

# Convergence of Stochastic Gradient Methods

Chih-Jen Lin  
National Taiwan University

Last updated: August 3, 2025

# Convergence of Stochastic Gradient Methods I

- For simplicity, we do not consider the regularization term
- Therefore,

$$f(\boldsymbol{\theta}) = \frac{1}{l} \sum_{i=1}^l \xi(\mathbf{z}^{L+1,i}(\boldsymbol{\theta}); \mathbf{y}^i, Z^{1,i})$$

- We further define

$$f_i(\boldsymbol{\theta}) = \xi(\mathbf{z}^{L+1,i}(\boldsymbol{\theta}); \mathbf{y}^i, Z^{1,i})$$

# Convergence of Stochastic Gradient Methods II

- Further, we consider the simplest version of stochastic gradient methods: at each step, an index  $\tilde{i}$  is chosen to have

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t)$$

Here  $t$  is the iteration index.

# Convergence of Stochastic Gradient Methods III

- Earlier we wrote

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \left( \frac{\boldsymbol{\theta}}{C} + \frac{1}{|S|} \nabla_{\boldsymbol{\theta}} \sum_{i:i \in S} \xi(\boldsymbol{\theta}; \mathbf{y}^i, Z^{1,i}) \right)$$

but here we need **iteration indices** for the analysis

- Our descriptions here are mainly based on <https://www.cs.cornell.edu/courses/cs4787/2019sp/notes/lecture5.pdf>

# Assumptions I

We assume that

- there is a constant  $G$  such that

$$\|\nabla f_i(\boldsymbol{\theta})\| \leq G, \forall i, \forall \boldsymbol{\theta} \quad (1)$$

- There is a constant  $L > 0$  such that

$$|\boldsymbol{u}^T \nabla^2 f(\boldsymbol{\theta}) \boldsymbol{u}| \leq L \|\boldsymbol{u}\|^2, \forall \boldsymbol{\theta}, \forall \boldsymbol{u} \quad (2)$$

# Compare New and Current Function Values I

- From Taylor's theorem,

$$\begin{aligned} & f(\boldsymbol{\theta}_{t+1}) \\ &= f(\boldsymbol{\theta}_t - \alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)) \\ &= f(\boldsymbol{\theta}_t) - \alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) + \quad (3) \\ & \quad \frac{1}{2} (\alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t))^T \nabla^2 f(\boldsymbol{\xi}_t) (\alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)), \end{aligned}$$

where  $\boldsymbol{\xi}_t$  is between  $\boldsymbol{\theta}_t$  and  $\boldsymbol{\theta}_{t+1}$

# Compare New and Current Function Values II

- From (2) and (1),

(3)

$$\begin{aligned} &\leq f(\boldsymbol{\theta}_t) - \alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) + \frac{\alpha_t^2 L}{2} \|\nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)\|^2 \\ &\leq f(\boldsymbol{\theta}_t) - \alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) + \frac{\alpha_t^2 G^2 L}{2} \end{aligned}$$

- We may not have that

$$-\alpha_t \nabla f_{\tilde{f}_{i_t}}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) < 0$$

# Compare New and Current Function Values III

- Earlier we had

$$-\alpha_t \nabla f(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) < 0$$

- Thus even with small  $\alpha_t$ , the function value may not decrease
- Instead we show the decrease **in expectation**

# Calculating the Expectation I

- The expectation is on the randomness of selecting  $\tilde{f}_t$

$$\begin{aligned} & E[f(\boldsymbol{\theta}_{t+1})] \\ & \leq E\left[f(\boldsymbol{\theta}_t) - \alpha_t \nabla f_{\tilde{f}_t}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) + \frac{\alpha_t^2 G^2 L}{2}\right] \\ & = E[f(\boldsymbol{\theta}_t)] - \alpha_t E[\nabla f_{\tilde{f}_t}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t)] + \frac{\alpha_t^2 G^2 L}{2} \quad (4) \end{aligned}$$

- Note that  $\alpha_t$  depends only on  $t$ , so is a constant in the expectation

