Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification

Authors :-

- Prateek Jain, Praneeth Netrapalli
- Sham M.Kakade
- Rahul Kidambi
- Aaron Sidford

Angad Singh(22B1211)

PROBLEM SETUP

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

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

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

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

ALGORITHM1:-MINIBATCH-TAIL AVERAGING-SGD

Input: Initial point $\mathbf{w}_0$, stepsize γ, minibatch size b, initial iterations s, total samples n.

1: **for** $t = 1, 2, .., \lfloor \frac{n}{b} \rfloor$ **do**

2: Sample "b" tuples $\{(x_{ti}, y_{ti})\}_{i=1}^{b} \sim \mathcal{D}^b$

3: $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \frac{\gamma}{b} \sum_{i=1}^{b} \widehat{\nabla L_{ti}}(\mathbf{w}_{t-1})$

Output: $\bar{\mathbf{w}} = \frac{1}{\lfloor \frac{n}{b} \rfloor - s} \sum_{i>s} \mathbf{w}_i$

ASSUMPTIONS

($\mathcal{A}1$) **Finite fourth moment:** The fourth moment tensor $\mathcal{M} = \mathbb{E}\left[\mathbf{x}^{\otimes 4}\right]$ exists and is finite.

($\mathcal{A}2$) **Strong convexity:** The Hessian of $L(\cdot)$, $\mathbf{H} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\top}\right]$ is positive definite i.e., $\mathbf{H} \succ 0$.

THEOREM

Theorem 1 *Consider the general mis-specified case of the LSR problem 1. Running Algorithm 1 with a batch size $b \geq 1$, step size $\gamma \leq \gamma_{b,max}/2$, number of unaveraged iterations s, total number of samples n, we obtain an iterate $\overline{\mathbf{w}}$ satisfying the following excess risk bound:*

$$\mathbb{E}\left[L(\overline{\mathbf{w}})\right] - L(\mathbf{w}^*) \leq \frac{2}{\gamma^2\mu^2} \cdot \frac{(1-\gamma\mu)^s}{\left(\frac{n}{b}-s\right)^2} \cdot \left(L(\mathbf{w}_0) - L(\mathbf{w}^*)\right) + 4 \cdot \frac{\widehat{\sigma^2_{MLE}}}{b \cdot (\frac{n}{b}-s)}. \tag{4}$$

In particular, with $\gamma = \gamma_{b,max}/2$, we have the following excess risk bound:

$$L(\overline{\mathbf{w}}) - L(\mathbf{w}^*) \leq \underbrace{\frac{2\kappa_b^2}{\left(\frac{n}{b}-s\right)^2} \exp\left(-\frac{s}{\kappa_b}\right)\left(L(\mathbf{w}_0) - L(\mathbf{w}^*)\right)}_{\mathfrak{T}_1} + \underbrace{4 \cdot \frac{\widehat{\sigma^2_{MLE}}}{b(\frac{n}{b}-s)}}_{\mathfrak{T}_2},$$

with $\kappa_b = \frac{R^2 \cdot \rho_m + (b-1)\|\mathbf{H}\|_2}{b\lambda_{min}(\mathbf{H})}$.

THEOREM(ZEROTH ORDER SGD)

$$
\begin{aligned}
&\mathbb{E}\left[L(\overline{\mathbf{w}}_{s+1,N})\right] - L(\mathbf{w}^*) \\
&\leq \frac{2\left\|\boldsymbol{\eta}_0\right\|^2}{\gamma^2\lambda_{\min}N^2(2-\gamma\lambda_{\min})}(1-\gamma\lambda_{\min})^{2(s+1)} \\
&\quad + \frac{2(1+d)\left(\lambda_{\max}^2+\sigma_H^2/b\right)}{\lambda_{\min}N^2(2-\gamma\lambda_{\min})}\left\|\boldsymbol{\eta}_0\right\|^2(1-\gamma\lambda_{\min})^{2s} \\
&= \frac{2\left\|\boldsymbol{\eta}_0\right\|^2(1-\gamma\lambda_{\min})^{2s}}{\lambda_{\min}N^2(2-\gamma\lambda_{\min})}\left[\frac{(1-\gamma\lambda_{\min})^2}{\gamma^2} + (1+d)\left(\lambda_{\max}^2+\frac{\sigma_H^2}{b}\right)\right]
\end{aligned}
$$

Optimizer: Zeroth-Order SGD Update Rule

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \gamma\, \hat{\boldsymbol{g}}_{t-1}$$

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

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

Where:

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

Derivation

Let $n_t = w_t - w^*$. Substitute $w = n + w^*$ into the update rule:

$$n_t = n_{t-1} - \gamma\, \hat{g}_{t-1}$$

Mini-batch Loss Terms in Gradient Estimator

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

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

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.

