How Predictive Can the Optimization Bounds Be in Deep Learning?

Rustem Islamov
University of Basel

How ~~Predictive~~ Useful Can the Optimization Bounds Be in Deep Learning?

Rustem Islamov
University of Basel

Background

How do we typically use optimization bounds?

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)
- To compare convergence of algorithms (e.g., accelerated vs non-accelerated or convergence with respect to problem constants)

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)
- To compare convergence of algorithms (e.g., accelerated vs non-accelerated or convergence with respect to problem constants)
- To talk about the optimality of algorithms for a certain class of problems (e.g., optimal asymptotic convergence)

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)
- To compare convergence of algorithms (e.g., accelerated vs non-accelerated or convergence with respect to problem constants)
- To talk about the optimality of algorithms for a certain class of problems (e.g., optimal asymptotic convergence)
- Something else…

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)
- To compare convergence of algorithms (e.g., accelerated vs non-accelerated or convergence with respect to problem constants)
- To talk about the optimality of algorithms for a certain class of problems (e.g., optimal asymptotic convergence)
- Something else…

This talk is not about such approaches…

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)
- To compare convergence of algorithms (e.g., accelerated vs non-accelerated or convergence with respect to problem constants)
- To talk about the optimality of algorithms for a certain class of problems (e.g., optimal asymptotic convergence)
- Something else…

This talk is not about such approaches…
because usually deep learning problems are highly non-convex
and typically are not covered by the convergence theory

Background

How do we typically use optimization bounds?

- To claim that algorithm works for a certain class of problems (e.g., smooth and convex or smooth and μ-PL)
- To compare convergence of algorithms (e.g., accelerated vs non-accelerated or convergence with respect to problem constants)
- To talk about the optimality of algorithms for a certain class of problems (e.g., optimal asymptotic convergence)
- Something else…

Can we extract useful information from optimization bounds to improve model training?

The Surprising Agreement Between Convex Optimization Theory and Learning-Rate Scheduling for Large Model Training

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, Francis Bach

International Conference on Machine Learning 2025

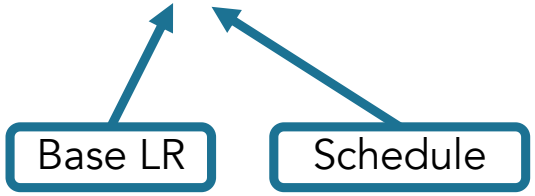
Setup of Schaipp et al

$$\min_{x \in \mathbb{R}^d} \Big[f(x) := \mathbb{E}_s[f(x, s)] \Big]$$

Setup of Schaipp et al

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \mathbb{E}_s[f(x, s)] \right]$$

$$x_{t+1} = x_t - \gamma \eta_t g_t, \quad g_t \in \partial f(x_t, s_t)$$



Setup of Schaipp et al

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \mathbb{E}_s[f(x,s)] \right]$$

$$x_{t+1} = x_t - \gamma \eta_t g_t, \quad g_t \in \partial f(x_t, s_t)$$

Base LR

Schedule

WSD

$$\eta_t = \begin{cases} 1 & 1 \leq t < T_0 \\ 1 - \frac{t - T_0}{T + 1 - T_0} & T_0 \leq t \leq T+1 \end{cases}$$

Cosine Annealing

$$\eta_t = \frac{1}{2}\left(1 + \cos\left(\pi \frac{t-1}{T}\right)\right)$$

Setup of Schaipp et al

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \mathbb{E}_s[f(x,s)] \right]$$

$$\forall\, s, \quad \forall\, x, y \in \mathbb{R}^d, \quad \forall\, g \in \partial f(x,s)$$
$$f(y,s) - f(x,s) \geq \langle g, y - x \rangle$$
$$\mathbb{E}\|g_t\|^2 \leq G^2 \quad \forall\, t \leq T$$

$$x_{t+1} = x_t - \gamma \eta_t g_t, \quad g_t \in \partial f(x_t, s_t)$$

Base LR

Schedule

WSD

$$\eta_t = \begin{cases} 1 & 1 \leq t < T_0 \\ 1 - \frac{t - T_0}{T + 1 - T_0} & T_0 \leq t \leq T + 1 \end{cases}$$

Cosine Annealing

$$\eta_t = \frac{1}{2}\left(1 + \cos\left(\pi \frac{t-1}{T}\right)\right)$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

Defazio et al., 2023

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

$$\mathcal{T}_1(\eta_{1:T}, D, T)$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

$$\mathcal{T}_2(\eta_{1:T}, G, T)$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \underbrace{\frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right]}_{\mathcal{T}_1(\eta_{1:T}, D, T)} + \underbrace{\frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2}_{\mathcal{T}_2(\eta_{1:T}, G, T)}$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \underbrace{\frac{1}{2\gamma \sum_{t=1}^T \eta_t} \left[D^2 \right.}_{\mathcal{T}_1(\eta_{1:T}, D, T)} + \underbrace{\left. \gamma^2 \sum_{t=1}^T \eta_t^2 \mathbb{E}\|g_t\|^2 \right]}_{\mathcal{T}_2(\eta_{1:T}, G, T)} + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^T \eta_t} \cdot \frac{1}{\sum_{t=k}^T \eta_t} \sum_{t=k}^T \eta_t^2 \mathbb{E}\|g_t\|^2$$

The optimal choice of the base LR: $\gamma^\star = \sqrt{\dfrac{\mathcal{T}_1(\eta_{1:T}, D, T)}{\mathcal{T}_2(\eta_{1:T}, G, T)}}$

Optimal Bound: $\mathbb{E}[f(x_T) - f(x^\star)] \leq 2\sqrt{\mathcal{T}_1(\eta_{1:T}, D, T)\mathcal{T}_2(\eta_{1:T}, G, T)}$

Main Convergence Bound

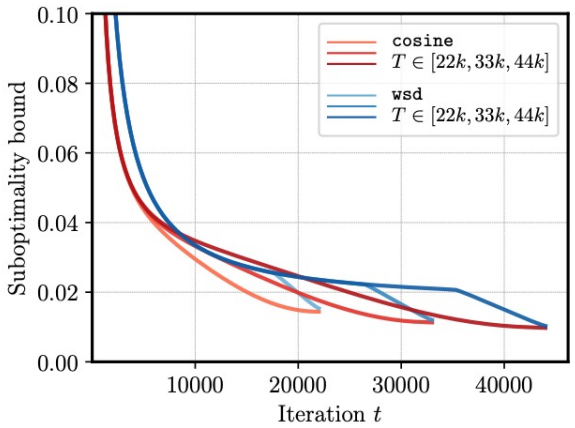
$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \underbrace{\frac{1}{2\gamma \sum_{t=1}^{T} \eta_t}\left[D^2}_{\mathcal{T}_1(\eta_{1:T}, D, T)} + \underbrace{\gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2\right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2}_{\mathcal{T}_2(\eta_{1:T}, G, T)}$$

The optimal choice of the base LR: $\gamma^\star = \sqrt{\dfrac{\mathcal{T}_1(\eta_{1:T}, D, T)}{\mathcal{T}_2(\eta_{1:T}, G, T)}}$ ← Decreases with training horizon

Optimal Bound: $\mathbb{E}[f(x_T) - f(x^\star)] \leq 2\sqrt{\mathcal{T}_1(\eta_{1:T}, D, T)\mathcal{T}_2(\eta_{1:T}, G, T)}$

Main Convergence Bound

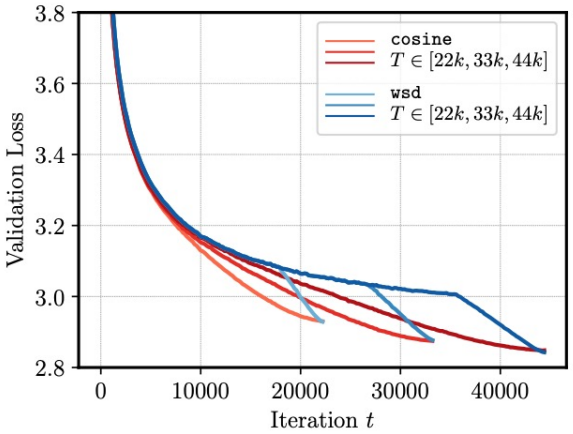
$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$



