Zeroth-Order SGD noiseless case on Linear Square Regression (LSR) with Mini-Batching and tail averaging

Sarvadnya Nandkumar Purkar(22B4232)
Angad Singh(22B1211)
Indian Institute of Technology, Bombay

Problem Setup

We consider the standard Linear Square Regression (LSR) problem with expected loss:

$$L(w) = \frac{1}{2}\mathbb{E}\left[(y - \langle w, x\rangle)^2\right],$$

and let w^* be the minimizer of $L(w)$. The corresponding population Hessian is $H = \mathbb{E}[xx^\top]$, and the loss admits the quadratic form:

$$L(w) = L(w^*) + \frac{1}{2}(w - w^*)^\top H(w - w^*).$$

Mini-batch Tail-Averaged SGD

Input: Initial point $w_0 \in \mathbb{R}^d$, step size $\gamma > 0$, mini-batch size b, burn-in iterations s, total number of samples n.
Let $T = \lfloor \frac{n}{b} \rfloor$ be the total number of iterations.

For each iteration $t = 1, 2, \ldots, T$, perform:

1. Sample a mini-batch $\{(x_{ti}, y_{ti})\}_{i=1}^b \sim \mathbb{D}^b$, where $\mathbb{D}$ is the data distribution.
2. Compute the gradient estimate:

$$g_t = \frac{1}{b}\sum_{i=1}^{b} \widehat{\nabla \mathcal{L}}(w_{t-1}; x_{ti}, y_{ti})$$

3. Update the iterate:

$$w_t = w_{t-1} - \gamma g_t$$

Output: The tail-averaged iterate:

$$\bar{w} = \frac{1}{T - s}\sum_{t=s+1}^{T} w_t$$

Optimizer: Zeroth-Order SGD Update Rule

$$w_t = w_{t-1} - \gamma \hat{g}_{t-1}$$

Zeroth-Order Gradient Estimator (Two-Point, Mini-Batch)

$$\hat{g}_{t-1} = \frac{1}{2\delta}\left[L_{B_{t-1}}(w_{t-1} + \delta u_{t-1}) - L_{B_{t-1}}(w_{t-1} - \delta u_{t-1})\right] u_{t-1}$$

Where:

- γ is the learning rate.
- δ is a small positive smoothing radius.
- $u_{t-1} \sim \mathcal{N}(0, I)$
- B_{t-1} is a mini-batch of size b, sampled at iteration $t-1$.
- $L_{B_{t-1}}(w) = \frac{1}{2b}\sum_{i \in B_{t-1}} (y_i - \langle w, x_i \rangle)^2$

Derivation

Let $\eta_t \ = w_t - w^*$. Substitute $w = \eta \ + w^*$ into the update rule:

$$\eta_t \ = \eta_{t-1} \ - \gamma \hat{g}_{t-1}$$

Mini-batch Loss Terms in Gradient Estimator

We approximate the loss on the mini-batch B_{t-1} as:

$$L_{B_{t-1}}(w_{t-1} \pm \delta u_{t-1}) = L_{B_{t-1}}(w^*) + \frac{1}{2}(\eta_{t-1} \ \pm \delta u_{t-1})^\top H_{B_{t-1}}(\eta_{t-1} \pm \delta u_{t-1})$$

where $H_{B_{t-1}} = \frac{1}{b}\sum_{i \in B_{t-1}} x_i x_i^\top$ is the empirical Hessian over the mini-batch.

$$= L_{B_{t-1}}(w^*) + \frac{1}{2}\left[\eta_{t-1}^\top H_{B_{t-1}} \eta_{t-1} \pm 2\delta \eta_{t-1}^\top H_{B_{t-1}} u_{t-1} + \delta^2 u_{t-1}^\top H_{B_{t-1}} u_{t-1}\right]$$

Loss Difference

$$L_{B_{t-1}}(w_{t-1} + \delta u_{t-1}) - L_{B_{t-1}}(w_{t-1} - \delta u_{t-1}) = \delta \cdot 2\eta_{t-1}^\top H_{B_{t-1}} u_{t-1}$$

Substitute into $\hat{g}_{t-1}$

$$\hat{g}_{t-1} = \frac{1}{2\delta} \cdot (2\delta \eta_{t-1}^\top H_{B_{t-1}} u_{t-1}) \cdot u_{t-1} = (\eta_{t-1}^\top H_{B_{t-1}} u_{t-1}) u_{t-1}$$