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

Loss Difference

$$\begin{aligned} &L_{B_{t-1}}(w_{t-1} + \delta u_{t-1}) - L_{B_{t-1}}(w_{t-1} - \delta u_{t-1}) \\ &= \delta \cdot 2 n_{t-1}^\top H_{B_{t-1}} u_{t-1} \end{aligned}$$

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

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

Outer Product Form

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

Update Rule for n_t

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

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:

$$\mathbb{E}[n_t | n_{t-1}] = (I - \gamma H) n_{t-1}$$

Noise Term

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

Full Update Equation

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

Bias Evolution (Mean Error)

$$\mathbb{E}[\boldsymbol{n}_t] = (\boldsymbol{I} - \gamma \boldsymbol{H})\mathbb{E}[\boldsymbol{n}_{t-1}]$$

Provided γ is chosen so that all eigenvalues of $(\boldsymbol{I} - \gamma \boldsymbol{H})$ lie within the unit circle, we have $\mathbb{E}[\boldsymbol{n}_t] \to 0$.

Variance Generation (Fluctuations)

The noise term:

$$\text{Noise}_t = \gamma(\boldsymbol{I} - \boldsymbol{u}_{t-1}\boldsymbol{u}_{t-1}^\top)\boldsymbol{H}_{B_{t-1}}\boldsymbol{n}_{t-1}$$

has zero conditional expectation given $\boldsymbol{n}_{t-1}$, and variance accumulates based on the empirical Hessian's fluctuations:

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

Now if we take the tail averaging for eta -

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

$$\mathbb{E}\left[\boldsymbol{\eta}_N \otimes \boldsymbol{\eta}_N\right] = \mathbb{E}\left[\left(\boldsymbol{\eta}_N^{\text{bias}} + \boldsymbol{\eta}_N^{\text{variance}}\right) \otimes \left(\boldsymbol{\eta}_N^{\text{bias}} + \boldsymbol{\eta}_N^{\text{variance}}\right)\right]$$

$$\preceq 2 \cdot \left(\mathbb{E}\left[\left(\boldsymbol{\eta}_N^{\text{bias}} \otimes \boldsymbol{\eta}_N^{\text{bias}}\right)\right] + \mathbb{E}\left[\left(\boldsymbol{\eta}_N^{\text{variance}} \otimes \boldsymbol{\eta}_N^{\text{variance}}\right)\right]\right).$$

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

$$\mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}\right] \preceq 2 \cdot \left(\mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}^{\text{bias}} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}^{\text{bias}}\right] + \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}^{\text{variance}} \otimes \bar{\boldsymbol{\eta}}_{s+1,N}^{\text{variance}}\right]\right)$$

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

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

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

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

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

$$\mathbb{E}\left[L(\overline{\mathbf{w}}_{s+1,N})\right] - L(\mathbf{w}^*) = \frac{1}{2}\cdot\langle \mathbf{H}, \mathbb{E}\left[\bar{\boldsymbol{\eta}}_{s+1,N}\otimes\bar{\boldsymbol{\eta}}_{s+1,N}\right]\rangle$$

$$\leq \frac{1}{\gamma N^2}\sum_{l=s+1}^{s+N}\mathrm{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l\otimes\boldsymbol{\eta}_l\right]\right)$$
$$\leq \frac{2}{\gamma N^2}\cdot\sum_{l=s+1}^{s+N}\left(\mathrm{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{bias}}\otimes\boldsymbol{\eta}_l^{\text{bias}}\right]\right) + \mathrm{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{variance}}\otimes\boldsymbol{\eta}_l^{\text{variance}}\right]\right)\right)$$

$$\mathbb{E}\left[L(\overline{\mathbf{w}}^{\text{bias}}_{s+1,N})\right] - L(\mathbf{w}^*) \overset{\text{def}}{=} \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\mathrm{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{bias}}\otimes\boldsymbol{\eta}_l^{\text{bias}}\right]\right).$$

$$\mathbb{E}\left[L(\overline{\mathbf{w}}^{\text{variance}}_{s+1,N})\right] - L(\mathbf{w}^*) \overset{\text{def}}{=} \frac{2}{\gamma N^2}\sum_{l=s+1}^{s+N}\mathrm{Tr}\left(\mathbb{E}\left[\boldsymbol{\eta}_l^{\text{variance}}\otimes\boldsymbol{\eta}_l^{\text{variance}}\right]\right)$$

