Bias-Variance Trade-offs for Averaged SGD in Least Mean Squares

Raghav Somani

July 4, 2018

Large scale optimization has been driven mostly by Stochastic Gradient Descent (SGD) based algorithms especially in machine learning problems. The interplay between optimization and generalization is therefore crucial. Under stochasticity, the generalization error can be broken down explicitly into two quantities, "Bias" and "Variance". Here is an attempt to help the reader understand one of the critical mathematical tools involved in a tight analysis of Averaged SGD on a simple toy model - Least Mean Square. The article is based on the novel techniques developed in [1].

A detailed asymptotic analysis of the averaged constant step size SGD algorithm for Least Mean Square has been done in [1] which has been presented in this article. We will first set up the problem and the algorithm formally for the reader's convenience.

1 Problem Setup

Throughout the article, we will follow certain conventions listed below.

- If $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ are random variables, we denote $\mathbf{H} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]$ as its second order moment matrix.
- The smallest eigenvalue of $\mathbf{H}$ is μ and is assumed to be strictly positive. Therefore, $\mathbf{H}$ is invertible.
- $\mathcal{M}(\mathbb{R}^d)$ denotes the set of all linear operators over $\mathbb{R}^d$ which is isomorphic to the space of matrices in $\mathbb{R}^{d\times d}$, and similarly $\mathcal{I} = \mathcal{M}(\mathcal{M}(\mathbb{R}^d))$ denotes an endomorphism on the space of matrices over $\mathbb{R}^d$. Therefore if $\mathcal{T} \in \mathcal{I}$, then it can be thought of as a matrix of matrices that is essentially in $\mathbb{R}^{d\times d\times d\times d}$. Just like matrix multiplication, we can define tensor-matrix multiplication as

$$(\mathcal{T}\mathbf{A}) = \sum_{(k,l)\in[d]\times[d]} \mathcal{T}_{(i,j),(k,l)}\mathbf{A}_{k,l} \tag{1.0.1}$$

  where $\mathbf{A} \in \mathbb{R}^{d\times d}$.

- We can define operator norms of tensors restricted to symmetric matrices with respect to Frobenius norm as

$$\|\mathcal{T}\|_{\text{op}} = \sup_{\mathbf{V}\in\mathcal{S}(\mathbb{R}^d),\|\mathbf{V}\|_F=1} \|\mathcal{T}\mathbf{V}\|_F \tag{1.0.2}$$

  where $\mathcal{S}(\mathbb{R}^d)$ is the set of all symmetric matrices on $\mathbb{R}^d$.

- We denote $\mathcal{T}_L$ and $\mathcal{T}_R$ as the left and right matrix multiplication operators respectively of the matrix $\mathbf{T} \in \mathbb{R}^{d\times d}$, defined as

$$\begin{aligned} \forall\ (i,j),(k,l), \qquad (\mathcal{T}_L)_{(i,j),(k,l)} &= \delta_{j,l}\mathbf{T}_{i,k} \\ \forall\ (i,j),(k,l), \qquad (\mathcal{T}_R)_{(i,j),(k,l)} &= \delta_{i,k}\mathbf{T}_{j,l} \end{aligned}$$

  One can use the tensor-matrix multiplication definition (1.0.1) to show that the above definitions of $\mathcal{T}_L$ and $\mathcal{T}_R$ of left and right multiplication operators of $\mathbf{T}$ satisfy

$$\begin{aligned} \mathcal{T}_L\mathbf{A} &= \mathbf{T}\mathbf{A} \\ \mathcal{T}_R\mathbf{A} &= \mathbf{A}\mathbf{T} \end{aligned} \tag{1.0.3}$$

- Let a linear operator on matrices be defined as

$$\mathcal{M}\mathbf{A} = \mathbb{E}\left[(\mathbf{x}^T\mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^T\right] \tag{1.0.4}$$

  then the elements in $\mathcal{M}$ can be written as

$$\mathcal{M}_{(i,j),(k,l)} = \mathbb{E}\left[\mathbf{x}^{(i)}\mathbf{x}^{(j)}\mathbf{x}^{(k)}\mathbf{x}^{(l)}\right] \tag{1.0.5}$$

  making $\mathcal{M}$, the fourth order moment tensor of the random variable $\mathbf{x} \in \mathbb{R}^d$.

- Define $\mathcal{T} = \mathcal{H}_L + \mathcal{H}_R - \eta\mathcal{M}$ with μ_T its smallest eigenvalue.
- $\rho_T = \|\mathbf{I} - \eta\mathcal{T}\|_{\text{op}}$, $\rho_H = \|\mathbf{I} - \eta\mathbf{H}\|_{\text{op}}$ and $\rho = \max\{\rho_T, \rho_H\}$.

The Least Mean Squares problem is to minimize the expected quadratic loss

$$f(\mathbf{w}) = \mathbb{E}\left[\frac{1}{2}\left\|\mathbf{x}^T\mathbf{w} - y\right\|_2^2\right] \tag{1.0.6}$$

Let $\mathbf{w}^*$ be the optimum solution to the problem (1.0.6). Because $\mathbf{H}$ is invertible, so $f(\mathbf{w})$ has a unique minimum that is $f^* = f(\mathbf{w}^*)$. The problem has 2 regimes based on the size of the domain of the random variables $\mathbf{x}$ and y.

