\gamma H)^{j-i} \\
&+ \sum_{j=0}^{i-1}(I-\gamma H)^{i-j}\mathbb{E}\left[M_{0,j}^T\eta_0\eta_0^T M_{0,j}\right]\Bigg) \text{ because of independence assumptions,} \\
=& \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right] + \sum_{j=i+1}^{n-1}\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right](I-\gamma H)^{j-i}\right) \\
&+ \frac{1}{n^2}\sum_{j=0}^{n-1}\left(\sum_{i=j+1}^{n-1}(I-\gamma H)^{i-j}\mathbb{E}\left[M_{0,j}^T\eta_0\eta_0^T M_{0,j}\right]\right) \\
=& \frac{1}{n^2}\sum_{i=0}^{n-1}\Bigg(\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right] \\
&+ \sum_{j=i+1}^{n-1}\left(\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right](I-\gamma H)^{j-i} + (I-\gamma H)^{j-i}\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right]\right)\Bigg) \\
&\text{by exchanging the role of } i \text{ and } j \text{ in the last equation,} \\
=& \frac{1}{n^2}\sum_{i=0}^{n-1}\Bigg(\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right] \\
&+ \mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right]\left((I-\gamma H) - (I-\gamma H)^{n-i}\right)(\gamma H)^{-1} \\
&+ (\gamma H)^{-1}\left((I-\gamma H) - (I-\gamma H)^{n-i}\right)\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right]\Bigg).
\end{aligned}$$

We only used the fact that X_i and X_j are independent as soon as $i \neq j$, so that we can condition on $X_1, \dots X_i$ to obtain $M_{1,i}(I-\gamma H)^{j-i}$. Now we need to express $\mathbb{E}\left[(I-\gamma X_i X_i^T)A(I-\gamma X_i X_i^T)\right]$ for A some matrix that is independent of X_i. Using the notation we introduced, we have immediately that

$$\begin{aligned}
\mathbb{E}\left[(I-\gamma X_i X_i^T)A(I-\gamma X_i X_i^T)\right] &= A - \gamma AH - \gamma HA + \gamma^2\mathbb{E}\left[X^T A X X X^T\right] \\
&= (I - \gamma H_R - \gamma H_L + \gamma^2 M)A \\
&= (I-\gamma T)A.
\end{aligned}$$

Then we have, with $\mathcal{F}_{i-1}$ the σ field generated by $X_1, \ldots, X_{i-1}$,

$$\begin{aligned}
\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T\right] &= \mathbb{E}\left[\mathbb{E}\left[M_{0,i}\eta_0\eta_0^T M_{0,i}^T|\mathcal{F}_{i-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(I-\gamma X_i X_i^T)M_{0,i-1}\eta_0\eta_0^T M_{0,i-1}^T(I-\gamma X_i X_i^T)|\mathcal{F}_{i-1}\right]\right] \\
&= \mathbb{E}\left[(I-\gamma T)M_{0,i-1}\eta_0\eta_0^T M_{0,i-1}\right] \\
&= (I-\gamma T)\mathbb{E}\left[M_{0,i-1}\eta_0\eta_0^T M_{0,i-1}\right].
\end{aligned}$$

and by iterating this process, we obtain

$$\begin{aligned}
\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] &= \frac{1}{n^2}\sum_{i=0}^{n-1}(I-\gamma T)^i E_0 \\
&\quad + \left((I-\gamma T)^i E_0\right)\left((I-\gamma H)-(I-\gamma H)^{n-i}\right)(\gamma H)^{-1} \\
&\quad + (\gamma H)^{-1}\left((I-\gamma H)-(I-\gamma H)^{n-i}\right)\left((I-\gamma T)^i E_0\right) \\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\Bigg(I + \left[(I-\gamma H)_L - (I-\gamma H)_L^{n-i}\right](\gamma H_L)^{-1} \\
&\quad + \left[(I-\gamma H)_R - (I-\gamma H)_R^{n-i}\right](\gamma H_R)^{-1}\Bigg)(I-\gamma T)^i E_0.
\end{aligned}$$