Goal

Derive bounds for the quantity

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

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:

$$\eta_t^{\text{bias}} = (I - \gamma H)\eta_{t-1}^{\text{bias}},$$

where:

- γ is the effective learning rate.
- $H = \mathbb{E}[xx^T]$ is the Hessian, symmetric and positive definite with eigenvalues $\lambda_{\min} > 0$ and $\lambda_{\max}$.
- I is the identity matrix.

Outer Product Expansion

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

Expectation

Assuming deterministic η_0:

$$\mathbb{E}[\eta_t^{\text{bias}}(\eta_t^{\text{bias}})^T] = (I-\gamma H)\mathbb{E}[\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^T](I-\gamma H).$$

Let $M_{t-1} = \mathbb{E}[\eta_{t-1}^{\text{bias}}(\eta_{t-1}^{\text{bias}})^T]$ and $P = I-\gamma H$.
Then:

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

Trace Bound

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

Spectral Norm of P

The eigenvalues of $P = I - \gamma H$ are $1 - \gamma\lambda_i(H)$. Thus,

$$\|P\|_2 = \max_i |1 - y\lambda_i(H)|.$$

Define:

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

Recursive Bound

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

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:

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

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

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:

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

where:

- γ is the learning rate parameter (which may incorporate dimension d).
- $\boldsymbol{u}_{t-1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is an isotropic random direction vector (standard Gaussian).
- $\boldsymbol{H}_{B_{t-1}} = \frac{1}{b}\sum_{i\in B_{t-1}} \boldsymbol{x}_i\boldsymbol{x}_i^\top$ is the random empirical Hessian calculated from the mini-batch B_{t-1} of size b. Note that $\mathbb{E}_{B_{t-1}}[\boldsymbol{H}_{B_{t-1}}] = \boldsymbol{H} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$.
- $\boldsymbol{\eta}_{t-1} = \boldsymbol{w}_{t-1} - \boldsymbol{w}^*$ 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 η_{t-1} and the specific mini-batch B_{t-1}. The only remaining randomness is $u := u_{t-1}$. Let $H_B := H_{B_{t-1}}$ for simpler notation in this step.

The conditional mean is:

$$\mathbb{E}_u[\eta_t^{\text{var}} \mid \eta_{t-1}, B_{t-1}] = \gamma\, \mathbb{E}_u\left[(I - uu^\top) H_B \eta_{t-1}\right] = \gamma(I - \mathbb{E}_u[uu^\top]) H_B \eta_{t-1} = \gamma(I - I) H_B \eta_{t-1} = 0.$$

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

$$\mathbb{E}_u[\eta_t^{\text{var}} (\eta_t^{\text{var}})^\top \mid \eta_{t-1}, B_{t-1}] = \text{Cov}_u(\eta_t^{\text{var}} \mid \eta_{t-1}, B_{t-1}).$$

Let $M_B = H_B \eta_{t-1} \eta_{t-1}^\top H_B$. Note that M_B is fixed when conditioning on η_{t-1} and B_{t-1}.
We compute the conditional covariance:

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

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.

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

Step 3: Apply Trace and Bound

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

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

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

$$\begin{aligned}\mathbb{E}[\mathrm{tr}(\boldsymbol{H}_B^2 \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)] &= \mathbb{E}_{\boldsymbol{\eta}_{t-1}}\left[\mathbb{E}_B[\mathrm{tr}(\boldsymbol{H}_B^2 \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top) \mid \boldsymbol{\eta}_{t-1}]\right] \\ &= \mathbb{E}_{\boldsymbol{\eta}_{t-1}}\left[\mathrm{tr}(\mathbb{E}_B[\boldsymbol{H}_B^2 \mid \boldsymbol{\eta}_{t-1}]\, \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)\right].\end{aligned}$$