Outer Product Form

$$\hat{g}_{t-1} = (u_{t-1}u_{t-1}^\top H_{B_{t-1}})\eta_{t-1}$$

Update Rule for n_t

$$\eta_t = \left[I - \gamma u_{t-1}u_{t-1}^\top H_{B_{t-1}}\right]\eta_{t-1}$$

Conditional Expectation

Assuming independence of u_{t-1} and $H_{B_{t-1}}$, and using $\mathbb{E}[u_{t-1}u_{t-1}^\top] = I$, we get:

$$\mathbb{E}[\eta_t|\eta_{t-1}] = (I - \gamma H_{B_{t-1}})\eta_{t-1}$$

Noise Term

$$\text{Noise}_t = \gamma\left(I - u_{t-1}u_{t-1}^\top\right)H_{B_{t-1}}\eta_{t-1}$$

Full Update Equation

$$\eta_t = (I - \gamma H_{B_{t-1}})\eta_{t-1} + \gamma\left(I - u_{t-1}u_{t-1}^\top\right)H_{B_{t-1}}\eta_{t-1}$$

Bias Evolution (Mean Error)

$$\mathbb{E}[\eta_t] = (I - \gamma H_{B_{t-1}})\mathbb{E}[\eta_{t-1}]$$

Provided γ is chosen so that all eigenvalues of $(I - \gamma H_{B_{t-1}})$ lie within the unit circle, we have $\mathbb{E}[\eta_t] \to 0$.

Variance Generation (Fluctuations)

The noise term:

$$\text{Noise}_t = \gamma\left(I - u_{t-1}u_{t-1}^\top\right)H_{B_{t-1}}\eta_{t-1}$$

has zero conditional expectation given η_{t-1}, and variance accumulates based on the empirical Hessian's fluctuations:

$$\text{Cov}(\eta_t) = \mathbb{E}[\eta_t\eta_t^\top] - \mathbb{E}[\eta_t]\mathbb{E}[\eta_t]^\top$$

Now if we take the tail averaging for η_t -

$$
\begin{aligned}
\bar{\eta}_{s+1,N} &= \frac{1}{N}\sum_{t=s+1}^{s+N}\eta_t \\
&= \frac{1}{N}\sum_{t=s+1}^{s+N}\left(\eta_t^{\text{bias}} + \eta_t^{\text{variance}}\right) \\
&= \bar{\eta}_{s+1,N}^{\text{bias}} + \bar{\eta}_{s+1,N}^{\text{variance}}.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\eta_N \otimes \eta_N\right] &= \mathbb{E}\left[\left(\eta_N^{\text{bias}} + \eta_N^{\text{variance}}\right) \otimes \left(\eta_N^{\text{bias}} + \eta_N^{\text{variance}}\right)\right] \\
&\preceq 2 \cdot \left(\mathbb{E}\left[\eta_N^{\text{bias}} \otimes \eta_N^{\text{bias}}\right] + \mathbb{E}\left[\eta_N^{\text{variance}} \otimes \eta_N^{\text{variance}}\right]\right).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[L(w_N)\right] - L(w^*) &= \frac{1}{2}\left\langle \mathbf{H},\, \mathbb{E}\left[\eta_N \otimes \eta_N\right]\right\rangle \\
&\leq \frac{1}{2}\left\langle \mathbf{H},\, 2 \cdot \left(\mathbb{E}\left[\eta_N^{\text{bias}} \otimes \eta_N^{\text{bias}}\right] + \mathbb{E}\left[\eta_N^{\text{variance}} \otimes \eta_N^{\text{variance}}\right]\right)\right\rangle \\
&= 2 \cdot \left(\frac{1}{2}\left\langle \mathbf{H},\, \mathbb{E}\left[\eta_N^{\text{bias}} \otimes \eta_N^{\text{bias}}\right]\right\rangle + \frac{1}{2}\left\langle \mathbf{H},\, \mathbb{E}\left[\eta_N^{\text{variance}} \otimes \eta_N^{\text{variance}}\right]\right\rangle\right) \\
&= 2 \cdot \left(\mathbb{E}\left[L(w_N^{\text{bias}})\right] - L(w^*) + \mathbb{E}\left[L(w_N^{\text{variance}})\right] - L(w^*)\right).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}\right] &\preceq 2 \cdot \left(\mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}^{\text{bias}} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}^{\text{bias}}\right] + \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}^{\text{variance}} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}^{\text{variance}}\right]\right) \\
L(\bar{\mathbf{w}}_{s+1,N}) - L(\mathbf{w}^*) &= \frac{1}{2}\left\langle \mathbf{H},\, \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}\right]\right\rangle \\
&\leq \left\langle \mathbf{H},\, \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}^{\text{bias}} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}^{\text{bias}}\right] + \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}^{\text{variance}} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}^{\text{variance}}\right]\right\rangle \\
\left\langle \mathbf{H},\, \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}\right]\right\rangle &\leq \frac{1}{N^2}\sum_{l=s+1}^{s+N}\left\langle \mathbf{H},\, \mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right](\gamma\mathbf{H})^{-1} + (\gamma\mathbf{H})^{-1}\mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right]\right\rangle \\
&\quad - \frac{1}{N^2}\sum_{l=s+1}^{s+N}\sum_{k=s+N+1}^{\infty}\left\langle \mathbf{H},\, \mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right](\mathbf{I}-\gamma\mathbf{H})^{k-l} + (\mathbf{I}-\gamma\mathbf{H})^{k-l}\mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right]\right\rangle \\
&\leq \frac{1}{N^2}\sum_{l=s+1}^{s+N}\left\langle \mathbf{H},\, \mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right](\gamma\mathbf{H})^{-1} + (\gamma\mathbf{H})^{-1}\mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right]\right\rangle \\
&= \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right]\right).
\end{aligned}
$$

