Optimization in Deep Learning

Bingjie YAN

bj.yan.pa@qq.com

ICT, CAS

Powered by @slidev

TOC

1. Optimization Problem & Deep Learning
2. Basic Definition (Convexity, Lipschitz Continuity, Smoothness, etc.)
3. Gradient Decent & Mini-batch Stochastic Gradient Decent
4. Convergence Analysis
5. (Optional) Convergence Analysis for Distributed Machine Learning
6. (Optional) Convergence Analysis for Federated Learning

Note:

1. [illegible]

2. [illegible] FL[illegible] Convergence Analysis [illegible] :)

3. [illegible]

Optimization Problem & Deep Learning

Optimization Problem Definition

□□□□□□□□□□□□□

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

□□ $f(x)$ □□□□□□□x □□□□□□□s.t. □□□□□□□□□

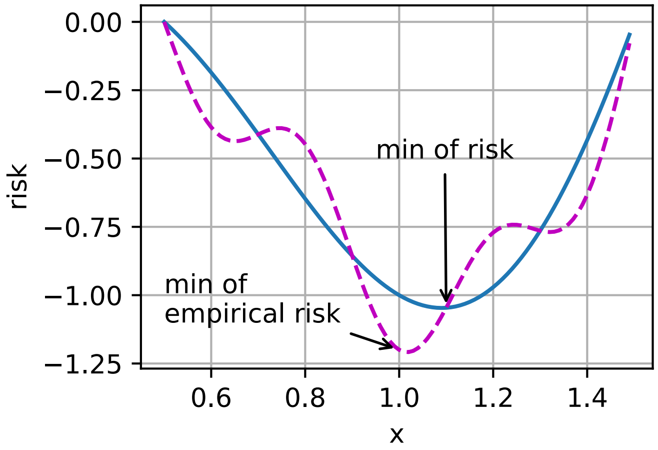
Optimization Problem & Deep Learning

[illegible] -> [illegible]

[illegible]

[illegible]

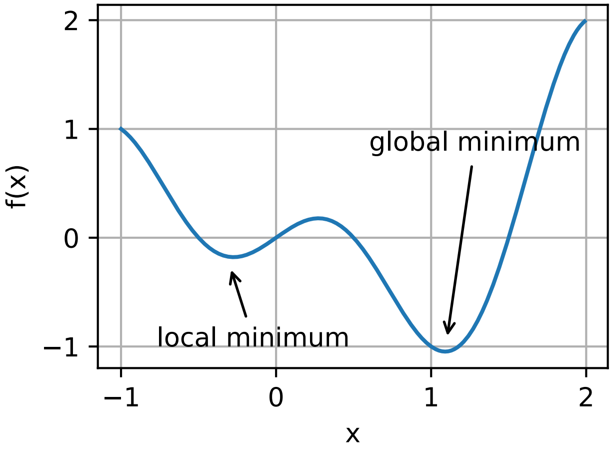
[illegible]😢



Challenges in Deep Learning Optimization

1. [illegible](local minimum)

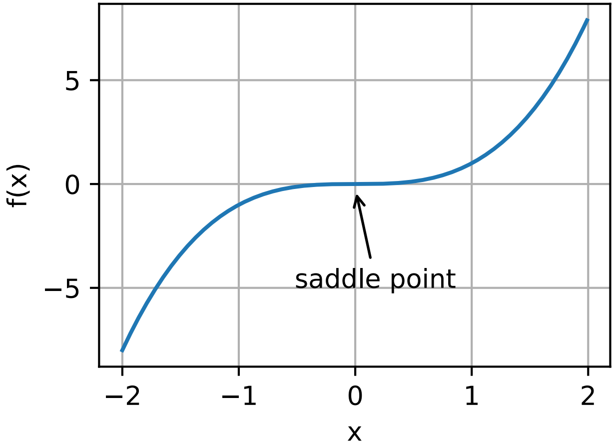
[illegible] $f(x)$[illegible] x [illegible] $f(x)$ [illegible] x [illegible] $f(x)$ [illegible] $f(x)$ [illegible] $f(x)$ [illegible] x [illegible] $f(x)$ [illegible]

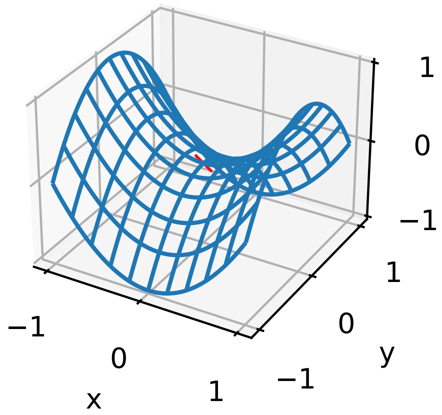


Challenges in Deep Learning Optimization

2. □□(saddle point)