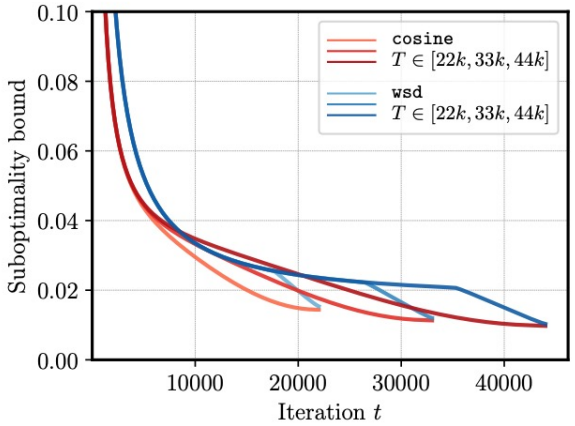
Theoretical Bound

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$



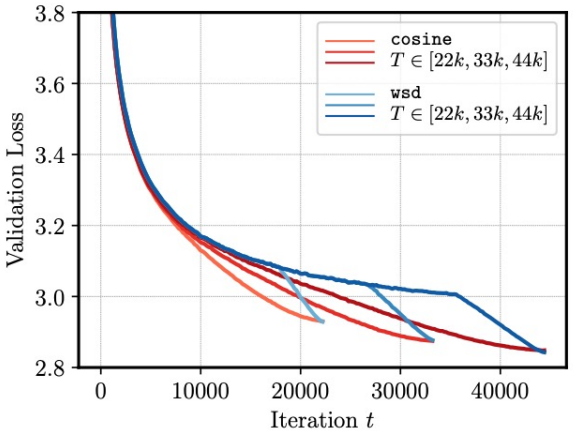
Real Loss Curves
(210M Llama, AdamW)



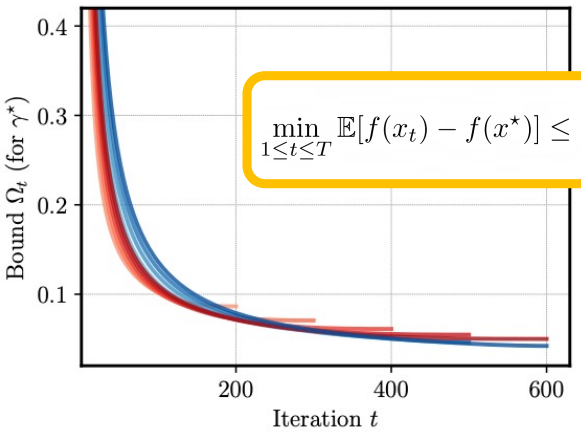
Theoretical Bound

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$



Real Loss Curves
(210M Llama, AdamW)



$$\min_{1 \leq t \leq T} \mathbb{E}[f(x_t) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 G^2 \sum_{t=1}^{T} \eta_t^2 \right]$$

Theoretical Bound

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

For constant schedule:

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathcal{O}\left(\frac{1 + \ln(T)}{\sqrt{T}} \right)$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

For constant schedule: $\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathcal{O}\left(\frac{1 + \ln(T)}{\sqrt{T}}\right)$

For WSD and cosine schedules: $\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

For constant schedule: $\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathcal{O}\left(\frac{1 + \ln(T)}{\sqrt{T}}\right)$

For WSD and cosine schedules: $\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$

Other observations study the cooldown phase (e.g., if γ is tuned, then linear decay is optimal)

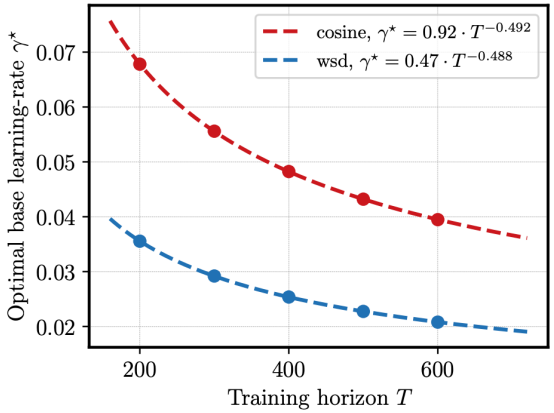
Applications

Setting: we have trained a model for T_1 steps, but later want to continue training up to T_2 steps. How to choose learning rate without re-tuning, if WSD schedule is used?

Applications

Setting: we have trained a model for T_1 steps, but later want to continue training up to T_2 steps. How to choose learning rate without re-tuning, if WSD schedule is used?

Solution: adjust the learning rate schedule between T_1 and T_2 as follows:

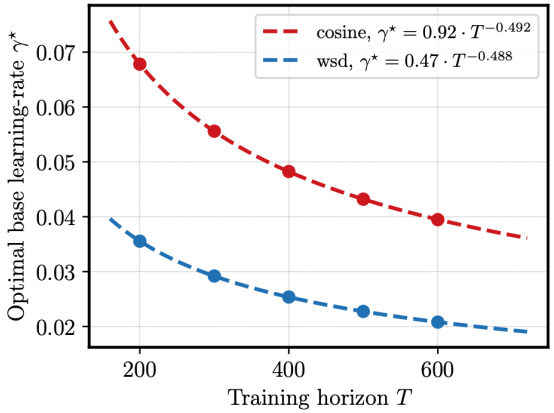


Applications

Setting: we have trained a model for T_1 steps, but later want to continue training up to T_2 steps. How to choose learning rate without re-tuning, if WSD schedule is used?

Solution: adjust the learning rate schedule between T_1 and T_2 as follows:

$$\eta_t = \begin{cases} 1 & 1 \le t \le T_1 \\ \rho & T_1 \le t < cT_2 \\ 1 - \frac{t - cT_2}{T_2 + 1 - cT_2} & cT_2 \le t \le T_2 + 1 \end{cases}$$

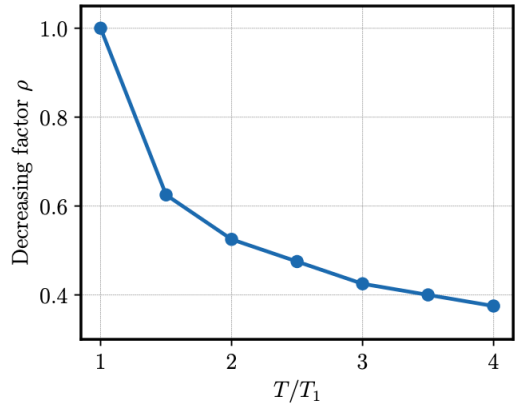


Applications

Setting: we have trained a model for T_1 steps, but later want to continue training up to T_2 steps. How to choose learning rate without re-tuning, if WSD schedule is used?

Solution: adjust the learning rate schedule between T_1 and T_2 as follows:

$$\eta_t = \begin{cases} 1 & 1 \le t \le T_1 \\ \rho & T_1 \le t < cT_2 \\ 1 - \frac{t - cT_2}{T_2 + 1 - cT_2} & cT_2 \le t \le T_2 + 1 \end{cases}$$



Applications

Setting: we have trained a model for T_1 steps, but later want to continue training up to T_2 steps. How to choose learning rate without re-tuning, if WSD schedule is used?