# Calculating the Expectation II

- Our expectation is on  $\tilde{i}_t, \forall t$ , so formally

$$\begin{aligned} & E[f(\boldsymbol{\theta}_{t+1})] \\ &= E_{\tilde{i}_0, \dots, \tilde{i}_t}[f(\boldsymbol{\theta}_{t+1})] \end{aligned}$$

Thus on the right-hand side we still need  $E[f(\boldsymbol{\theta}_t)]$  instead of just  $f(\boldsymbol{\theta}_t)$  because

$$\begin{aligned} & E[f(\boldsymbol{\theta}_t)] \\ &= E_{\tilde{i}_0, \dots, \tilde{i}_{t-1}}[f(\boldsymbol{\theta}_t)] \end{aligned}$$

# Calculating the Expectation III

- Next we investigate the term

$$E[\nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t)]$$

by checking the expected value of  $\nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t)$  given  $\boldsymbol{\theta}_t$ :

$$\begin{aligned} & E_{\tilde{i}_t}[\nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t) | \boldsymbol{\theta}_t] \\ &= \sum_{i=1}^I \nabla f_i(\boldsymbol{\theta}_t) P(\tilde{i}_t = i | \boldsymbol{\theta}_t) \\ &= \sum_{i=1}^I \nabla f_i(\boldsymbol{\theta}_t) \cdot \frac{1}{I} = \nabla f(\boldsymbol{\theta}_t) \end{aligned}$$

# Calculating the Expectation IV

- Therefore,

$$\begin{aligned} & E_{\tilde{i}_0, \dots, \tilde{i}_t} [\nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t)] \\ &= E_{\tilde{i}_0, \dots, \tilde{i}_{t-1}} [E_{\tilde{i}_t} [\nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) | \tilde{i}_0, \dots, \tilde{i}_{t-1}]] \end{aligned}$$

This is the same as

$$\begin{aligned} & E_{\tilde{i}_0, \dots, \tilde{i}_{t-1}} [E_{\tilde{i}_t} [\nabla f_{\tilde{i}_t}(\boldsymbol{\theta}_t)^T \nabla f(\boldsymbol{\theta}_t) | \boldsymbol{\theta}_t]] \\ &= E_{\tilde{i}_0, \dots, \tilde{i}_{t-1}} [\|\nabla f(\boldsymbol{\theta}_t)\|^2] \end{aligned}$$

# Calculating the Expectation V

- Therefore,

$$\begin{aligned} & E[f(\boldsymbol{\theta}_{t+1})] \\ & \leq E[f(\boldsymbol{\theta}_t)] - \alpha_t E[\|\nabla f(\boldsymbol{\theta}_t)\|^2] + \frac{\alpha_t^2 G^2 L}{2} \end{aligned}$$

# Calculating the Expectation VI

- We rearrange the terms and sum up over  $T$  iterations

$$\begin{aligned} & \sum_{t=0}^{T-1} \alpha_t E[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\ \leq & \sum_{t=0}^{T-1} \left( E[f(\boldsymbol{\theta}_t)] - E[f(\boldsymbol{\theta}_{t+1})] + \frac{\alpha_t^2 G^2 L}{2} \right) \\ = & E[f(\boldsymbol{\theta}_0)] - E[f(\boldsymbol{\theta}_T)] + \sum_{t=0}^{T-1} \frac{\alpha_t^2 G^2 L}{2} \end{aligned}$$

# Calculating the Expectation VII

This can be further written as

$$\begin{aligned} &= f(\boldsymbol{\theta}_0) - E[f(\boldsymbol{\theta}_T)] + \sum_{t=0}^{T-1} \frac{\alpha_t^2 G^2 L}{2} \\ &\leq f(\boldsymbol{\theta}_0) - f^* + \sum_{t=0}^{T-1} \frac{\alpha_t^2 G^2 L}{2}, \end{aligned}$$

where  $f^*$  is the global optimum of  $f$

- The left-hand side is a sum of all  $T$  iterations
- We need to re-write it in a way of a single iteration