$$
\begin{aligned}
&\left\langle \mathbf{H},\, \mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right](\mathbf{I}-\gamma\mathbf{H})^{k-l} + (\mathbf{I}-\gamma\mathbf{H})^{k-l}\mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right]\right\rangle \\
&= 2\operatorname{Tr}\left[\mathbf{H}(\mathbf{I}-\gamma\mathbf{H})^{k-l}\mathbb{E}\left[\boldsymbol{\eta}_l \otimes \boldsymbol{\eta}_l\right]\right] \geq 0.
\end{aligned}
$$

$$L(\mathbf{w}) - L(\mathbf{w}^*) = \frac{1}{2}\operatorname{Tr}\left(\mathbf{H}\cdot(\boldsymbol{\eta}\otimes\boldsymbol{\eta})\right), \quad \text{with } \boldsymbol{\eta} = \mathbf{w} - \mathbf{w}^*$$

$$\begin{aligned}
\mathbb{E}\left[L(\bar{\mathbf{w}}_{s+1,N})\right] - L(\mathbf{w}^*) &= \frac{1}{2}\cdot\langle \mathbf{H},\, \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}\otimes\bar{\boldsymbol{\eta}}_{s+1,N}\right]\rangle \\
&\leq \frac{1}{\gamma N^2}\sum_{l=s+1}^{s+N}\operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l\otimes\boldsymbol{\eta}_l\right]\right) \\
&\leq \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\left(\operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{bias}}\otimes\boldsymbol{\eta}_l^{\text{bias}}\right]\right) + \operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{variance}}\otimes\boldsymbol{\eta}_l^{\text{variance}}\right]\right)\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[L(\bar{\mathbf{w}}_{s+1,N}^{\text{bias}})\right] - L(\mathbf{w}^*) &\triangleq \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{bias}}\otimes\boldsymbol{\eta}_l^{\text{bias}}\right]\right), \\
\mathbb{E}\left[L(\bar{\mathbf{w}}_{s+1,N}^{\text{variance}})\right] - L(\mathbf{w}^*) &\triangleq \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{variance}}\otimes\boldsymbol{\eta}_l^{\text{variance}}\right]\right)
\end{aligned}$$

Goal

Derive bounds for the quantity

$$\operatorname{tr}\left(\mathbb{E}\left[\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^\top\right]\right),$$

which represents the expected squared Euclidean norm of the bias component of the error vector at step t, i.e., $\mathbb{E}[\|\eta_t^{\text{bias}}\|^2]$.

Bias Recurrence

The bias evolves deterministically:

$$\eta_t^{\text{bias}} = \left(I - \gamma H_{B_{t-1}}\right)\eta_{t-1}^{\text{bias}},$$

where:

- γ is the effective learning rate.
- $H_{B_{t-1}} = \frac{1}{b}\sum_{i\in B_{t-1}} x_i x_i^\top$ is the Hessian, symmetric and positive definite with eigenvalues $\lambda_{\min} > 0$ and $\lambda_{\max}$.
- I is the identity matrix.

Outer Product Expansion

$$\begin{aligned}\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^\top &= (I-\gamma H_{B_{t-1}})\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^\top(I-\gamma H_{B_{t-1}})^\top \\ &= (I-\gamma H_{B_{t-1}})\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^\top(I-\gamma H_{B_{t-1}}) \quad \text{(since } H_{B_{t-1}} \text{ symmetric)}.\end{aligned}$$

Expectation

Assuming deterministic η_0:

$$\mathbb{E}\left[\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^\top\right] = (I-\gamma H_{B_{t-1}})\,\mathbb{E}\left[\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^\top\right]\,(I-\gamma H_{B_{t-1}}).$$

Let $M_{t-1} = \mathbb{E}\left[\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^\top\right]$ and $P = I-\gamma H_{B_{t-1}}$. Then:

$$\mathbb{E}\left[\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^\top\right] = PM_{t-1}P.$$

Trace Bound

$$\begin{aligned}\operatorname{tr}\left(\mathbb{E}[\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^\top]\right) &= \operatorname{tr}(PM_{t-1}P) \\ &= \operatorname{tr}(P^2M_{t-1}) \quad \text{(cyclic property of trace)} \\ &\le \|P^2\|_2 \operatorname{tr}(M_{t-1}) \quad \text{(another property of trace)} \\ &= \|P\|_2^2 \operatorname{tr}(M_{t-1}).\end{aligned}$$