Solution: adjust the learning rate schedule between T_1 and T_2 as follows:

$$\eta_t = \begin{cases} 1 & 1 \le t \le T_1 \\ \rho & T_1 \le t < cT_2 \\ 1 - \frac{t - cT_2}{T_2 + 1 - cT_2} & cT_2 \le t \le T_2 + 1 \end{cases}$$

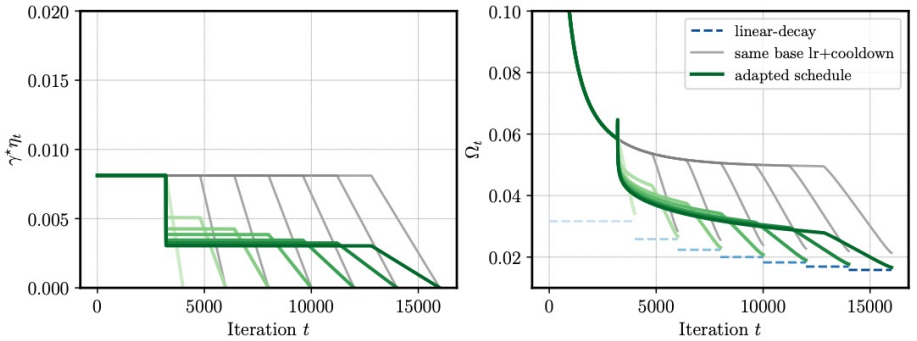


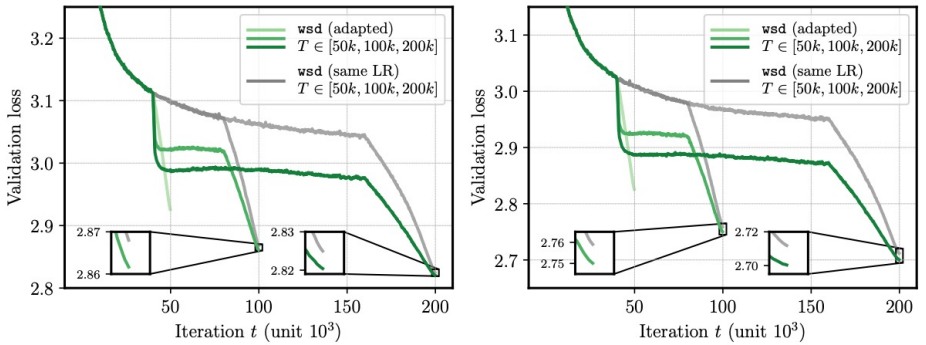
Figure 9. Transfering the learning-rate schedule from horizon $T_1 = 4000$ to $T_2 \in [1.5T_1, 4T_1]$ (see also Fig. 22, left). Decreasing the learning rate **(green)** after the short run (at iteration 3200) leads to significant better bound Ω_t as keeping it constant **(grey)**. Dashed horizontal lines **(blue)** mark bounds for linear-decay schedule with tuned $\gamma^\star$.

Applications

Setting: we have trained a model for T_1 steps, but later want to continue training up to T_2 steps. How to choose learning rate without re-tuning, if WSD schedule is used?

Solution: adjust the learning rate schedule between T_1 and T_2 as follows:

$$\eta_t = \begin{cases} 1 & 1 \le t \le T_1 \\ \rho & T_1 \le t < cT_2 \\ 1 - \frac{t - cT_2}{T_2 + 1 - cT_2} & cT_2 \le t \le T_2 + 1 \end{cases}$$



(a) 124M model (b) 210M model

Figure 10. Transferring the learning-rate schedule from horizon $T_1 = 50\,000$ to $T_2 \in [2T_1, 4T_1]$. Decreasing the base learning-rate **(green)** after 40k steps leads to small improvements in validation loss compared to keeping it the same **(grey)**.

Optimal Linear Decay Learning Rate Schedules and Further Refinements

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, Konstantin Mishchenko
arXiv preprint arXiv: 2310.07831

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathbb{E}\left[\frac{1}{2w_{1:T}}\left(D^2 + \sum_{t=1}^{T} w_t^2 \|g_t\|^2\right)\right]$$

$$x_{t+1} = x_t - \frac{w_t w_{t+1:T}}{w_{1:T}} g_t, \quad g_t \in \partial f(x_t, s_t)$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathbb{E}\left[\frac{1}{2w_{1:T}}\left(D^2 + \sum_{t=1}^{T} w_t^2 \|g_t\|^2\right)\right]$$

$$x_{t+1} = x_t - \frac{w_t w_{t+1:T}}{w_{1:T}} g_t, \quad g_t \in \partial f(x_t, s_t)$$

For a given sequence $\|g_1\|^2,\ldots,\|g_T\|^2$ the RHS of the bound is minimized for

$$w_t = \|g_t\|^{-2} \frac{D}{\sqrt{\sum_{p=1}^{T} \|g_p\|^{-2}}}$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \mathbb{E}\left[\frac{1}{2w_{1:T}}\left(D^2 + \sum_{t=1}^{T} w_t^2 \|g_t\|^2\right)\right]$$

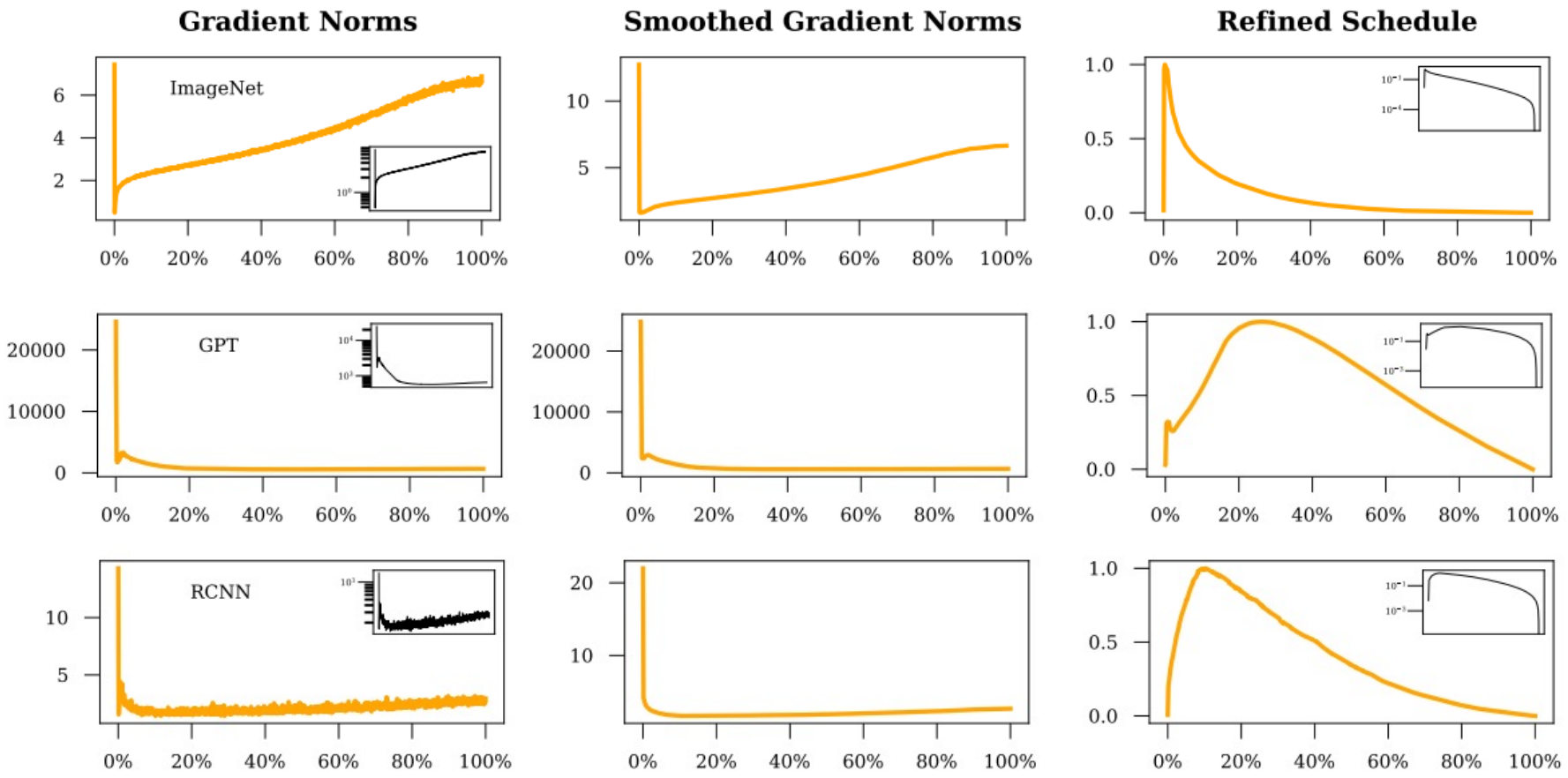
$$x_{t+1} = x_t - \frac{w_t w_{t+1:T}}{w_{1:T}} g_t, \quad g_t \in \partial f(x_t, s_t)$$