[illegible]





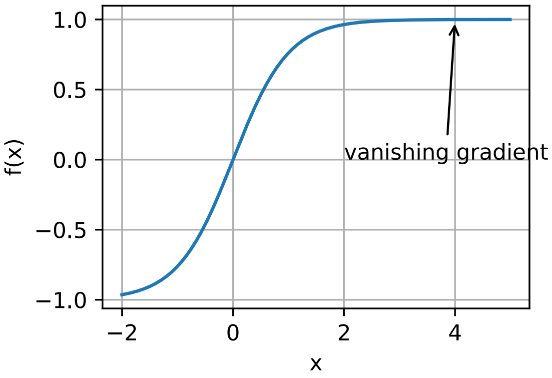
Challenges in Deep Learning Optimization

3. [illegible]

[illegible]

[illegible]🌰[illegible]

[illegible] $f(x) = \tanh(x)$ [illegible] $x = 4$ [illegible] $f'(x) = 1 - \tanh^2(x)$[illegible] $f'(4) = 0.0013$[illegible] $x = 4$ [illegible]



Challenges in Deep Learning Optimization

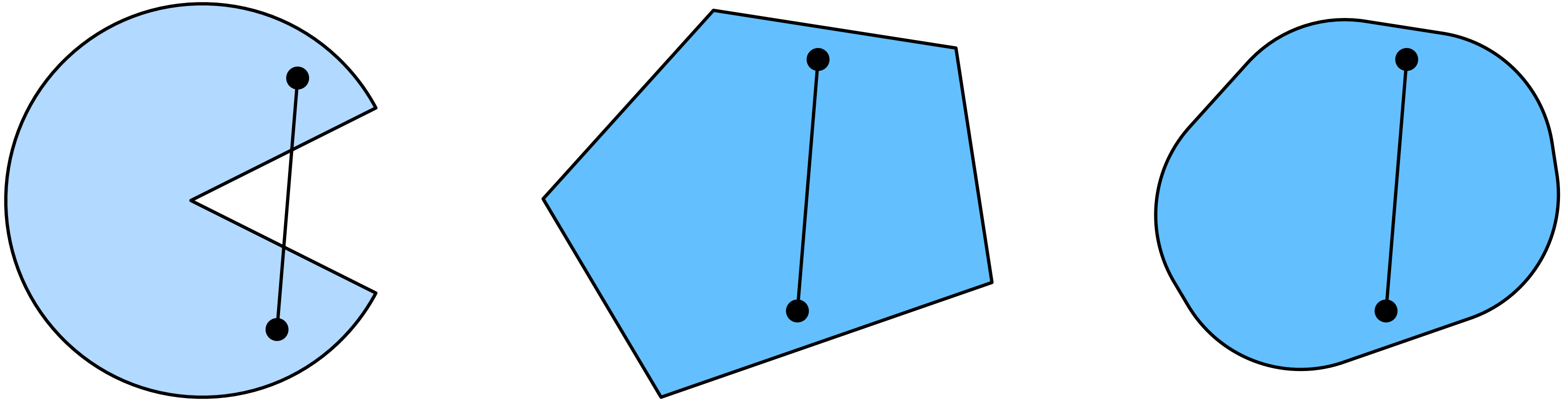
Basic Definition

[illegible]Lipschitz [illegible]

Convex Set

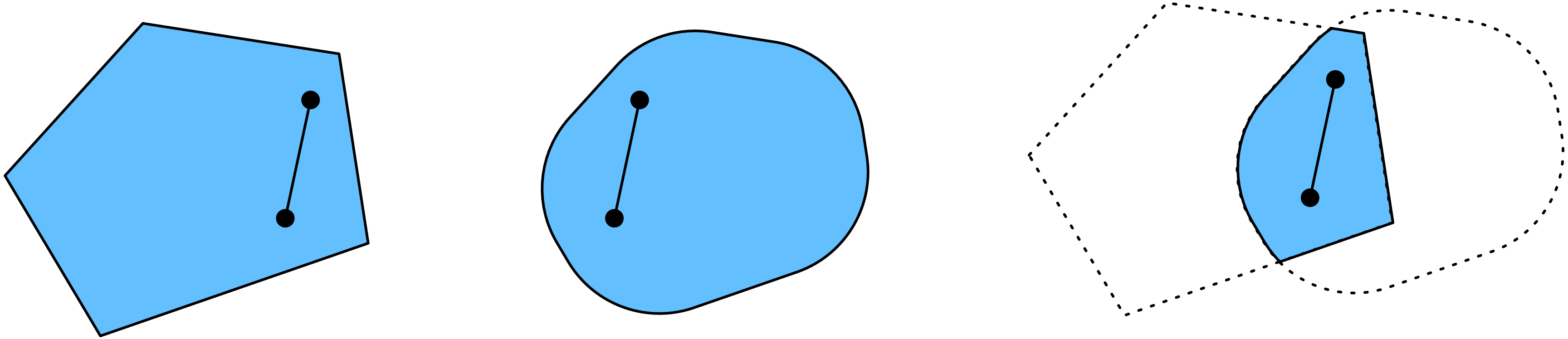
[illegible] $\mathcal{X}$ [illegible] $x, y \in \mathcal{X}$ [illegible] $\alpha \in [0, 1]$[illegible]