Spectral Norm of P

The eigenvalues of $P = I-\gamma H_{B_{t-1}}$ are $1-\gamma\lambda_i(H_{B_{t-1}})$. Thus,

$$\|P\|_2 = \max_i |1-\gamma\lambda_i(H_{B_{t-1}})|.$$

Define:

$$\rho(\gamma) := \|P\|_2 = \max\left(|1-\gamma\lambda_{\min}|, |1-\gamma\lambda_{\max}|\right).$$

Recursive Bound

$$\begin{aligned}\operatorname{tr}\left(\mathbb{E}[\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^\top]\right) &\le \rho(\gamma)^2 \operatorname{tr}\left(\mathbb{E}[\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^\top]\right) \\ &\le \rho(\gamma)^{2t} \operatorname{tr}(\eta_0\eta_0^\top) \\ &= \rho(\gamma)^{2t}\,\|\eta_0\|^2,\end{aligned}$$

assuming η_0 is deterministic.

Conservative Learning Rate

If $0 < \gamma \leq \frac{1}{\lambda_{\max}}$, then all eigenvalues $1 - \gamma\lambda_i$ lie in $[0, 1]$. So:

$$\rho(\gamma) = 1 - \gamma\lambda_{\min},$$

and the bound becomes:

$$\mathbb{E}[\|\eta_t^{\text{bias}}\|^2] \leq (1 - \gamma\lambda_{\min})^{2t}\|\eta_0\|^2.$$

$$\mathbb{E}\left[L\left(\overline{\mathbf{w}}_{s+1,N}^{\text{bias}}\right)\right] - L(\mathbf{w}^*) = \frac{2}{\gamma N^2}\sum_{t=s+1}^{s+N}\mathbb{E}\left[\|\eta_t\|^2\right] \leq \frac{2}{\gamma N^2}\sum_{t=s+1}^{\infty}\mathbb{E}\left[\|\eta_t\|^2\right]$$

We start with the bound:

$$\mathbb{E}\left[L\left(\overline{\mathbf{w}}_{s+1,N}^{\text{bias}}\right)\right] - L(\mathbf{w}^*) \leq \frac{2}{\gamma N^2}\sum_{t=s+1}^{\infty}\mathbb{E}\left[\|\eta_t\|^2\right]$$

Using the bound:

$$\mathbb{E}[\|\eta_t^{\text{bias}}\|^2] \leq (1 - \gamma\lambda_{\min})^{2t}\|\eta_0\|^2,$$

we substitute:

$$\begin{aligned}
\mathbb{E}\left[L\left(\overline{\mathbf{w}}_{s+1,N}^{\text{bias}}\right)\right] - L(\mathbf{w}^*) &\leq \frac{2}{\gamma N^2}\sum_{t=s+1}^{\infty}(1 - \gamma\lambda_{\min})^{2t}\|\eta_0\|^2 \\
&= \frac{2}{\gamma N^2}\|\eta_0\|^2\sum_{t=s+1}^{\infty}(1 - \gamma\lambda_{\min})^{2t} \\
&= \frac{2}{\gamma N^2}\|\eta_0\|^2(1 - \gamma\lambda_{\min})^{2(s+1)}\sum_{t=0}^{\infty}(1 - \gamma\lambda_{\min})^{2t} \\
&= \frac{2}{\gamma N^2}(1 - \gamma\lambda_{\min})^{2(s+1)}\|\eta_0\|^2 \cdot \frac{1}{1 - (1 - \gamma\lambda_{\min})^2} \\
&= \frac{2}{\gamma N^2} \cdot \frac{(1 - \gamma\lambda_{\min})^{2(s+1)}}{1 - (1 - 2\gamma\lambda_{\min} + \gamma^2\lambda_{\min}^2)}\|\eta_0\|^2 \\
&= \frac{2}{\gamma N^2} \cdot \frac{(1 - \gamma\lambda_{\min})^{2(s+1)}}{2\gamma\lambda_{\min} - \gamma^2\lambda_{\min}^2}\|\eta_0\|^2 \\
&= \frac{2}{\gamma^2\lambda_{\min}N^2(2 - \gamma\lambda_{\min})}(1 - \gamma\lambda_{\min})^{2(s+1)}\|\eta_0\|^2
\end{aligned}$$

Let $\boldsymbol{\eta}_t^{\text{var}}$ denote the noise term representing the fluctuation around the mean update direction at step t. Based on the decomposition derived from the mini-batch Zeroth-Order (ZO) update rule, this term is given by:

$$\boldsymbol{\eta}_t^{\text{var}} = \gamma(I - \boldsymbol{u}_{t-1}\boldsymbol{u}_{t-1}^\top)H_{B_{t-1}}\boldsymbol{\eta}_{t-1},$$

where:

- γ is the learning rate parameter (which may incorporate dimension d).
- $\boldsymbol{u}_{t-1} \sim \mathcal{N}(0, I)$ is an isotropic random direction vector (standard Gaussian).
- $H_{B_{t-1}} = \frac{1}{b}\sum_{i\in B_{t-1}} \boldsymbol{x}_i\boldsymbol{x}_i^\top$ is the random empirical Hessian calculated from the mini-batch B_{t-1} of size b. Note that $\mathbb{E}_{B_{t-1}}[H_{B_{t-1}}] = H = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$.
- $\boldsymbol{\eta}_{t-1} = \boldsymbol{w}_{t-1} - \boldsymbol{w}^*$ is the total error vector at step $t-1$.

We are interested in bounding the expected squared Euclidean norm of this noise term:

$$\mathbb{E}[\|\boldsymbol{\eta}_t^{\text{var}}\|^2] = \operatorname{tr}\left(\mathbb{E}[\boldsymbol{\eta}_t^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top]\right).$$

The outer expectation $\mathbb{E}$ is taken over all sources of randomness, which include the random direction $\boldsymbol{u}$ and the mini-batch sampling B.

Step 1: Conditional Second Moment

We first compute the expectation conditioned on the state $\boldsymbol{\eta}_{t-1}$ and the specific mini-batch B_{t-1}. The only remaining randomness is $\boldsymbol{u} := \boldsymbol{u}_{t-1}$. Let $H_B := H_{B_{t-1}}$ for simpler notation in this step(for below part only) .

The conditional mean is:

$$\mathbb{E}_{\boldsymbol{u}}[\boldsymbol{\eta}_t^{\text{var}} \mid \boldsymbol{\eta}_{t-1}, B_{t-1}] = \gamma\mathbb{E}_{\boldsymbol{u}}\left[(I - \boldsymbol{u}\boldsymbol{u}^\top)H_B\boldsymbol{\eta}_{t-1}\right] = \gamma(I - \mathbb{E}_{\boldsymbol{u}}[\boldsymbol{u}\boldsymbol{u}^\top])H_B\boldsymbol{\eta}_{t-1} = \gamma(I-I)H_B\boldsymbol{\eta}_{t-1} = 0.$$

Since the noise term is conditionally zero-mean, its conditional second moment equals its conditional covariance:

$$\mathbb{E}_{\boldsymbol{u}}[\boldsymbol{\eta}_t^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top \mid \boldsymbol{\eta}_{t-1}, B_{t-1}] = \operatorname{Cov}_{\boldsymbol{u}}(\boldsymbol{\eta}_t^{\text{var}} \mid \boldsymbol{\eta}_{t-1}, B_{t-1}).$$

Let $M_B = H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B$. Note that M_B is fixed when conditioning on $\boldsymbol{\eta}_{t-1}$ and B_{t-1}.

We compute the conditional covariance:

$$\begin{aligned}
\operatorname{Cov}_{\boldsymbol{u}}(\boldsymbol{\eta}_t^{\text{var}} \mid \boldsymbol{\eta}_{t-1}, B_{t-1}) &= \mathbb{E}_{\boldsymbol{u}}\left[\left(\gamma(I - \boldsymbol{u}\boldsymbol{u}^\top)H_B\boldsymbol{\eta}_{t-1}\right)\left(\gamma(I - \boldsymbol{u}\boldsymbol{u}^\top)H_B\boldsymbol{\eta}_{t-1}\right)^\top\right] \\
&= \gamma^2\mathbb{E}_{\boldsymbol{u}}\left[(I - \boldsymbol{u}\boldsymbol{u}^\top)H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B(I - \boldsymbol{u}\boldsymbol{u}^\top)^\top\right] \\
&= \gamma^2\mathbb{E}_{\boldsymbol{u}}\left[(I - \boldsymbol{u}\boldsymbol{u}^\top)M_B(I - \boldsymbol{u}\boldsymbol{u}^\top)\right] \quad \text{(since } I - \boldsymbol{u}\boldsymbol{u}^\top \text{ and } M_B \text{ are symmetric)} \\
&= \gamma^2\left(M_B + \operatorname{tr}(M_B)I\right) \quad \text{(using Gaussian identity } \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top M\boldsymbol{u}\boldsymbol{u}^\top] = 2M + \operatorname{tr}(M)I) \\
&= \gamma^2\left(H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B + \operatorname{tr}(H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B)I\right) \\
&= \gamma^2\left(H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B + \operatorname{tr}(H_B^2\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)I\right) \quad \text{(using cyclic property of trace).}
\end{aligned}$$