1. Single pass through the data, where each example is seen once and considered as an i.i.d. sample.
2. Multiple passes through the data, which is of interest when the number of data points is finite.

The first case is explicitly studied in [1] which we discuss in detail from here on.

1.1 Averaged SGD with constant step size

Let $\mathbf{w}_0 \in \mathbb{R}^d$ be an initial point and at each iteration i, we sample an i.i.d. instance $(\mathbf{x}_i, y_i)$ of $\mathbf{x}$ and y respectively. Let η be the constant step size for an SGD update

$$\begin{aligned}
\mathbf{w}_i &= \mathbf{w}_{i-1} - \eta\nabla f(\mathbf{w}_{i-1}; \mathbf{x}_i, y_i) \\
&= \mathbf{w}_{i-1} - \eta\mathbf{x}_i(\mathbf{x}_i^T\mathbf{w}_{i-1} - y_i) \\
\bar{\mathbf{w}}_i &= \frac{1}{i+1}\sum_{k=0}^{i}\mathbf{w}_k \\
&= \frac{1}{i+1}\mathbf{w}_i + \frac{i}{i+1}\bar{\mathbf{w}}_{i-1}
\end{aligned} \tag{1.1.1}$$

Let us define a few more variables

- $\varepsilon_i = \mathbf{x}_i^T\mathbf{w}^* - y_i \implies \mathbb{E}[\varepsilon_i\mathbf{x}_i] = 0$ since $\nabla f(w^*) = \mathbf{0}$.
- $\Delta_i = \mathbf{w}_i - \mathbf{w}^*$, $\bar{\Delta}_i = \bar{\mathbf{w}}_i - \mathbf{w}^*$

Therefore we get

$$\Delta_i = (\mathbf{I} - \eta\mathbf{x}_i\mathbf{x}_i^T)\Delta_{i-1} + \eta\varepsilon_i\mathbf{x}_i \tag{1.1.2}$$

an operator style of writing a recursive update rule.

Let us introduce $\mathbf{M}_{k,j} = \left(\prod_{i=k+1}^{j}\left(\mathbf{I} - \eta\mathbf{x}_i\mathbf{x}_i^T\right)\right)^T \in \mathbb{R}^{d\times d}$ as a compact matrix operator acting on the terms in (1.1.2). Upon unrolling the recursion (1.1.2), we get

$$\begin{aligned}
\Delta_n &= \eta\sum_{k=1}^{n}\mathbf{M}_{k,n}\mathbf{x}_k\varepsilon_k + M_{0,n}\Delta_0 \\
\bar{\Delta}_n &= \frac{\eta}{n}\sum_{j=0}^{n-1}\sum_{k=1}^{j}\mathbf{M}_{k,j}\mathbf{x}_k\varepsilon_k + \frac{1}{n}\sum_{j=0}^{n-1}\mathbf{M}_{0,j}\Delta_0
\end{aligned}$$

$$= \frac{\eta}{n}\sum_{k=1}^{n-1}\left(\sum_{j=k}^{n-1}\mathbf{M}_{k,j}\right)\mathbf{x}_k\varepsilon_k + \frac{1}{n}\sum_{j=0}^{n-1}\mathbf{M}_{0,j}\Delta_0 \qquad \text{(Re-arranging summations)} \tag{1.1.3}$$

We can clearly see that the error $\bar{\Delta}_n$ can be decomposed in two terms, one dominated by Δ_0 and the other depending on the noise ε_k's. The cross term in $\mathbb{E}\left[\bar{\Delta}_n\bar{\Delta}_n^T\right]$ can be separately written as

$$\frac{\eta}{n^2}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{x}_k\varepsilon_k\Delta_0^T\mathbf{M}_{0,p}\right] \tag{1.1.4}$$

When $p < k$, then we take the expectation inside and write the above term as

$$\frac{\eta}{n^2}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{x}_k\varepsilon_k\Delta_0^T\mathbf{M}_{0,p}\right] = \frac{\eta}{n^2}\mathbf{M}_{k,j}\mathbb{E}\left[\mathbf{x}_k\varepsilon_k\right]\Delta_0^T\mathbf{M}_{0,p} = \mathbf{0}$$

Else, $\mathbf{x}_k$ appears in $\mathbf{M}_{0,p}$ and we will have a term that can be expressed as $G\left(\mathbb{E}\left[\mathbf{x}_k\varepsilon_k\Delta_0^T\mathbf{x}_k\mathbf{x}_k^T\right]\right)$ where G is a linear operator, which is $\mathbf{0}$ as soon as

$$\forall\ 1 \le i, j, k \le d, \quad \mathbb{E}\left[\mathbf{x}^{(i)}\mathbf{x}^{(j)}\mathbf{x}^{(k)}\varepsilon\right] = 0$$

Because we consider the least squares problem, we have

$$f_n - f^* = \frac{1}{2}\mathbb{E}\left[(\bar{\mathbf{w}}_n - \mathbf{w}^*)^T\mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right](\bar{\mathbf{w}}_n - \mathbf{w}^*)\right] = \operatorname{Tr}\left[\mathbf{H}\mathbb{E}\left[\bar{\Delta}_n\bar{\Delta}_n^T\right]\right] \tag{1.1.5}$$

Since we want a Bias-Variance trade-off, we achieve it by considering two scenarios

1. The bias term $\varDelta^{\text{bias}}$ which is the covariance matrix when $\varepsilon_i = 0\ \forall\ i$.
2. The variance term $\varDelta^{\text{variance}}$ which is the covariance matrix when we start at the solution itself, $\Delta_0 = \mathbf{0}$.