For a given sequence $\|g_1\|^2,\ldots,\|g_T\|^2$ the RHS of the bound is minimized for

$$w_t = \|g_t\|^{-2} \frac{D}{\sqrt{\sum_{p=1}^{T} \|g_p\|^{-2}}}$$

Strategy: 1. perform one run using a baseline schedule to record gradient norms
2. use median smoothing filter on the gradient norms
3. perform another run with a refined schedule

Applications



Applications

Problem	Stepwise	Cosine	Linear
CIFAR10	4.53 ±.03	**4.27** ±.04	**4.35** ±.05
CIFAR100	22.78 ±.10	22.59 ±.09	**22.11** ±.08
ImageNet	23.51 ±.07	**23.10** ±.06	**23.11** ±.07
IWSLT14	7.43 ±.01	7.17 ±.009	7.10 ±.01
GPT	19.70 ±.03	18.65 ±.02	18.60 ±.02
RoBERTa	4.36 ±.01	4.07 ±.000	3.94 ±.007
DLRM	20.95 ±.008	**20.94** ±.005	**20.94** ±.006
MRI	8.97 ±.02	8.90 ±.03	8.88 ±.02
ViT	26.27 ±.33	**24.56** ±.15	24.82 ±.31
RCNN	61.76 ±.06	**61.00** ±.04	**60.98** ±.02

Applications

Problem	Stepwise	Cosine	Linear	Refinement
CIFAR10	4.53 ±.03	**4.27** ±.04	**4.35** ±.05	-
CIFAR100	22.78 ±.10	22.59 ±.09	**22.11** ±.08	-
ImageNet	23.51 ±.07	**23.10** ±.06	**23.11** ±.07	**23.12** ±0.03
IWSLT14	7.43 ±.01	7.17 ±.009	7.10 ±.01	**6.92** ±.03
GPT	19.70 ±.03	18.65 ±.02	18.60 ±.02	**18.29** ±.005
RoBERTa	4.36 ±.01	4.07 ±.000	3.94 ±.007	**3.86** ±.005
DLRM	20.95 ±.008	**20.94** ±.005	**20.94** ±.006	**20.94** ±.009
MRI	8.97 ±.02	8.90 ±.03	8.88 ±.02	**8.85** ±.01
ViT	26.27 ±.33	**24.56** ±.15	24.82 ±.31	25.53 ±.16
RCNN	61.76 ±.06	**61.00** ±.04	**60.98** ±.02	61.21 ±.03

Applications

Problem	Stepwise	Cosine	Linear	Refinement
CIFAR10	4.53 ±.03	**4.27** ±.04	**4.35** ±.05	-
CIFAR100	22.78 ±.10	22.59 ±.09	**22.11** ±.08	-
ImageNet	23.51 ±.07	**23.10** ±.06	**23.11** ±.07	**23.12** ±0.03
IWSLT14	7.43 ±.01	7.17 ±.009	7.10 ±.01	**6.92** ±.03
GPT	19.70 ±.03	18.65 ±.02	18.60 ±.02	**18.29** ±.005
RoBERTa	4.36 ±.01	4.07 ±.000	3.94 ±.007	**3.86** ±.005
DLRM	20.95 ±.008	**20.94** ±.005	**20.94** ±.006	**20.94** ±.009
MRI	8.97 ±.02	8.90 ±.03	8.88 ±.02	**8.85** ±.01
ViT	26.27 ±.33	**24.56** ±.15	24.82 ±.31	25.53 ±.16
RCNN	61.76 ±.06	**61.00** ±.04	**60.98** ±.02	61.21 ±.03

Note: for Adam, instead of squared L2 gradient norm, the refinement uses L1 gradient norm

Towards joint scaling laws with optimal batch size schedules

Jiaxiang Li, Zhiqi Bu, Shiyun Xu

arXiv preprint arXiv: 2607.27731

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

Defazio et al., 2023

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

Assumption: $\mathbb{E}\|g_t\|^2 \leq G^2 + \frac{X}{B_t}$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

Assumption: $\mathbb{E}\|g_t\|^2 \leq G^2 + \frac{X}{B_t}$

To simplify the derivations, they move to continuous regime

$$\hat{f}_{\text{dyn}}(T) \approx f(x^\star) + \frac{D^2}{2\gamma \int_0^T \eta_t \mathrm{d}t} + \frac{G^2 \gamma}{2} \int_0^T \left(\frac{\eta_t^2}{\int_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t + \frac{X\gamma}{2} \int_0^T \left(\frac{\eta_t^2 / B_t}{\sum_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t$$

Main Convergence Bound

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq \frac{1}{2\gamma \sum_{t=1}^{T} \eta_t} \left[D^2 + \gamma^2 \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2 \right] + \frac{\gamma}{2} \sum_{k=1}^{T-1} \frac{\eta_k}{\sum_{t=k+1}^{T} \eta_t} \cdot \frac{1}{\sum_{t=k}^{T} \eta_t} \sum_{t=k}^{T} \eta_t^2 \mathbb{E}\|g_t\|^2$$

Assumption: $\mathbb{E}\|g_t\|^2 \leq G^2 + \frac{X}{B_t}$

To simplify the derivations, they move to continuous regime

$$\hat{f}_{\text{dyn}}(T) \approx f(x^\star) + \frac{D^2}{2\gamma \int_0^T \eta_t \mathrm{d}t} + \frac{G^2\gamma}{2} \int_0^T \left(\frac{\eta_t^2}{\int_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t + \frac{X\gamma}{2} \int_0^T \left(\frac{\eta_t^2 / B_t}{\sum_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t$$

$$\text{s.t.} \int_0^T B_t \mathrm{d}t = N$$

Batch Size Scheduling

$$\hat{f}_{\mathrm{dyn}}(T) \approx f(x^\star) + \frac{D^2}{2\gamma \int_0^T \eta_t \mathrm{d}t} + \frac{G^2\gamma}{2} \int_0^T \left(\frac{\eta_t^2}{\int_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t + \frac{X\gamma}{2} \int_0^T \left(\frac{\eta_t^2 / B_t}{\sum_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t \quad \text{s.t.} \quad \int_0^T B_t \mathrm{d}t = N$$

Closed-form solution: $B_t^{\mathrm{optim}} = \frac{N}{2} \frac{\eta_t}{\sqrt{\int_0^T \eta_k \mathrm{d}k \int_t^T \eta_k \mathrm{d}k}}$

Batch Size Scheduling

$$\hat{f}_{\mathrm{dyn}}(T) \approx f(x^\star) + \frac{D^2}{2\gamma \int_0^T \eta_t \mathrm{d}t} + \frac{G^2\gamma}{2} \int_0^T \left(\frac{\eta_t^2}{\int_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t + \frac{X\gamma}{2} \int_0^T \left(\frac{\eta_t^2 / B_t}{\sum_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t \quad \text{s.t.} \quad \int_0^T B_t \mathrm{d}t = N$$

Closed-form solution: $B_t^{\mathrm{optim}} = \frac{N}{2} \frac{\eta_t}{\sqrt{\int_0^T \eta_k \mathrm{d}k \int_t^T \eta_k \mathrm{d}k}}$

Independent of the base LR

Batch Size Scheduling

$$\hat{f}_{\mathrm{dyn}}(T) \approx f(x^\star) + \frac{D^2}{2\gamma \int_0^T \eta_t \mathrm{d}t} + \frac{G^2\gamma}{2} \int_0^T \left(\frac{\eta_t^2}{\int_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t + \frac{X\gamma}{2} \int_0^T \left(\frac{\eta_t^2 / B_t}{\sum_t^T \eta_k \mathrm{d}k} \right) \mathrm{d}t \quad \text{s.t.} \quad \int_0^T B_t \mathrm{d}t = N$$

Closed-form solution: $B_t^{\mathrm{optim}} = \frac{N}{2} \frac{\eta_t}{\sqrt{\int_0^T \eta_k \mathrm{d}k \int_t^T \eta_k \mathrm{d}k}}$

Independent of the base LR

Note: this result departs from the common simplification that the training dynamics are governed by the ratio η_t / B_t. The authors support this claim empirically as well.