We use the identity for the expected squared empirical Hessian:

$$\mathbb{E}_B[\boldsymbol{H}_B^2] = \boldsymbol{H}^2 + \frac{1}{b}\boldsymbol{V}_H, \quad \text{where } \boldsymbol{V}_H = \mathbb{E}[(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{H})^2] = \mathbb{E}[(\boldsymbol{x}\boldsymbol{x}^\top)^2] - \boldsymbol{H}^2.$$

Note that $\boldsymbol{V}_H$ represents the variance of the Hessian estimator. Substitute this back:

$$\mathbb{E}[\mathrm{tr}(\boldsymbol{H}_B^2 \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)] = \mathbb{E}_{\boldsymbol{\eta}_{t-1}}\left[\mathrm{tr}\left(\left(\boldsymbol{H}^2 + \frac{1}{b}\boldsymbol{V}_H\right)\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\right)\right].$$

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

$$\mathrm{tr}\left(\left(\boldsymbol{H}^2 + \frac{1}{b}\boldsymbol{V}_H\right)\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\right) \le \left\|\boldsymbol{H}^2 + \frac{1}{b}\boldsymbol{V}_H\right\|_2 \cdot \mathrm{tr}(\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top) = \left\|\boldsymbol{H}^2 + \frac{1}{b}\boldsymbol{V}_H\right\|_2 \cdot \|\boldsymbol{\eta}_{t-1}\|^2.$$

We commonly bound the spectral norm using the triangle inequality and assuming a bound on the Hessian variance, $\|\boldsymbol{V}_H\|_2 \le \sigma_H^2$:

$$\left\|\boldsymbol{H}^2 + \frac{1}{b}\boldsymbol{V}_H\right\|_2 \le \|\boldsymbol{H}^2\|_2 + \frac{1}{b}\|\boldsymbol{V}_H\|_2 \le \lambda_{\max}(\boldsymbol{H})^2 + \frac{\sigma_H^2}{b}.$$

Applying this bound:

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

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

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

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

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, $\lambda_{\max}(\boldsymbol{H})^2$.
- 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, $\mathbb{E}[\|\boldsymbol{\eta}_{t-1}\|^2]$, 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:

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

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

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

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

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

Let $C_{\text{var}} = \gamma^2(1+d)\left(\lambda_{\max}(\boldsymbol{H})^2 + \frac{\sigma_{\mu}^2}{b}\right)$. Then:

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

Summing over the iterations:

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

$$= C_{\text{var}} \sum_{k=s}^{s+N-1} \mathbb{E}[\|\boldsymbol{\eta}_k\|^2].$$

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

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

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

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

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

Substituting the definition of C_{var}:

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

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

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

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

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

First order

$$\mathbb{E}\left[L(\overline{\mathbf{w}})\right] - L(\mathbf{w}^*) \leq \frac{2}{\gamma^2\mu^2} \cdot \frac{(1-\gamma\mu)^s}{\left(\frac{n}{b}-s\right)^2} \cdot \left(L(\mathbf{w}_0) - L(\mathbf{w}^*)\right) + 4 \cdot \frac{\widehat{\sigma^2_{MLE}}}{b \cdot (\frac{n}{b}-s)}.$$

Zeroth order

$$\begin{aligned}
&\mathbb{E}\left[L(\overline{\mathbf{w}}_{s+1,N})\right] - L(\mathbf{w}^*) \\
&\leq \frac{2\left\|\boldsymbol{\eta}_0\right\|^2}{\gamma^2\lambda_{\min}N^2(2-\gamma\lambda_{\min})}(1-\gamma\lambda_{\min})^{2(s+1)} \\
&\quad + \frac{2(1+d)\left(\lambda^2_{\max} + \sigma^2_H/b\right)}{\lambda_{\min}N^2(2-\gamma\lambda_{\min})}\left\|\boldsymbol{\eta}_0\right\|^2(1-\gamma\lambda_{\min})^{2s} \\
&= \frac{2\left\|\boldsymbol{\eta}_0\right\|^2(1-\gamma\lambda_{\min})^{2s}}{\lambda_{\min}N^2(2-\gamma\lambda_{\min})}\left[\frac{(1-\gamma\lambda_{\min})^2}{\gamma^2} + (1+d)\left(\lambda^2_{\max} + \frac{\sigma^2_H}{b}\right)\right]
\end{aligned}$$

THANK YOU