Therefore,

$$f_n - f^* = \operatorname{Tr}\left[\mathbf{H}\varDelta^{\text{bias}}\right] + \operatorname{Tr}\left[\mathbf{H}\varDelta^{\text{variance}}\right] \tag{1.1.6}$$

Even when it is not true, we can still use Minkowski's inequality to have

$$f_n^{\text{total}} - f^* \le 2(f_n^{\text{bias}} - f^*) + 2(f_n^{\text{variance}} - f^*) \tag{1.1.7}$$

1.2 Bias term

Assuming $\varepsilon_i = 0\ \forall\ i$, we have

$$\bar{\Delta}_n = \frac{1}{n}\sum_{j=0}^{n-1}\mathbf{M}_{0,j}\Delta_0 \tag{1.2.1}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\bar{\Delta}_n\bar{\Delta}_n^T\right] &= \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,j}^T\right] \\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T + \sum_{j=i+1}^{n-1}\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\mathbf{M}_{i,j}^T + \sum_{j=0}^{i-1}\mathbf{M}_{j,i}\mathbf{M}_{0,j}\Delta_0\Delta_0^T\mathbf{M}_{0,j}^T\right]\right) \\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right] + \sum_{j=i+1}^{n-1}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right](\mathbf{I} - \eta\mathbf{H})^{j-i} + \sum_{j=0}^{i-1}(\mathbf{I} - \eta\mathbf{H})^{i-j}\mathbb{E}\left[\mathbf{M}_{0,j}\Delta_0\Delta_0^T\mathbf{M}_{0,j}^T\right]\right) \\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right] + \sum_{j=i+1}^{n-1}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right](\mathbf{I} - \eta\mathbf{H})^{j-i}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{n^2}\sum_{i=0}^{n-1}\left(\sum_{j=0}^{i-1}(\mathbf{I}-\eta\mathbf{H})^{i-j}\mathbb{E}\left[\mathbf{M}_{0,j}\Delta_0\Delta_0^T\mathbf{M}_{0,j}^T\right]\right)\\
&=\frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]+\sum_{j=i+1}^{n-1}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right](\mathbf{I}-\eta\mathbf{H})^{j-i}\right)\\
&\quad+\frac{1}{n^2}\sum_{j=0}^{n-1}\left(\sum_{i=j+1}^{n-1}(\mathbf{I}-\eta\mathbf{H})^{i-j}\mathbb{E}\left[\mathbf{M}_{0,j}\Delta_0\Delta_0^T\mathbf{M}_{0,j}^T\right]\right)\\
&=\frac{1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]+\sum_{i=0}^{n-1}\left(\sum_{j=i+1}^{n-1}\left(\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right](\mathbf{I}-\eta\mathbf{H})^{j-i}+(\mathbf{I}-\eta\mathbf{H})^{j-i}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]\right)\right)\\
&=\frac{1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]+\sum_{i=0}^{n-1}\left(\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]\left((\mathbf{I}-\eta\mathbf{H})-(\mathbf{I}-\eta\mathbf{H})^{n-i}\right)(\eta\mathbf{H})^{-1}\right)\\
&\quad+\sum_{i=0}^{n-1}\left((\eta\mathbf{H})^{-1}\left((\mathbf{I}-\eta\mathbf{H})-(\mathbf{I}-\eta\mathbf{H})^{n-i}\right)\mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]\right) \qquad (1.2.2)
\end{aligned}
$$

To simplify (1.2.2) further, we can use the definition of $\mathcal{T}$. Let $\mathbf{A}$ be any matrix, then

$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{I}-\eta\mathbf{x}_i\mathbf{x}_i^T)\mathbf{A}(I-\eta\mathbf{x}_i\mathbf{x}_i^T)\right]&=\mathbf{A}-\eta(\mathbf{AH}+\mathbf{HA})+\eta^2\mathbb{E}\left[(\mathbf{x}^T\mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^T\right]\\
&=(\mathbf{I}-\eta\mathcal{H}_R-\eta\mathcal{H}_L+\eta^2\mathcal{M})\mathbf{A}\\
&=(\mathbf{I}-\eta\mathcal{T})\mathbf{A}\\
\therefore\ \mathbb{E}\left[\mathbf{M}_{0,i}\Delta_0\Delta_0^T\mathbf{M}_{0,i}^T\right]&=(\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0 \qquad (1.2.3)
\end{aligned}
$$