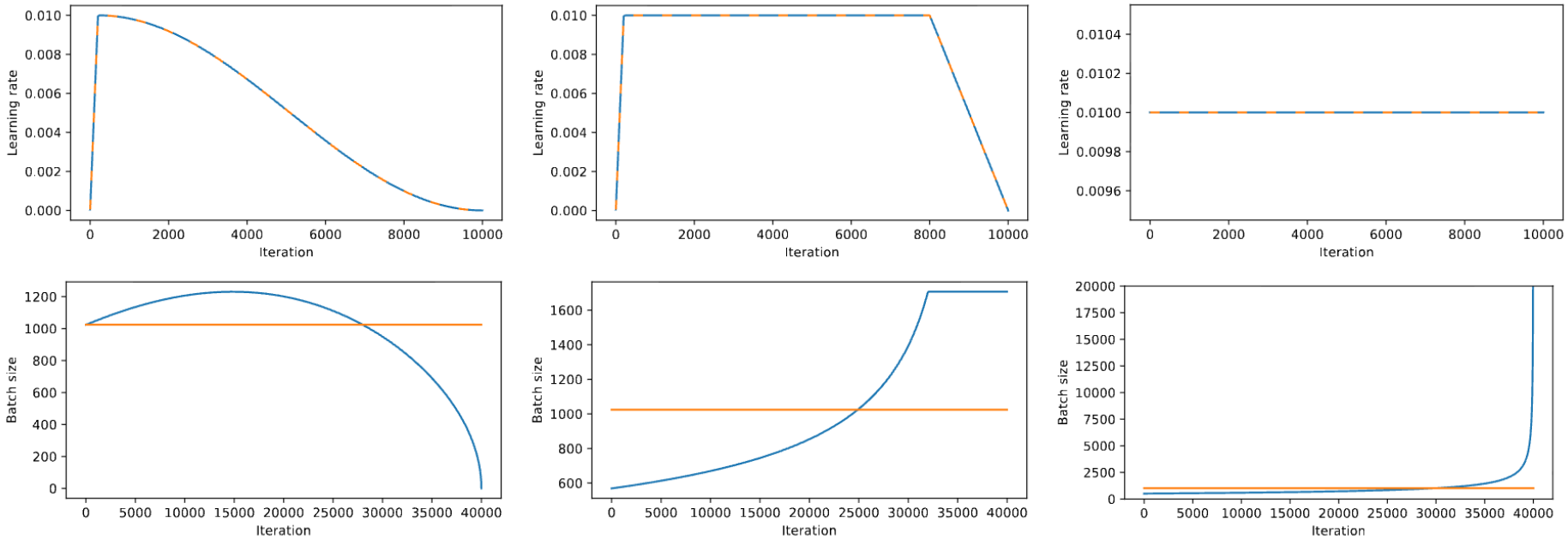
Batch Size Scheduling

$$B_t^{\text{optim}} = \frac{N}{2} \frac{\eta_t}{\sqrt{\int_0^T \eta_k \mathrm{d}k \int_t^T \eta_k \mathrm{d}k}}$$

learning rate	η_t formula	B_t^{optim} formula
constant	η	$\frac{B_{\text{static}}}{2}\sqrt{\frac{T}{T-t}}$
cosine	$\frac{\eta}{2}(\cos(\frac{\pi t}{T})+1)$	$\frac{B_{\text{static}}\cos^2\left(\frac{\pi t}{2T}\right)}{\sqrt{1-\frac{t}{T}-\frac{1}{\pi}\sin\frac{\pi t}{T}}}$
linear	$\eta(1-t/T)$	B_{static}
WSD	$\begin{cases}\eta & \text{if } t \le cT \\ \eta\frac{T-t}{T-cT} & \text{if } t > cT\end{cases}$	$\begin{cases}\frac{B_{\text{static}}T}{\sqrt{(T+cT-2t)(T+cT)}} & \text{if } t \le cT \\ \frac{B_{\text{static}}}{\sqrt{1-c^2}} & \text{if } t > cT\end{cases}$

Batch Size Scheduling

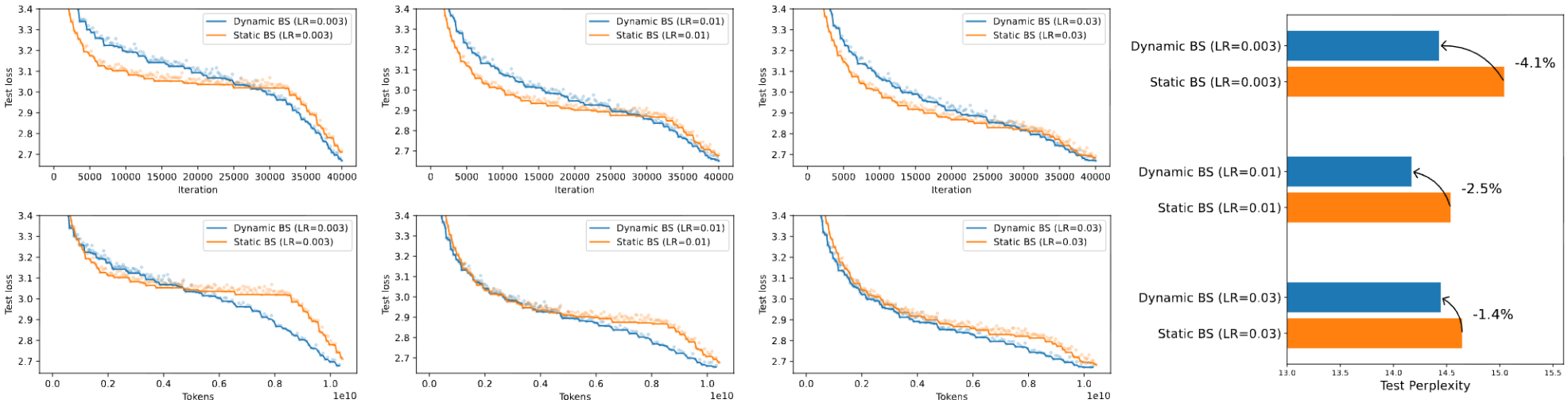
$$B_t^{\text{optim}} = \frac{N}{2} \frac{\eta_t}{\sqrt{\int_0^T \eta_k \mathrm{d}k \int_t^T \eta_k \mathrm{d}k}}$$



Batch Size Scheduling

Let $Z_t = \frac{\eta_t}{\int_t^T \eta_k \mathrm{d}k}$, then

$$\hat{f}_{\mathrm{dyn}}(T) - \hat{f}_{\mathrm{stat}}(T) = \frac{XT}{2B_{\mathrm{stat}}} \left[\left(\frac{\int_0^T Z_t \mathrm{d}t}{T} \right)^2 - \frac{\int_0^T Z_t^2 \mathrm{d}t}{T} \right] \leq 0$$

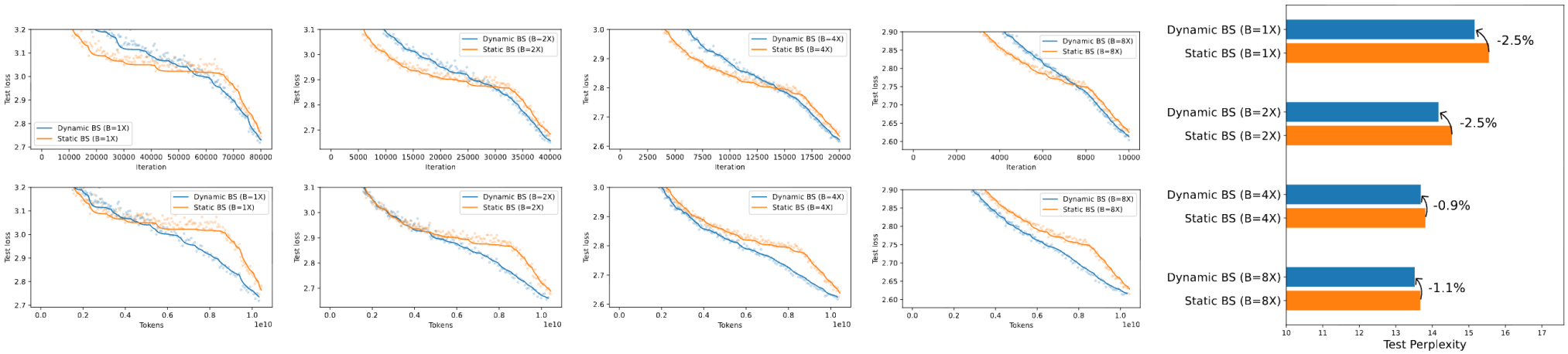


Llama3-1B model on FineWeb-Edu, Batch Size 256

Batch Size Scheduling

Let $Z_t = \frac{\eta_t}{\int_t^T \eta_k \mathrm{d}k}$, then

$$\hat{f}_{\mathrm{dyn}}(T) - \hat{f}_{\mathrm{stat}}(T) = \frac{XT}{2B_{\mathrm{stat}}} \left[\left(\frac{\int_0^T Z_t \mathrm{d}t}{T} \right)^2 - \frac{\int_0^T Z_t^2 \mathrm{d}t}{T} \right] \leq 0$$