Let us define

$$\begin{aligned}
A_n &= -\frac{1}{n^2}\sum_{i=0}^{n-1}\left((\gamma H_R)^{-1}(I-\gamma H)_R^{n-i} + (\gamma H_L)^{-1}(I-\gamma H)_L^{n-i}\right)\left((I-\gamma T)^i E_0\right) \\
\|A_n\|_F &\leq \frac{2d}{n\gamma\mu}\rho^n \|E_0\|_F,
\end{aligned}$$

which is decaying exponentially. We now have

$$\begin{aligned}
\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] &= \frac{1}{n^2}\sum_{i=0}^{n-1}\left(I + (I-\gamma H_L)(\gamma H_L)^{-1} + (I-\gamma H_R)(\gamma H_R)^{-1}\right)(I-\gamma T)^i E_0 + A_n \\
&= \frac{1}{\gamma^2 n^2}\left(H_L^{-1} + H_R^{-1} - \gamma I\right)T^{-1}\left(I-(I-\gamma T)^n\right)E_0 + A_n.
\end{aligned}$$

Again, we have some exponential terms, that we will regroup in B_n with

$$\begin{aligned}
B_n &= -\frac{1}{\gamma^2 n^2}\left(H_L^{-1} + H_R^{-1} - \gamma I\right)T^{-1}(I-\gamma T)^n E_0 \\
\|B_n\|_F &\leq \frac{d}{n^2\gamma^2\mu_T}\rho_T^n\left(\frac{2}{\mu}-\gamma\right)\|E_0\|_F,
\end{aligned}$$

and we have

$$\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] = \frac{1}{n^2\gamma^2}\left(H_L^{-1} + H_R^{-1} - \gamma I\right)T^{-1}E_0 + A_n + B_n.$$

We can bound $A_n + B_n$ by

$$\|A_n + B_n\|_F \leq \frac{d\rho^n\|E_0\|_F}{\gamma n}\left(\frac{2}{\mu} + \frac{1}{\mu_T n\gamma}\left(\frac{2}{\mu}-\gamma\right)\right),$$

which completes the first assertion of Theorem 1.

3.3 Proof for the variance term

Let assume now that $\eta_0 = 0$, then we have

$$\bar{\eta}_n = \frac{\gamma}{n}\sum_{k=1}^{n-1}\left(\sum_{j=k}^{n-1} M_{k,j}\right)X_k\varepsilon_k,$$

and

$$\begin{aligned}\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] &= \frac{\gamma^2}{n^2}\mathbb{E}\left[\sum_{k,l=1}^{n-1}\left(\sum_{j=k}^{n-1}M_{k,j}\right)X_k\varepsilon_k\varepsilon_l X_l^T\left(\sum_{p=l}^{n-1}M_{l,p}^T\right)\right]\\ &= \frac{\gamma^2}{n^2}\mathbb{E}\left[\sum_{k=1}^{n-1}\left(\sum_{j=k}^{n-1}M_{k,j}\right)X_k\varepsilon_k\varepsilon_k X_k^T\left(\sum_{p=k}^{n-1}M_{k,p}^T\right)\right].\end{aligned}$$

Indeed, we can remove terms where $k \neq l$: if we have for instance $l < k$, then $X_l\varepsilon_l$ will be independent from the rest of the terms and as $\mathbb{E}\left[X_l\varepsilon_l\right] = 0$, the term will be 0.