where $\mathcal{E}_0=\Delta_0\Delta_0^T$. Using (1.2.3) in (1.2.2) we get

$$
\begin{aligned}
\mathbb{E}\left[\bar{\Delta}_n\bar{\Delta}_n^T\right]&=\frac{1}{n^2}\sum_{i=0}^{n-1}(I-\eta\mathcal{T})^i\mathcal{E}_0+\sum_{i=0}^{n-1}\left((\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0\left((\mathbf{I}-\eta\mathbf{H})-(\mathbf{I}-\eta\mathbf{H})^{n-i}\right)(\eta\mathbf{H})^{-1}\right)\\
&\quad+\sum_{i=0}^{n-1}\left((\eta\mathbf{H})^{-1}\left((\mathbf{I}-\eta\mathbf{H})-(\mathbf{I}-\eta\mathbf{H})^{n-i}\right)(\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0\right)\\
&=\frac{1}{n^2}\sum_{i=0}^{n-1}(\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0+\left((\mathbf{I}-\eta\mathcal{H})_R-(\mathbf{I}-\eta\mathcal{H})_R^{n-i}\right)(\eta\mathcal{H}_R)^{-1}(\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0\\
&\quad+\sum_{i=0}^{n-1}\left((\mathbf{I}-\eta\mathcal{H})_L-(\mathbf{I}-\eta\mathcal{H})_L^{n-i}\right)(\eta\mathcal{H}_L)^{-1}(\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0\\
&=\frac{1}{n^2}\sum_{i=0}^{n-1}\left[\mathbf{I}+\left((\mathbf{I}-\eta\mathcal{H})_L-(\mathbf{I}-\eta\mathcal{H})_L^{n-i}\right)(\eta\mathcal{H}_L)^{-1}+\left((\mathbf{I}-\eta\mathcal{H})_R-(\mathbf{I}-\eta\mathcal{H})_R^{n-i}\right)(\eta\mathcal{H}_R)^{-1}\right](\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0\\
&\qquad (1.2.4)
\end{aligned}
$$

We can now try to analyze separate term in (1.2.4). Let us define A_n as

$$
A_n=-\frac{1}{n^2}\sum_{i=0}^{n}\left[(\eta\mathcal{H}_R)^{-1}(\mathbf{I}-\eta\mathcal{H})_R^{n-i}+(\eta\mathcal{H}_L)^{-1}(\mathbf{I}-\eta\mathcal{H})_L^{n-i}\right](\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0 \qquad (1.2.5)
$$

Using (1.2.4) and (1.2.5) we have

$$
\begin{aligned}
\mathbb{E}\left[\Delta_n\Delta_n^T\right]&=\frac{1}{n^2}\sum_{i=0}^{n}\left[\mathbf{I}+(\mathbf{I}-\eta\mathcal{H})_L(\eta\mathcal{H}_L)^{-1}+(\mathbf{I}-\eta\mathcal{H})_R(\eta\mathcal{H}_R)^{-1}\right](\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0+A_n\\
&=\frac{1}{n^2}\sum_{i=0}^{n}\left[\mathbf{I}+\frac{1}{\eta}(\mathcal{H}_L^{-1}-\eta\mathbf{I})+\frac{1}{\eta}(\mathcal{H}_R^{-1}-\eta\mathbf{I})\right](\mathbf{I}-\eta\mathcal{T})^i\mathcal{E}_0+A_n
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\eta n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta \mathbf{I}\right] \sum_{i=0}^{n-1} (\mathbf{I} - \eta \mathcal{T})^i \mathcal{E}_0 + A_n \\
&= \frac{1}{\eta^2 n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta \mathbf{I}\right] \mathcal{T}^{-1} \left[\mathbf{I} - (\mathbf{I} - \eta \mathcal{T})^n\right] \mathcal{E}_0 + A_n
\end{aligned} \tag{1.2.6}
$$

Defining B_n as

$$
B_n = -\frac{1}{\eta^2 n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta \mathbf{I}\right] \mathcal{T}^{-1} (\mathbf{I} - \eta \mathcal{T})^n \mathcal{E}_0 \tag{1.2.7}
$$

Using (1.2.7) in (1.2.6) we get

$$
\mathbb{E}\left[\Delta_n \Delta_n^T\right] = \frac{1}{\eta^2 n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta \mathbf{I}\right] \mathcal{T}^{-1} \mathcal{E}_0 + A_n + B_n \tag{1.2.8}
$$

We will now analyze the Frobenius norms of A_n and B_n and show that they decay exponentially.