Step 2: Total Expectation

Now, we take the expectation over the remaining randomness in $\boldsymbol{\eta}_{t-1}$ (inherited from previous steps) and the mini-batch sampling B_{t-1}. Let $\mathbb{E}_{\eta,B}$ denote this expectation.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\eta}_t^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top] &= \mathbb{E}_{\eta_{t-1},B_{t-1}}\left[\mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{\eta}_t^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top \mid \boldsymbol{\eta}_{t-1}, B_{t-1}\right]\right] \\
&= \mathbb{E}_{\eta_{t-1},B_{t-1}}\left[\gamma^2\left(H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B + \operatorname{tr}(H_B^2\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)I\right)\right] \\
&= \gamma^2\left(\mathbb{E}[H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B] + \mathbb{E}[\operatorname{tr}(H_B^2\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)]I\right) \quad \text{(linearity of } \mathbb{E}\text{)} \\
&= \gamma^2\left(\mathbb{E}[H_B\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top H_B] + \operatorname{tr}(\mathbb{E}[H_B^2\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top])I\right) \quad \text{(linearity of trace).}
\end{aligned}$$

Step 3: Apply Trace and Bound

We apply the *trace* operator to get the expected squared norm:

$$\begin{aligned}
\mathbb{E}[\|\eta_t^{\text{var}}\|^2] &= \operatorname{tr}(\mathbb{E}[\eta_t^{\text{var}}(\eta_t^{\text{var}})^\top]) \\
&= \operatorname{tr}\left(\gamma^2\left(\mathbb{E}[H_B^\top\eta_{t-1}\eta_{t-1}^\top H_B] + \operatorname{tr}(\mathbb{E}[H_B^2\eta_{t-1}\eta_{t-1}^\top])I\right)\right) \\
&= \gamma^2\left[\operatorname{tr}(\mathbb{E}(H_B\eta_{t-1}\eta_{t-1}^\top H_B)) + \operatorname{tr}(\operatorname{tr}(\mathbb{E}[H_B^2\eta_{t-1}\eta_{t-1}^\top])I)\right] \\
&= \gamma^2\left[\mathbb{E}[\operatorname{tr}(H_B\eta_{t-1}\eta_{t-1}^\top H_B)] + d\cdot\operatorname{tr}(\mathbb{E}[H_B^2\eta_{t-1}\eta_{t-1}^\top])\right] \quad \text{(trace properties)} \\
&= \gamma^2\left[\mathbb{E}[\operatorname{tr}(H_B^2\eta_{t-1}\eta_{t-1}^\top)] + d\cdot\mathbb{E}[\operatorname{tr}(H_B^2\eta_{t-1}\eta_{t-1}^\top)]\right] \quad \text{(cyclic property)} \\
&= \gamma^2(1+d)\,\mathbb{E}[\operatorname{tr}(H_B^2\eta_{t-1}\eta_{t-1}^\top)].
\end{aligned}$$

Now we bound the remaining expectation term. Let $\mathbb{E}_B$ denote expectation only over the mini-batch B.

$$\mathbb{E}[\operatorname{tr}(H_B^2\eta_{t-1}\eta_{t-1}^\top)] = \mathbb{E}_{\eta_{t-1}}\left[\mathbb{E}_B[\operatorname{tr}(H_B^2\eta_{t-1}\eta_{t-1}^\top) \mid \eta_{t-1}]\right] = \mathbb{E}_{\eta_{t-1}}\left[\operatorname{tr}(\mathbb{E}_B[H_B^2 \mid \eta_{t-1}]\eta_{t-1}\eta_{t-1}^\top)\right].$$

We use the identity for the expected squared empirical Hessian:

$$\mathbb{E}_B[H_B^2] = H^2 + \frac{1}{b}V_H, \quad \text{where } V_H = \mathbb{E}[(xx^\top - H)^2] = \mathbb{E}[(xx^\top)^2] - H^2.$$

Note that V_H represents the variance of the Hessian estimator. Substitute this back:

$$\mathbb{E}[\operatorname{tr}(H_B^2\eta_{t-1}\eta_{t-1}^\top)] = \mathbb{E}_{\eta_{t-1}}\left[\operatorname{tr}\left(\left(H^2 + \frac{1}{b}V_H\right)\eta_{t-1}\eta_{t-1}^\top\right)\right].$$

Using the trace inequality $\operatorname{tr}(XY) \le \|X\|_2\operatorname{tr}(Y)$ for symmetric X and PSD $Y = \eta_{t-1}\eta_{t-1}^\top$ (note $\operatorname{tr}(Y) = \|\eta_{t-1}\|^2$):

$$\operatorname{tr}\left(\left(H^2 + \frac{1}{b}V_H\right)\eta_{t-1}\eta_{t-1}^\top\right) \le \left\|H^2 + \frac{1}{b}V_H\right\|_2\cdot\operatorname{tr}(\eta_{t-1}\eta_{t-1}^\top) = \left\|H^2 + \frac{1}{b}V_H\right\|_2\cdot\|\eta_{t-1}\|^2.$$