By using mostly the same method as for the bias term, we obtain that

$$\begin{aligned}\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] =&\frac{\gamma^2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k}^{n-1}(I-\gamma T)^{j-k}\,\Sigma_0\\ &+\left((I-\gamma H)-(I-\gamma H)^{n-j}\right)(\gamma H)^{-1}\left((I-\gamma T)^{j-k}\,\Sigma_0\right)\\ &+\left((I-\gamma T)^{j-k}\,\Sigma_0\right)\left((I-\gamma H)-(I-\gamma H)^{n-j}\right)(\gamma H)^{-1}\\ =&\frac{\gamma^2}{n^2}\sum_{j=1}^{n-1}\sum_{k=1}^{j}(I-\gamma T)^{j-k}\,\Sigma_0\\ &+\left((I-\gamma H)-(I-\gamma H)^{n-j}\right)(\gamma H)^{-1}\left((I-\gamma T)^{j-k}\,\Sigma_0\right)\\ &+\left((I-\gamma T)^{j-k}\,\Sigma_0\right)\left((I-\gamma H)-(I-\gamma H)^{n-j}\right)(\gamma H)^{-1}\\ =&\frac{\gamma^2}{n^2}\sum_{j=1}^{n-1}\left(I-(I-\gamma T)^j\right)(\gamma T)^{-1}\Sigma_0\\ &+\left((I-\gamma H)-(I-\gamma H)^{n-j}\right)(\gamma H)^{-1}\left(I-(I-\gamma T)^j\right)(\gamma T)^{-1}\Sigma_0\\ &+\left(I-(I-\gamma T)^j\right)(\gamma T)^{-1}\Sigma_0\left((I-\gamma H)-(I-\gamma H)^{n-j}\right)(\gamma H)^{-1}.\end{aligned}$$

As for the bias, we can bound some terms:

$$\begin{aligned}C_n &= \frac{\gamma^2}{n^2}\sum_{j=1}^{n-1}\left((I-\gamma H)_L^{n-j}(\gamma H_L)^{-1}+(I-\gamma H)_R^{n-j}(\gamma H_R)^{-1}\right)(I-\gamma T)^j(\gamma T)^{-1}\Sigma_0\\ \left\|C_n\right\|_F &\leq \frac{2d}{n\mu\mu_T}\rho^n\left\|\Sigma_0\right\|_F.\end{aligned}$$

Now we have,

$$\begin{aligned}\mathbb{E}\left[\bar{\eta}_n\bar{\eta}_n^T\right] =&\frac{1}{n^2}\sum_{j=1}^{n-1}\left(H_L^{-1}+H_R^{-1}-\gamma I\right)\left(I-(I-\gamma T)^j\right)T^{-1}\Sigma_0+C_n\\ =&\frac{1}{n}\left(H_L^{-1}+H_R^{-1}-\gamma I\right)T^{-1}\Sigma_0+D_n+C_n,\end{aligned}$$

where D_n is defined by

$$\begin{aligned}D_n =& -\frac{1}{n^2}\sum_{j=1}^{n-1}\left(H_L^{-1}+H_R^{-1}-\gamma I\right)(I-\gamma T)^jT^{-1}\Sigma_0\\ =& -\frac{1}{\gamma n^2}\left(H_L^{-1}+H_R^{-1}-\gamma I\right)(I-\gamma T)T^{-2}\Sigma_0+D_n'.\end{aligned}$$

D_n' are again exponentially decreasing terms:

$$\begin{aligned}D_n' =&\frac{1}{\gamma n^2}\left(H_L^{-1}+H_R^{-1}-\gamma I\right)(I-\gamma T)^nT^{-2}\Sigma_0\\ \left\|D_n'\right\|_F &\leq \frac{d}{\gamma^2\mu_T^2 n}\left(\frac{2}{\mu}-\gamma\right)\rho_T^n\left\|\Sigma_0\right\|_F,\end{aligned}$$

so that we have

$$\mathbb{E}\left[\bar{\eta}_n \bar{\eta}_n^T\right] = \frac{1}{n}\left(H_L^{-1} + H_R^{-1} - \gamma I\right) T^{-1}\Sigma_0 - \frac{1}{\gamma n^2}\left(H_L^{-1} + H_R^{-1} - \gamma I\right)(I - \gamma T)T^{-2}\Sigma_0 + C_n + D'_n. \tag{3.2}$$

We can bound $C_n + D'_n$ by

$$\|C_n + D'_n\|_F \leq \frac{d\rho^n \|\Sigma_0\|_F}{n}\left(\frac{1}{n\gamma\mu_T^2}\left(\frac{2}{\mu} - \gamma\right) + \frac{2}{\mu\mu_T}\right).$$

This concludes the proof of Theorem 2.

References

Bach, F. and E. Moulines (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Adv. NIPS*.