$$
\begin{aligned}
\|A_n\|_F &= \frac{1}{n^2}\left\|\sum_{i=0}^{n}\left[(\eta \mathcal{H}_R)^{-1}(\mathbf{I} - \eta \mathcal{H})_R^{n-i} + (\eta \mathcal{H}_L)^{-1}(\mathbf{I} - \eta \mathcal{H})_L^{n-i}\right](\mathbf{I} - \eta \mathcal{T})^i \mathcal{E}_0\right\|_F \\
&\le \frac{1}{n^2}\sum_{i=0}^{n}\left\|\left[(\eta \mathcal{H}_R)^{-1}(\mathbf{I} - \eta \mathcal{H})_R^{n-i} + (\eta \mathcal{H}_L)^{-1}(\mathbf{I} - \eta \mathcal{H})_L^{n-i}\right](\mathbf{I} - \eta \mathcal{T})^i \mathcal{E}_0\right\|_F \\
&\le \frac{1}{n^2}\sum_{i=0}^{n}\left\|(\eta \mathcal{H}_R)^{-1}(\mathbf{I} - \eta \mathcal{H})_R^{n-i}(\mathbf{I} - \eta \mathcal{T})^i \mathcal{E}_0\right\|_F + \left\|(\eta \mathcal{H}_L)^{-1}(\mathbf{I} - \eta \mathcal{H})_L^{n-i}(\mathbf{I} - \eta \mathcal{T})^i \mathcal{E}_0\right\|_F \\
&\le \frac{1}{n^2}\sum_{i=0}^{n}\left[\frac{1}{\eta}\left\|\mathcal{H}_R^{-1}\right\|_{\text{op}} \|(\mathbf{I} - \eta \mathcal{H})_R\|_{\text{op}}^{n-i} \|\mathbf{I} - \eta \mathcal{T}\|_{\text{op}}^{i} \|\mathcal{E}_0\|_F + \frac{1}{\eta}\left\|\mathcal{H}_L^{-1}\right\|_{\text{op}} \|(\mathbf{I} - \eta \mathcal{H})_L\|_{\text{op}}^{n-i} \|\mathbf{I} - \eta \mathcal{T}\|_{\text{op}}^{i} \|\mathcal{E}_0\|_F\right] \\
&= \frac{\rho^n}{\eta n^2}\sum_{i=0}^{n}\left[\left\|\mathcal{H}_R^{-1}\right\|_{\text{op}} + \left\|\mathcal{H}_L^{-1}\right\|_{\text{op}}\right] \|\mathcal{E}_0\|_F
\end{aligned} \tag{1.2.9}
$$

From the definition of operator norms of tensors

$$
\left\|\mathcal{H}_R^{-1}\right\|_{\text{op}} = \sup_{\mathbf{A} \ne \mathbf{0}} \frac{\left\|\mathcal{H}_R^{-1}\mathbf{A}\right\|_F}{\|\mathbf{A}\|_F} = \sup_{\mathbf{A} \ne \mathbf{0}} \frac{\left\|\mathbf{A}\mathbf{H}^{-1}\right\|_F}{\|\mathbf{A}\|_F} \le \sup_{\mathbf{A} \ne \mathbf{0}} \frac{\|\mathbf{A}\|_2 \left\|\mathbf{H}^{-1}\right\|_F}{\|\mathbf{A}\|_F} = \left\|\mathbf{H}^{-1}\right\|_F \le \frac{d}{\mu} \tag{1.2.10}
$$

$$
\left\|\mathcal{H}_L^{-1}\right\|_{\text{op}} = \sup_{\mathbf{A} \ne \mathbf{0}} \frac{\left\|\mathcal{H}_L^{-1}\mathbf{A}\right\|_F}{\|\mathbf{A}\|_F} = \sup_{\mathbf{A} \ne \mathbf{0}} \frac{\left\|\mathbf{H}^{-1}\mathbf{A}\right\|_F}{\|\mathbf{A}\|_F} \le \sup_{\mathbf{A} \ne \mathbf{0}} \frac{\left\|\mathbf{H}^{-1}\right\|_2 \|\mathbf{A}\|_F}{\|\mathbf{A}\|_F} = \left\|\mathbf{H}^{-1}\right\|_2 \le \frac{1}{\mu} \tag{1.2.11}
$$

Combining (1.2.10) and (1.2.11) we get

$$
\left\|\mathcal{H}_R^{-1}\right\|_{\text{op}} + \left\|\mathcal{H}_L^{-1}\right\|_{\text{op}} \le \frac{d+1}{\mu} \le \frac{2d}{\mu} \tag{1.2.12}
$$

Using (1.2.12) in (1.2.9) we have

$$
\|A_n\|_F \le \frac{2d\rho^n}{n\eta\mu} \tag{1.2.13}
$$

Assuming $\eta \le \frac{2}{\text{Tr}[\mathbf{H}]} \le \frac{2}{\rho_H} \le \frac{2}{\mu}$, we have $\left\|\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta \mathbf{I}\right\|_{\text{op}} \le d\left(\frac{2}{\mu} - \eta\right)$. Therefore, we can bound the Frobenius norm of B_n as

$$
\|B_n\|_F \le \frac{d\rho^n}{\eta^2 n^2 \mu_T}\left(\frac{2}{\mu} - \eta\right) \|\mathcal{E}_0\|_F \tag{1.2.14}
$$

Using (1.2.13) and (1.2.14) we get

$$
\|A_n + B_n\|_F \le \frac{d\rho^n}{\eta n}\left(\frac{2}{\mu} + \frac{1}{n\mu_T\eta}\left(\frac{2}{\mu} - \eta\right)\right) \tag{1.2.15}
$$