Llama3-1B model on FineWeb-Edu, B=1X means 256

Scaling Laws

Using the bounds, the authors provide **scaling laws** – how to scale hyperparameters with token budget N

$$\gamma^\star(N_{\text{large}}, T_{\text{large}}) = \gamma^\star(N_{\text{small}}, T_{\text{small}}) \times (T_{\text{large}}/T_{\text{small}})^{-1/2}$$

$$B_t(T_{\text{large}}) = \frac{N_{\text{large}}}{2} \frac{\eta_t^\star}{\sqrt{\int_0^T \eta_k^\star \mathrm{d}k \int_t^T \eta_k^\star \mathrm{d}k}}$$

Scaling Laws

Using the bounds, the authors provide **scaling laws** – how to scale hyperparameters with token budget N

$$\gamma^\star(N_{\text{large}}, T_{\text{large}}) = \gamma^\star(N_{\text{small}}, T_{\text{small}}) \times (T_{\text{large}}/T_{\text{small}})^{-1/2}$$

$$B_t(T_{\text{large}}) = \frac{N_{\text{large}}}{2} \frac{\eta_t^\star}{\sqrt{\int_0^T \eta_k^\star \mathrm{d}k \int_t^T \eta_k^\star \mathrm{d}k}}$$

But how to set number of iterations?

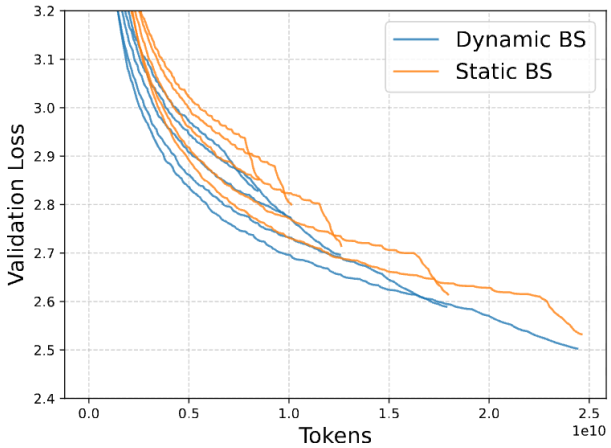
Scaling Laws

Using the bounds, the authors provide **scaling laws** – how to scale hyperparameters with token budget N

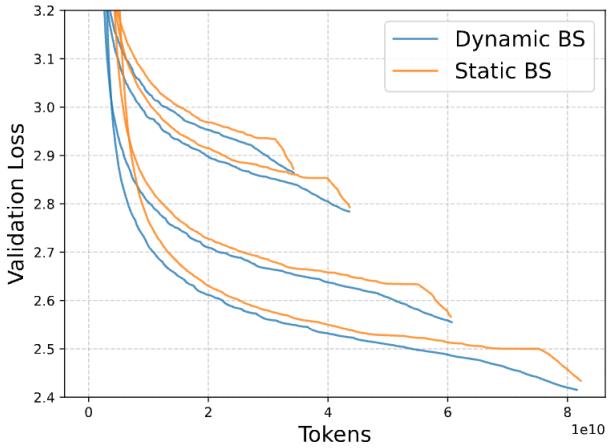
$$\gamma^\star(N_{\text{large}}, T_{\text{large}}) = \gamma^\star(N_{\text{small}}, T_{\text{small}}) \times (T_{\text{large}}/T_{\text{small}})^{-1/2}$$

$$B_t(T_{\text{large}}) = \frac{N_{\text{large}}}{2} \frac{\eta_t^\star}{\sqrt{\int_0^T \eta_k^\star \mathrm{d}k \int_t^T \eta_k^\star \mathrm{d}k}}$$

But how to set number of iterations?



Llama3



Qwen3 MoE

On the Role of Batch Size in Stochastic Conditional Gradient Methods

Rustem Islamov, Roman Macháček, Aurelien Lucchi, Antonio Silveti-Falls,
Eduard Gorbunov, Volkan Cevher

International Conference on Machine Learning 2026

Setup of Islamov et al

Instead of using bounded gradients, the authors propose to study the algorithms under assumptions

$$\forall\ x, y \in \mathcal{X}, \qquad (i)\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \qquad (ii)\mu(f(x) - f^\star) \leq \|\nabla f(x)\|_*$$

$$(iii)\mathbb{E}_s[\|\nabla f(x, s) - \nabla f(x)\|_2^2] \leq \sigma^2/B \qquad (iv)\|x\|_* \leq \rho\|x\|_2$$

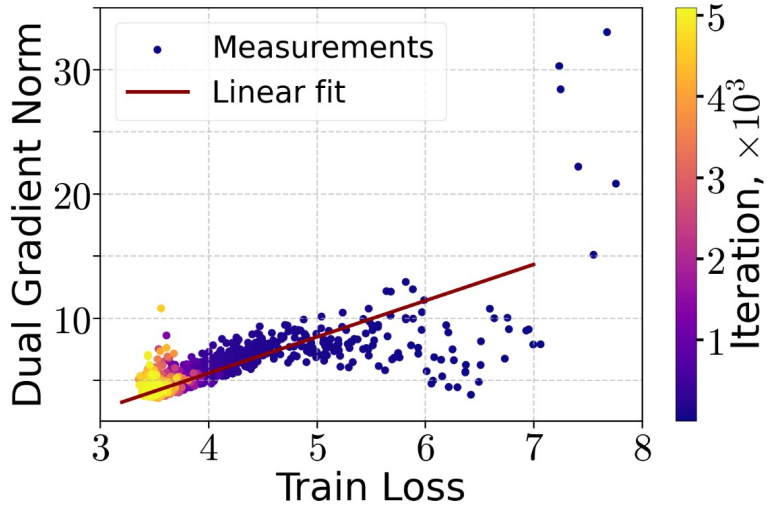
Setup of Islamov et al

Instead of using bounded gradients, the authors propose to study the algorithms under assumptions

$$\forall\ x, y \in \mathcal{X},$$

$$(i)\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \qquad (ii)\mu(f(x) - f^\star) \leq \|\nabla f(x)\|_*$$

$$(iii)\mathbb{E}_s[\|\nabla f(x, s) - \nabla f(x)\|_2^2] \leq \sigma^2/B \qquad (iv)\|x\|_* \leq \rho\|x\|_2$$



NanoGPT 124M

Setup of Islamov et al

Instead of using bounded gradients, the authors propose to study the algorithms under assumptions

$$\forall\ x, y \in \mathcal{X}, \qquad (i)\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \qquad (ii)\mu(f(x) - f^\star) \leq \|\nabla f(x)\|_*$$
$$(iii)\mathbb{E}_s[\|\nabla f(x, s) - \nabla f(x)\|_2^2] \leq \sigma^2/B \qquad (iv)\|x\|_* \leq \rho\|x\|_2$$

Algorithm 1 Stochastic Conditional Gradient (SCG)

Input: $x_0, m_0 \in \mathcal{X}$, parameters $\alpha, \beta \in (0, 1), \eta > 0$
for $t = 0, \ldots, T + 1$ **do**
  sample $\xi_k \sim \mathcal{D}$
  compute $m_{t+1} = (1 - \alpha)m_t + \alpha\nabla f(x_t; s_t)$
  compute $d_{t+1} = \arg\min_{d \in \mathcal{X}} \langle m_{t+1}, d\rangle$ s.t. $\|d\| \leq 1$
  compute $x_{t+1} = (1 - \beta)x_t + \beta\eta d_{t+1}$
end for

Main Convergence Bound

Let $m_0 = \nabla f(x_0, s_0)$, and

$$\beta \cong 1/T,\ \eta \cong 1/\mu,\ \alpha \cong \min\left\{1, \frac{(\varepsilon\mu)^2}{(\rho\sigma)^2}\right\},\ 2\|x_0\| \leq \eta,\ T \cong \max\left\{1, \frac{L}{\varepsilon\mu^2}, \frac{\rho\sigma}{\varepsilon\mu\sqrt{B}}, \frac{L(\rho\sigma)^2}{\mu B(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{B^{3/2}(\varepsilon\mu)^3}\right\}$$

Then, $\mathbb{E}[f(x_T) - f^\star] \leq \varepsilon$.