$$\alpha x + (1 - \alpha) y \in \mathcal{X}$$



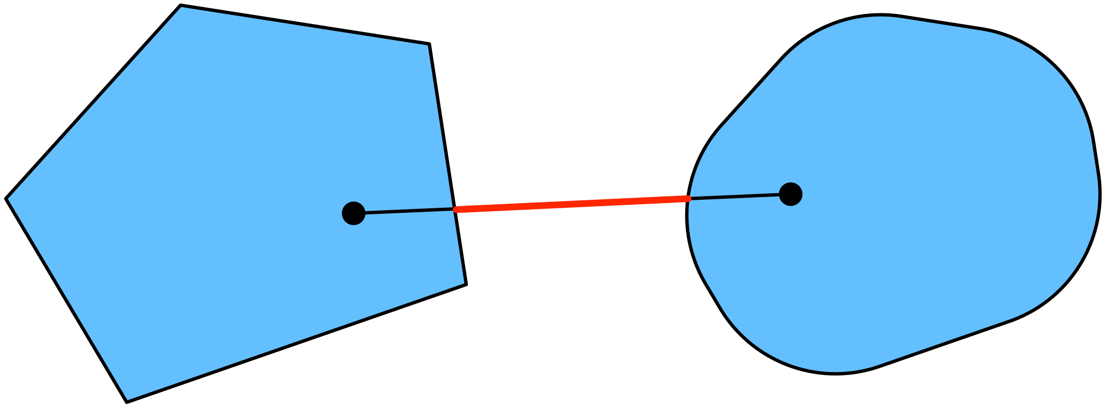
Convex Set

[illegible] $\mathcal{X}$ [illegible] $\mathcal{Y}$[illegible] $\mathcal{X} \cap \mathcal{Y}$ [illegible]



Convex Set

[illegible] $\mathcal{X}$ [illegible] $\mathcal{Y}$ [illegible] $\mathcal{X} \cup \mathcal{Y}$ [illegible]

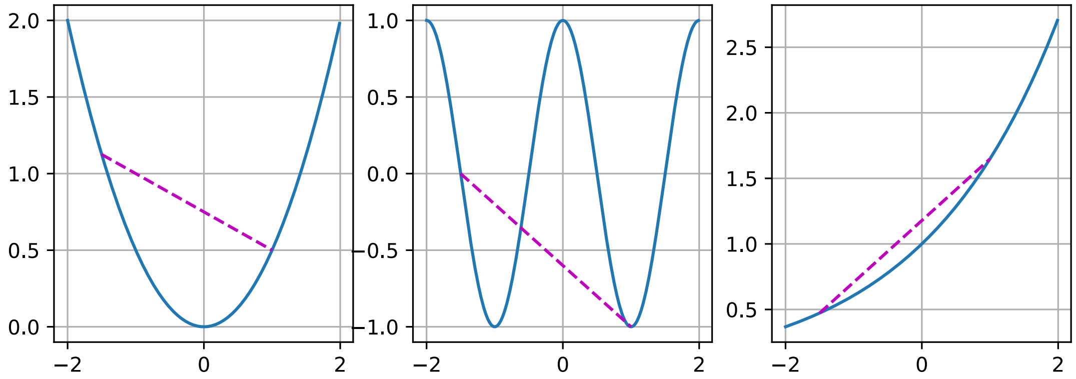


Convex Function

[illegible] $\mathcal{X}$ [illegible] $x, y \in \mathcal{X}$ [illegible] $\alpha \in [0, 1]$[illegible]

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

[illegible] f [illegible]



Properties of Convex Function

1. [illegible]

[illegible] f [illegible] $x^* \in \mathcal{X}$ [illegible] f [illegible] x^* [illegible] f [illegible]

[illegible]

[illegible] x^* [illegible] f [illegible] p[illegible] $x \in \mathcal{X}$ [illegible] $0 < |x - x^*| < p$ [illegible] $f(x) \geq f(x^*)$[illegible]