Therefore if $\rho < 1$, then term $A_n + B_n$ goes to 0 exponentially. Then (1.2.8) becomes

$$\mathbb{E}\left[\Delta_n \Delta_n^T\right] = \frac{1}{\eta^2 n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right]\mathcal{T}^{-1}\mathcal{E}_0 + \mathcal{O}\left(\frac{\rho^n}{n}\right) \tag{1.2.16}$$

Asymptotically we can write the bias term as

$$\begin{aligned} \lim_{n\to\infty} n^2 \mathrm{Tr}\left[\mathbf{H}\mathbb{E}\left[\Delta_0\Delta_0^T\right]\right] &= \lim_{n\to\infty} \mathrm{Tr}\left[\frac{1}{\eta^2}\mathbf{H}(\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I})\mathcal{T}^{-1}\mathcal{E}_0\right] \\ &= \frac{1}{\eta^2}\Delta_0^T\mathbf{H}\Delta_0 \end{aligned} \tag{1.2.17}$$

1.3 Variance term

Consider stating at the solution itself, i.e., $\Delta_0 = \mathbf{0}$, then we are only left with the variance term that depends on the distribution of ε_i's. In that case re-writing (1.1.3),

$$\Delta_n = \frac{\eta}{n}\sum_{k=1}^{n-1}\left(\sum_{j=k}^{n-1}\mathbf{M}_{k,j}\right)\mathbf{x}_k\varepsilon_k \tag{1.3.1}$$

Therefore,

$$\mathbb{E}\left[\Delta_n\Delta_n^T\right] = \frac{\eta^2}{n^2}\mathbb{E}\left[\sum_{k=1}^{n-1}\left[\left(\sum_{j=k}^{n-1}\mathbf{M}_{k,j}\right)\mathbf{x}_k\varepsilon_k^2\mathbf{x}_k^T\left(\sum_{p=k}^{n-1}\mathbf{M}_{k,p}^T\right)\right]\right] \tag{1.3.2}$$

For notational simplicity, let us denote $\mathbf{\Sigma}_k = \mathbf{x}_k\varepsilon_k^2\mathbf{x}_k^T$. Then we can re-write (1.3.2) as

$$\begin{aligned} \mathbb{E}\left[\Delta_n\Delta_n^T\right] &= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k}^{n-1}\sum_{p=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{\Sigma}_k\mathbf{M}_{k,p}^T\right] \\ &= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{j>p\geq k}\mathbb{E}\left[\mathbf{M}_{p,j}\mathbf{M}_{k,p}\mathbf{\Sigma}_k\mathbf{M}_{k,p}^T\right] + \sum_{k\leq j<p}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{\Sigma}_k\mathbf{M}_{k,j}^T\mathbf{M}_{p,j}^T\right] + \sum_{l=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,l}\mathbf{\Sigma}_k\mathbf{M}_{k,l}^T\right]\right] \\ &= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{j>p\geq k}(\mathbf{I}-\eta\mathbf{H})^{j-p}\mathbb{E}\left[\mathbf{M}_{k,p}\mathbf{\Sigma}_k\mathbf{M}_{k,p}^T\right] + \sum_{k\leq j<p}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{\Sigma}_k\mathbf{M}_{k,j}^T\right](\mathbf{I}-\eta\mathbf{H})^{p-j} + \sum_{l=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,l}\mathbf{\Sigma}_k\mathbf{M}_{k,l}^T\right]\right] \\ &= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{p=k}^{n-1}\left(\sum_{j=p+1}^{n-1}(\mathbf{I}-\eta\mathbf{H})^{j-p}\right)\mathbb{E}\left[\mathbf{M}_{k,p}\mathbf{\Sigma}_k\mathbf{M}_{k,p}^T\right] + \sum_{j=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{\Sigma}_k\mathbf{M}_{k,j}^T\right]\left(\sum_{p=j+1}^{n-1}(\mathbf{I}-\eta\mathbf{H})^{p-j}\right)\right] \\ &\qquad + \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{l=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,l}\mathbf{\Sigma}_k\mathbf{M}_{k,l}^T\right]\right] \\ &= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{p=k}^{n-1}\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-p}\right](\eta\mathbf{H})^{-1}\mathbb{E}\left[\mathbf{M}_{k,p}\mathbf{\Sigma}_k\mathbf{M}_{k,p}^T\right]\right] \\ &\qquad + \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{j=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,j}\mathbf{\Sigma}_k\mathbf{M}_{k,j}^T\right]\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1} + \sum_{l=k}^{n-1}\mathbb{E}\left[\mathbf{M}_{k,l}\mathbf{\Sigma}_k\mathbf{M}_{k,l}^T\right]\right] \\ &= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{p=k}^{n-1}\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-p}\right](\eta\mathbf{H})^{-1}(\mathbf{I}-\eta\mathcal{T})^{p-k}\mathbf{\Sigma}_0\right] \qquad \left(\text{Where } \mathbf{\Sigma}_0 = \mathbb{E}\left[\varepsilon^2\mathbf{x}\mathbf{x}^T\right]\right) \\ &\qquad + \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\left[\sum_{j=k}^{n-1}(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1} + \sum_{l=k}^{n-1}(\mathbf{I}-\eta\mathcal{T})^{l-k}\mathbf{\Sigma}_0\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k}^{n-1}\left[(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0 + \left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1}(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0\right] \\
&\qquad + \frac{\eta^2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k}^{n-1}\left[(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1}\right] \\
&= \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\sum_{k=1}^{j}\left[(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0 + \left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1}(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0\right] \\
&\qquad + \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\sum_{k=1}^{j}\left[(\mathbf{I}-\eta\mathcal{T})^{j-k}\mathbf{\Sigma}_0\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1}\right] \qquad \text{(Re-arranging summations)} \\
&= \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 + \left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1}\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0\right] \\
&\qquad + \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0\left[(\mathbf{I}-\eta\mathbf{H}) - (\mathbf{I}-\eta\mathbf{H})^{n-j}\right](\eta\mathbf{H})^{-1}\right] \\
&= \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 + \left[(\mathbf{I}-\eta\mathcal{H})_L - (\mathbf{I}-\eta\mathcal{H})_L^{n-j}\right](\eta\mathcal{H}_L)^{-1}\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0\right] \\
&\qquad + \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[(\mathbf{I}-\eta\mathcal{H})_R - (\mathbf{I}-\eta\mathcal{H})_R^{n-j}\right](\eta\mathcal{H}_R)^{-1}\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 \\
&= \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 \\
&+ \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\left[(\mathbf{I}-\eta\mathcal{H})_L - (\mathbf{I}-\eta\mathcal{H})_L^{n-j}\right](\eta\mathcal{H}_L)^{-1} + \left[(\mathbf{I}-\eta\mathcal{H})_R - (\mathbf{I}-\eta\mathcal{H})_R^{n-j}\right](\eta\mathcal{H}_R)^{-1}\right]\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0
\end{aligned}
\tag{1.3.3}
$$