Based on Kovalev et al., 2025

Main Convergence Bound

Let $m_0 = \nabla f(x_0, s_0)$, and

$\beta \cong 1/T,\ \eta \cong 1/\mu,\ \alpha \cong \min\left\{1, \frac{(\varepsilon\mu)^2}{(\rho\sigma)^2}\right\},\ 2\|x_0\| \leq \eta,\ T \cong \max\left\{1, \frac{L}{\varepsilon\mu^2}, \frac{\rho\sigma}{\varepsilon\mu\sqrt{B}}, \frac{L(\rho\sigma)^2}{\mu B(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{B^{3/2}(\varepsilon\mu)^3}\right\}$

Then, $\mathbb{E}[f(x_T) - f^\star] \leq \varepsilon$.

The token budget is fixed: $N = TB \cong \max\left\{B, \frac{LB}{\mu^2\varepsilon}, \frac{\rho\sigma\sqrt{B}}{\varepsilon\mu}, \frac{L(\rho\sigma)^2}{\mu(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{\sqrt{B}(\varepsilon\mu)^3}\right\}$

Main Convergence Bound

Let $m_0 = \nabla f(x_0, s_0)$, and

$\beta \cong 1/T,\ \eta \cong 1/\mu,\ \alpha \cong \min\left\{1, \frac{(\varepsilon\mu)^2}{(\rho\sigma)^2}\right\},\ 2\|x_0\| \leq \eta,\ T \cong \max\left\{1, \frac{L}{\varepsilon\mu^2}, \frac{\rho\sigma}{\varepsilon\mu\sqrt{B}}, \frac{L(\rho\sigma)^2}{\mu B(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{B^{3/2}(\varepsilon\mu)^3}\right\}$

Then, $\mathbb{E}[f(x_T) - f^\star] \leq \varepsilon$.

The token budget is fixed: $N = TB \cong \max\left\{B, \frac{LB}{\mu^2\varepsilon}, \frac{\rho\sigma\sqrt{B}}{\varepsilon\mu}, \frac{L(\rho\sigma)^2}{\mu(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{\sqrt{B}(\varepsilon\mu)^3}\right\}$

Therefore, the error if lower bounded: $\varepsilon \cong \left\{\frac{\rho\sigma}{\mu B^{1/6}T^{1/3}}, \frac{L^{1/3}(\rho\sigma)^{2/3}}{\mu^{4/3}T^{1/3}}, \frac{LB}{\mu^2 T}\right\}$

Main Convergence Bound

Let $m_0 = \nabla f(x_0, s_0)$, and

$\beta \cong 1/T,\ \eta \cong 1/\mu,\ \alpha \cong \min\left\{1, \frac{(\varepsilon\mu)^2}{(\rho\sigma)^2}\right\},\ 2\|x_0\| \leq \eta,\ T \cong \max\left\{1, \frac{L}{\varepsilon\mu^2}, \frac{\rho\sigma}{\varepsilon\mu\sqrt{B}}, \frac{L(\rho\sigma)^2}{\mu B(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{B^{3/2}(\varepsilon\mu)^3}\right\}$

Then, $\mathbb{E}[f(x_T) - f^\star] \leq \varepsilon$.

The token budget is fixed: $N = TB \cong \max\left\{B, \frac{LB}{\mu^2\varepsilon}, \frac{\rho\sigma\sqrt{B}}{\varepsilon\mu}, \frac{L(\rho\sigma)^2}{\mu(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{\sqrt{B}(\varepsilon\mu)^3}\right\}$

Therefore, the error if lower bounded: $\varepsilon \cong \left\{\frac{\rho\sigma}{\mu B^{1/6}T^{1/3}}, \frac{L^{1/3}(\rho\sigma)^{2/3}}{\mu^{4/3}T^{1/3}}, \frac{LB}{\mu^2 T}\right\}$

U-shaped error floor

Main Convergence Bound

Let $m_0 = \nabla f(x_0, s_0)$, and

$\beta \cong 1/T,\ \eta \cong 1/\mu,\ \alpha \cong \min\left\{1, \dfrac{(\varepsilon\mu)^2}{(\rho\sigma)^2}\right\},\ 2\|x_0\| \leq \eta,\ T \cong \max\left\{1, \dfrac{L}{\varepsilon\mu^2}, \dfrac{\rho\sigma}{\varepsilon\mu\sqrt{B}}, \dfrac{L(\rho\sigma)^2}{\mu B(\varepsilon\mu)^3}, \dfrac{(\rho\sigma)^3}{B^{3/2}(\varepsilon\mu)^3}\right\}$

Then, $\mathbb{E}[f(x_T) - f^\star] \leq \varepsilon$.

Equalizing the last two terms we get: $\dfrac{L^{1/3}(\rho\sigma)^{2/3}}{\mu^{4/3}T^{1/3}} = \dfrac{LB}{\mu^2 T} \Longrightarrow B = T^{2/3} \times \left(\dfrac{\mu\rho\sigma}{L}\right)^{2/3}$

Main Convergence Bound

Let $m_0 = \nabla f(x_0, s_0)$, and

$\beta \cong 1/T,\ \eta \cong 1/\mu,\ \alpha \cong \min\left\{1, \frac{(\varepsilon\mu)^2}{(\rho\sigma)^2}\right\},\ 2\|x_0\| \leq \eta,\ T \cong \max\left\{1, \frac{L}{\varepsilon\mu^2}, \frac{\rho\sigma}{\varepsilon\mu\sqrt{B}}, \frac{L(\rho\sigma)^2}{\mu B(\varepsilon\mu)^3}, \frac{(\rho\sigma)^3}{B^{3/2}(\varepsilon\mu)^3}\right\}$

Then, $\mathbb{E}[f(x_T) - f^\star] \leq \varepsilon$.

Equalizing the last two terms we get: $\frac{L^{1/3}(\rho\sigma)^{2/3}}{\mu^{4/3}T^{1/3}} = \frac{LB}{\mu^2 T} \Longrightarrow B = T^{2/3} \times \left(\frac{\mu\rho\sigma}{L}\right)^{2/3}$

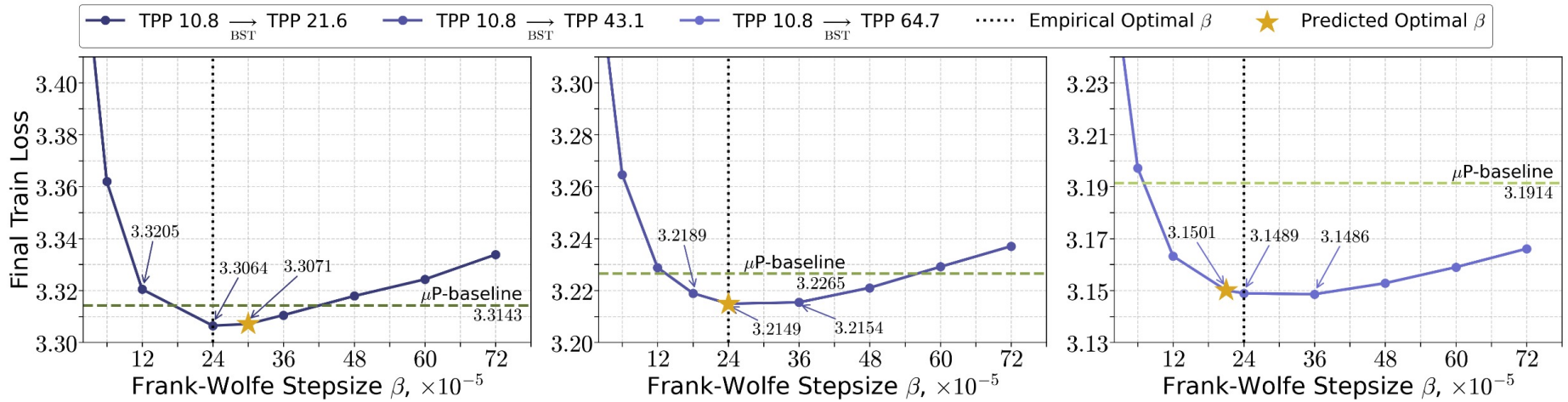
This leads to BST scaling rule:

$$B_{\text{large}} = B_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{2/3} \quad \beta_{\text{large}} = \beta_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{-1/3}$$

$$\alpha_{\text{large}} = \alpha_{\text{small}}$$

Applications

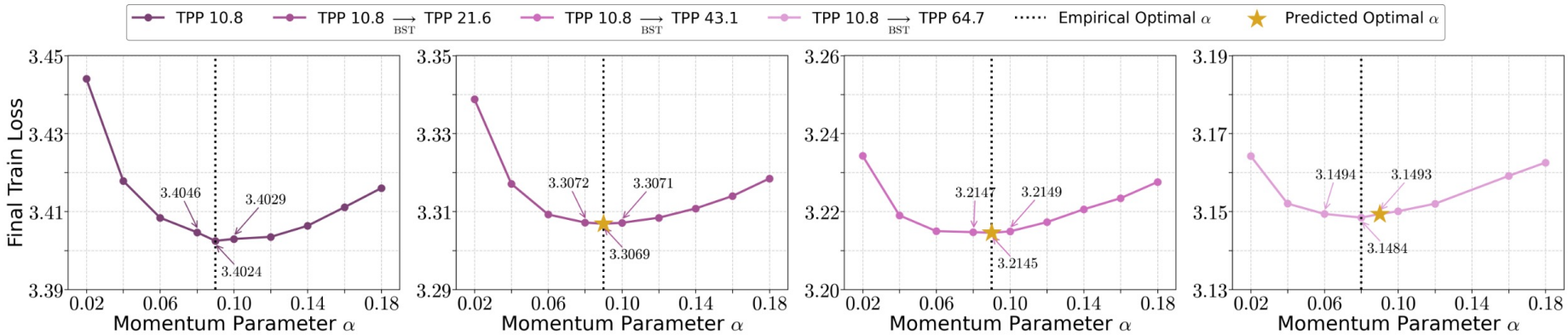
$$B_{\text{large}} = B_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{2/3} \quad \beta_{\text{large}} = \beta_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{-1/3}$$
$$\alpha_{\text{large}} = \alpha_{\text{small}}$$



HP Transfer from NanoGPT 124M

Applications

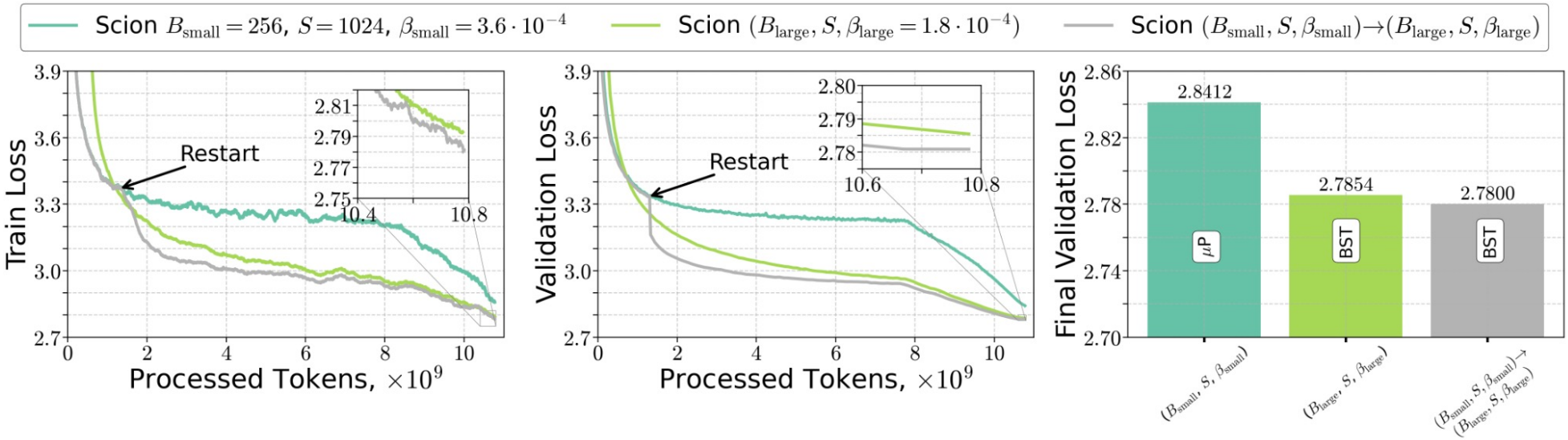
$$B_{\text{large}} = B_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{2/3} \quad \beta_{\text{large}} = \beta_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{-1/3}$$
$$\alpha_{\text{large}} = \alpha_{\text{small}}$$



HP Transfer from NanoGPT 124M

Applications

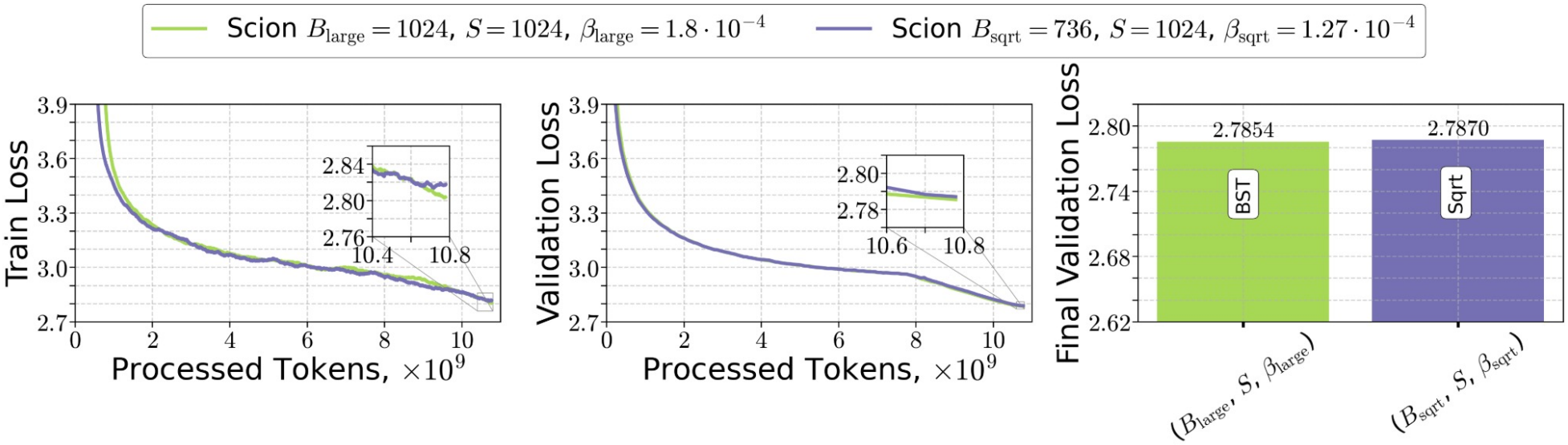
$$B_{\text{large}} = B_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{2/3} \quad \beta_{\text{large}} = \beta_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{-1/3}$$
$$\alpha_{\text{large}} = \alpha_{\text{small}}$$



HP Transfer from NanoGPT 124M to NanoGPT 1B

Applications

$$B_{\text{large}} = B_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{2/3} \quad \beta_{\text{large}} = \beta_{\text{small}} \times (T_{\text{large}}/T_{\text{small}})^{-1/3}$$
$$\alpha_{\text{large}} = \alpha_{\text{small}}$$



HP Transfer from NanoGPT 124M to NanoGPT 1B

Conclusion

- Which set of assumptions is the most predictive?
- How to distinguish between various training setups (over-parametrized vs under-parametrized; one-epoch vs multi-epoch training)
- How model configuration/size should be incorporated into optimization bounds?

Thank you!

Questions? Comments?