We commonly bound the spectral norm using the triangle inequality and assuming a bound on the Hessian variance, $\|V_H\|_2 \leq \sigma_H^2$:

$$\left\| H^2 + \frac{1}{b} V_H \right\|_2 \leq \|H^2\|_2 + \frac{1}{b}\|V_H\|_2 \leq \lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b}.$$

Applying this bound:

$$\mathbb{E}[\mathrm{tr}(H_B^2 \eta_{t-1} \eta_{t-1}^\top)] \leq \left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b} \right) \mathbb{E}[\|\eta_{t-1}\|^2].$$

$$\mathbb{E}[\mathrm{tr}(H_B^2 \eta_{t-1} \eta_{t-1}^\top)] \leq \mathbb{E}_{\eta_{t-1}} \left[\left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b} \right) \|\eta_{t-1}\|^2 \right] = \left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b} \right) \mathbb{E}[\|\eta_{t-1}\|^2].$$

Here, σ_H^2 is a constant bounding $\|\mathbb{E}[(xx^\top)^2] - H^2\|_2$.

Combining the results, we obtain the final bound on the expected squared norm of the variance term:

$$\begin{aligned} \mathbb{E}[\|\eta_t^{\mathrm{var}}\|^2] &= \gamma^2 (1+d)\, \mathbb{E}[\mathrm{tr}(H_B^2 \eta_{t-1} \eta_{t-1}^\top)] \\ &\leq \gamma^2 (1+d) \left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b} \right) \mathbb{E}[\|\eta_{t-1}\|^2]. \end{aligned}$$

This bound shows that the variance introduced by the ZO estimation noise at each step:

- Is proportional to the square of the learning rate, γ^2.
- Grows approximately linearly with the dimension d.
- Depends on the maximum curvature of the loss landscape, $\lambda_{\max}(H)^2$.
- Includes a term related to the variance of the mini-batch Hessian estimator, σ_H^2/b,
  which decreases as the mini-batch size b increases.
- Is scaled by the expected squared total error from the previous step, $\mathbb{E}[\|\eta_{t-1}\|^2]$, highlighting its multiplicative nature.

We are interested in bounding the expected optimization error (excess loss) attributed to the accumulated variance, averaged over N iterates starting from step $s{+}1$. Let $\overline{\mathbf{w}}_{s+1,N}$ represent some form of averaged iterate over $\mathbf{w}_{s+1}, \ldots, \mathbf{w}_{s+N}$. Under certain analysis frameworks (e.g., for Polyak-Ruppert averaging or related analyses), the contribution of the variance terms to the excess loss can be related to the sum of the expected squared norms of the noise terms.

Let's assume the relevant quantity to bound is proportional to the sum of the expected squared norms of the noise terms:

$$\text{Error}_{\text{variance}} \propto \sum_{l=s+1}^{s+N} \mathbb{E}\left[\left\|\boldsymbol{\eta}_l^{\text{variance}}\right\|^2\right]$$

We need to bound $\sum_{l=s+1}^{s+N} \mathbb{E}\left[\left\|\boldsymbol{\eta}_l^{\text{variance}}\right\|^2\right]$.

From the previous analysis of the noise term $\boldsymbol{\eta}_l^{\text{var}} = \gamma(\mathbf{I} - \mathbf{u}_{l-1}\mathbf{u}_{l-1}^\top)\mathbf{H}_{B_{l-1}}\boldsymbol{\eta}_{l-1}$ in the mini-batch setting (assuming Gaussian $\mathbf{u}$), we derived the bound:

$$\mathbb{E}\left[\left\|\boldsymbol{\eta}_l^{\text{variance}}\right\|^2\right] \leq \gamma^2(1+d)\left(\lambda_{\max}(\mathbf{H})^2 + \frac{\sigma_H^2}{b}\right)\mathbb{E}\left[\|\boldsymbol{\eta}_{l-1}\|^2\right]$$

where $\boldsymbol{\eta}_{l-1} = \mathbf{w}_{l-1} - \mathbf{w}^*$ is the *total* error vector at step $l-1$, d is the dimension, b is the mini-batch size, $\lambda_{\max}(\mathbf{H})$ is the max eigenvalue of the population Hessian, and σ_H^2 relates to the variance of the Hessian estimator (e.g., $\sigma_H^2 \approx \|\mathbb{E}[(\mathbf{x}\mathbf{x}^\top)^2] - \mathbf{H}^2\|_2$).