We can now try to analyze separate terms in (1.3.3). Let us define C_n as

$$
C_n = \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[(I-\eta\mathcal{H})_L^{n-j}(\eta\mathcal{H}_L)^{-1} + (I-\eta\mathcal{H})_R^{n-j}(\eta\mathcal{H}_R)^{-1}\right](\mathbf{I}-\eta\mathcal{T})^{j}(\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 \tag{1.3.4}
$$

Using the definition of C_n, (1.3.3) becomes

$$
\begin{aligned}
\mathbb{E}\left[\Delta_n\Delta_n^T\right] &= \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 \\
&+ \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[(\mathbf{I}-\eta\mathcal{H})_L(\eta\mathcal{H}_L)^{-1} + (\mathbf{I}-\eta\mathcal{H})_R(\eta\mathcal{H}_R)^{-1}\right]\left[\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right](\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 + C_n \\
&= \frac{\eta^2}{n^2}\left[\mathbf{I} + (\mathbf{I}-\eta\mathcal{H})_L(\eta\mathcal{H}_L)^{-1} + (\mathbf{I}-\eta\mathcal{H})_R(\eta\mathcal{H}_R)^{-1}\right]\sum_{j=1}^{n-1}\left(\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right)(\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0 + C_n \\
&= \frac{\eta}{n^2}\left[\mathbf{I} + (\eta\mathcal{H}_L)^{-1} - \mathbf{I} + (\eta\mathcal{H}_R)^{-1} - \mathbf{I}\right]\sum_{j=1}^{n-1}\left(\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right)\mathcal{T}^{-1}\mathbf{\Sigma}_0 + C_n \\
&= \frac{1}{n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right]\sum_{j=1}^{n-1}\left(\mathbf{I}-(\mathbf{I}-\eta\mathcal{T})^{j}\right)\mathcal{T}^{-1}\mathbf{\Sigma}_0 + C_n
\end{aligned}
\tag{1.3.5}
$$

Let us define D_n as

$$
D_n = -\frac{1}{n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right]\sum_{j=1}^{n-1}(\mathbf{I}-\eta\mathcal{T})^{j}\mathcal{T}^{-1}\mathbf{\Sigma}_0
$$

$$= -\frac{1}{\eta n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right]\left[(\mathbf{I}-\eta\mathcal{T}) - (\mathbf{I}-\eta\mathcal{T})^n\right]\mathcal{T}^{-2}\mathbf{\Sigma}_0 \tag{1.3.6}$$

Further, lets define E_n as

$$E_n = \frac{1}{\eta n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right](\mathbf{I}-\eta\mathcal{T})^n\mathcal{T}^{-2}\mathbf{\Sigma}_0 \tag{1.3.7}$$

Using (1.3.6) and (1.3.7) in (1.3.5), we get

$$\mathbb{E}\left[\Delta_n\Delta_n^T\right] = \frac{1}{n}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right]\mathcal{T}^{-1}\mathbf{\Sigma}_0 - \frac{1}{\eta n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right](\mathbf{I}-\eta\mathcal{T})\mathcal{T}^{-2}\mathbf{\Sigma}_0 + C_n + E_n \tag{1.3.8}$$

We will now analyze the Frobenius norms of C_n and E_n and show that they decay exponentially.

$$\begin{aligned}
\|C_n\|_F &\le \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left\|\left[(I-\eta\mathcal{H})_L^{n-j}(\eta\mathcal{H}_L)^{-1} + (I-\eta\mathcal{H})_R^{n-j}(\eta\mathcal{H}_R)^{-1}\right](\mathbf{I}-\eta\mathcal{T})^j(\eta\mathcal{T})^{-1}\mathbf{\Sigma}_0\right\|_F \\
&\le \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\|\mathbf{I}-\eta\mathcal{H}_L\|_{\text{op}}^{n-j}\left\|(\eta\mathcal{H}_L)^{-1}\right\|_{\text{op}}\|\mathbf{I}-\eta\mathcal{T}\|_{\text{op}}^j\left\|(\eta\mathcal{T})^{-1}\right\|_{\text{op}}\|\mathbf{\Sigma}_0\|_F\right] \\
&\quad + \frac{\eta^2}{n^2}\sum_{j=1}^{n-1}\left[\|\mathbf{I}-\eta\mathcal{H}_R\|_{\text{op}}^{n-j}\left\|(\eta\mathcal{H}_R)^{-1}\right\|_{\text{op}}\|\mathbf{I}-\eta\mathcal{T}\|_{\text{op}}^j\left\|(\eta\mathcal{T})^{-1}\right\|_{\text{op}}\|\mathbf{\Sigma}_0\|_F\right] \\
&\le \frac{\eta^2\rho^n}{n^2\eta^2}\|\mathbf{\Sigma}_0\|_F\, n\left[\left\|\mathcal{H}_L^{-1}\right\|_{\text{op}}\left\|\mathcal{T}^{-1}\right\|_{\text{op}} + \left\|\mathcal{H}_R^{-1}\right\|_{\text{op}}\left\|\mathcal{T}^{-1}\right\|_{\text{op}}\right] \\
&\le \frac{\rho^n}{n\mu_T}\|\mathbf{\Sigma}_0\|_F\left[\left\|\mathcal{H}_L^{-1}\right\|_{\text{op}} + \left\|\mathcal{H}_R^{-1}\right\|_{\text{op}}\right] \\
&\le \frac{2d\rho^n}{n\mu\mu_T}\|\mathbf{\Sigma}_0\|_F
\end{aligned} \tag{1.3.9}$$

Similarly,

$$\begin{aligned}
\|E_n\|_F &\le \frac{1}{\eta n^2}\left\|\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right](\mathbf{I}-\eta\mathcal{T})^n\mathcal{T}^{-2}\mathbf{\Sigma}_0\right\|_F \\
&\le \frac{d\rho_T^n}{\eta\mu_T^2 n^2}\left(\frac{2}{\mu} - \eta\right)\|\mathbf{\Sigma}_0\|_F
\end{aligned} \tag{1.3.10}$$

Using (1.3.9) and (1.3.10) we get

$$\begin{aligned}
\|C_n + E_n\|_F &\le \frac{2d\rho^n}{n\mu\mu_T}\|\mathbf{\Sigma}_0\|_F + \frac{d\rho_T^n}{\eta\mu_T^2 n^2}\left(\frac{2}{\mu} - \eta\right)\|\mathbf{\Sigma}_0\|_F \\
&\le \frac{d\rho^n}{n}\|\mathbf{\Sigma}_0\|_F\left[\frac{2}{\mu\mu_T} + \frac{1}{n\eta\mu_T^2}\left(\frac{2}{\mu} - \eta\right)\right]
\end{aligned} \tag{1.3.11}$$

Using (1.3.11) we have

$$\mathbb{E}\left[\Delta_n\Delta_n^T\right] = \frac{1}{n}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right]\mathcal{T}^{-1}\mathbf{\Sigma}_0 - \frac{1}{\eta n^2}\left[\mathcal{H}_L^{-1} + \mathcal{H}_R^{-1} - \eta\mathbf{I}\right](\mathbf{I}-\eta\mathcal{T})\mathcal{T}^{-2}\mathbf{\Sigma}_0 + \mathcal{O}\left(\frac{\rho^n}{n}\right) \tag{1.3.12}$$

Taking limits we get

$$\lim_{\eta\to 0} n\text{Tr}\left[\mathbf{H}\mathbb{E}\left[\Delta_n\Delta_n^T\right]\right] \simeq \mathbb{E}\left[\varepsilon\mathbf{x}^T\mathbf{H}^{-1}\mathbf{x}\right] \tag{1.3.13}$$

Further if ε and $\mathbf{x}$ are independent, and if $\mathbb{E}\left[\varepsilon^2\right] = \sigma^2$, we get

$$\lim_{n\to\infty} n\text{Tr}\left[\mathbf{H}\mathbb{E}\left[\Delta_n\Delta_n^T\right]\right] \simeq d\sigma^2 \tag{1.3.14}$$

which matches the Cramer-Rao bound for such a problem.

References

[1] A. Défossez and F. Bach. Constant Step Size Least-Mean-Square: Bias-Variance Trade-offs and Optimal Sampling Distributions. *ArXiv e-prints*, November 2014.