Let

$$C_{\text{var}} = \gamma^2(1+d)\left(\lambda_{\max}(\mathbf{H})^2 + \frac{\sigma_H^2}{b}\right)$$

Then:

$$\mathbb{E}\left[\left\|\boldsymbol{\eta}_l^{\text{variance}}\right\|^2\right] \leq C_{\text{var}} \cdot \mathbb{E}\left[\|\boldsymbol{\eta}_{l-1}\|^2\right]$$

Summing over the iterations:

$$\begin{aligned}\sum_{l=s+1}^{s+N} \mathbb{E}\left[\left\|\boldsymbol{\eta}_l^{\text{variance}}\right\|^2\right] &\leq \sum_{l=s+1}^{s+N} C_{\text{var}} \cdot \mathbb{E}\left[\|\boldsymbol{\eta}_{l-1}\|^2\right] \\ &= C_{\text{var}} \sum_{k=s}^{s+N-1} \mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right].\end{aligned}$$

To proceed, we need a bound on the sum of the total expected squared errors $\sum_{k=s}^{s+N-1} \mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right]$.

Assuming the algorithm converges, the total error $\mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right]$ typically decreases (possibly to a non-zero noise floor if the learning rate is constant). A common, simple bound, especially if the error is decreasing or stable around iteration s, is:

$$\sum_{k=s}^{s+N-1} \mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right] \leq N \cdot \sup_{k \geq s} \mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right] \quad \text{or perhaps} \quad \sum_{k=s}^{s+N-1} \mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right] \leq N \cdot \mathbb{E}\left[\|\boldsymbol{\eta}_s\|^2\right] \quad \text{(if error is decreasin}$$

Alternatively, if we know $\mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right]$ converges to some value $\mathcal{E}_\infty$, then for large s, the sum is approximately $N \cdot \mathcal{E}_\infty$.

Using the simpler bound $\sum_{k=s}^{s+N-1} \mathbb{E}\left[\|\boldsymbol{\eta}_k\|^2\right] \leq N \cdot \mathbb{E}\left[\|\boldsymbol{\eta}_s\|^2\right]$ (assuming error doesn't significantly increase after step s):

$$\sum_{l=s+1}^{s+N} \mathbb{E}\left[\left\|\boldsymbol{\eta}_l^{\text{variance}}\right\|^2\right] \leq C_{\text{var}} \cdot N \cdot \mathbb{E}\left[\left\|\boldsymbol{\eta}_s\right\|^2\right].$$

Substituting the definition of C_{var}:

$$\sum_{l=s+1}^{s+N} \mathbb{E}\left[\|\eta_l^{\text{variance}}\|^2\right] \leq N\gamma^2(1+d)\left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b}\right)\mathbb{E}\left[\|\eta_s\|^2\right].$$

If we plug this into the specific error expression provided in the original template (assuming its correctness for the analysis context):

$$\begin{aligned}
\mathbb{E}\left[L(\mathbb{W}_{s+1,N}^{\text{variance}})\right] - L(w^*) &\triangleq \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\mathbb{E}\left[\|\eta_l^{\text{variance}}\|^2\right] \\
&\leq \frac{2}{\gamma N^2}\left[N\gamma^2(1+d)\left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b}\right)\mathbb{E}\left[\|\eta_s\|^2\right]\right] \\
&= \frac{2\gamma(1+d)}{N}\left(\lambda_{\max}(H)^2 + \frac{\sigma_H^2}{b}\right)\mathbb{E}\left[\|\eta_s\|^2\right].
\end{aligned}$$

This bound highlights that the average error contribution from variance over N steps:

- Decreases as $\frac{1}{N}$ (the averaging effect)
- Increases linearly with the learning rate γ
- Increases with dimension d
- Depends on the Hessian properties $(\lambda_{\max}(H), \sigma_H^2)$ and mini-batch size b
- Depends on the magnitude of the total error $\mathbb{E}\left[\|\eta_s\|^2\right]$ at the beginning of the averaging window

This is our final expression

$$\begin{aligned}
&\mathbb{E}\left[L(\overline{w}_{s+1,N})\right] - L(w^*) \\
&\leq \frac{2\|\eta_0\|^2}{\gamma^2\lambda_{\min}N^2(2-\gamma\lambda_{\min})}(1-\gamma\lambda_{\min})^{2(s+1)} \\
&\quad + \frac{2(1+d)(\lambda_{\max}^2 + \sigma_H^2/b)}{\lambda_{\min}N(2-\gamma\lambda_{\min})}\|\eta_0\|^2(1-\gamma\lambda_{\min})^{2s} \\
&= \frac{2\|\eta_0\|^2(1-\gamma\lambda_{\min})^{2s}}{\lambda_{\min}N^2(2-\gamma\lambda_{\min})}\left[\frac{(1-\gamma\lambda_{\min})^2}{\gamma^2} + (1+d)N\left(\lambda_{\max}^2 + \frac{\sigma_H^2}{b}\right)\right]
\end{aligned}$$