[illegible] x^* [illegible] f [illegible] $x \in \mathcal{X}$ [illegible] $f(x) < f(x^*)$[illegible] $\alpha = 1 - \frac{p}{|x-x^*|}$ [illegible]

$$
\begin{aligned}
f(\alpha x + (1-\alpha)x^*) &\leq \alpha f(x) + (1-\alpha) f(x^*) \\
&< \alpha f(x^*) + (1-\alpha) f(x^*) \\
&= f(x^*)
\end{aligned}
$$

[illegible] x^* [illegible] f [illegible] x^* [illegible] f [illegible]

Properties of Convex Function

2. [illegible]

3. [illegible]

…

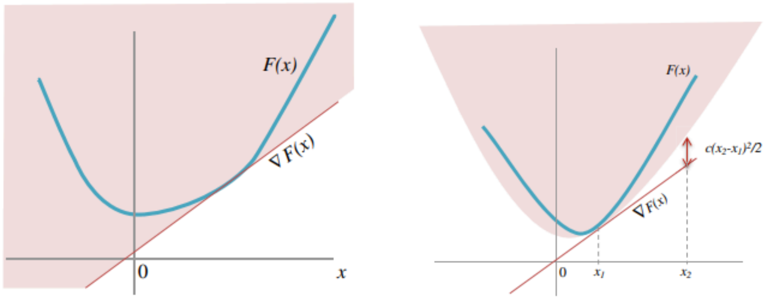
Strong Convex

[illegible] μ-[illegible]
[illegible]μ-[illegible]

[illegible] $f : R^d \to R$[illegible] $||\cdot||$[illegible] $x, y \in R^d$[illegible]

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}||y - x||^2, \quad \forall x, y \in \mathcal{D}$$

[illegible] f [illegible] $||\cdot||$ [illegible] μ-[illegible]



(a) Convex Function (b) Strongly Convex Function

Fig. 2.3 Illustration of a convex function and a c-strongly convex function

Strong Convex

[illegible]μ-[illegible] μ[illegible] μ-[illegible]

[illegible] $y = x^*$ [illegible] $\nabla f(y) = \nabla f(x^*) = 0$[illegible]

$$\langle \nabla f(x), x - x^* \rangle \geq \frac{\mu}{2} ||x - x^*||^2$$

[illegible] f [illegible] μ-[illegible] $f - \frac{\mu}{2}||\cdot||^2$ [illegible]

Lipschitz Continuity

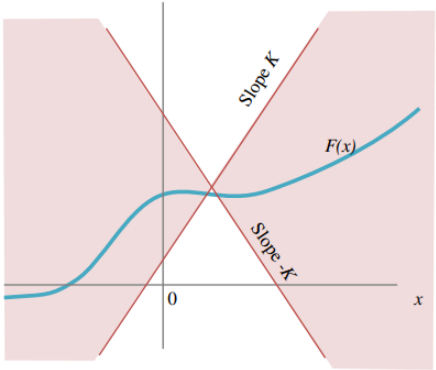
L-Lipschitz ☒☒☒ Lipschitz □□□

☒☒☒☒☒☒ $f : R^d \to R$☒☒☒ $||\cdot||$☒☒☒☒☒☒☒ $L > 0$☒☒☒☒☒☒☒ $x, y \in R^d$□□□□□□□□□

$$|f(x) - f(y)| \leq L||x - y||$$

☒☒☒☒ f ☒☒☒ $||\cdot||$ ☒ L-Lipschitz □□□□

Fig. 2.1 Illustration of Lipschitz continuity. If a scalar function $F(x)$ is K-Lipschitz continuous, then for any point x, the function lies inside the region bounded by lines of slope K and $-K$ that pass through the point $(x, F(x))$



Smoothness

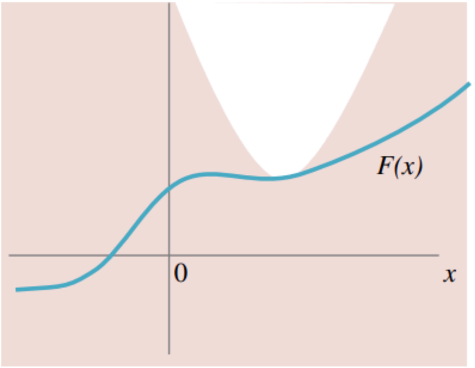
[illegible]

[illegible] $f : R^d \rightarrow R$ [illegible] $||\cdot||$ [illegible] $L > 0$ [illegible] $x, y \in R^d$ [illegible]

$$f(x) - f(y) \leq \nabla f(y)^\top (x - y) + \frac{L}{2}||x - y||^2$$

[illegible] f [illegible] $||\cdot||$ [illegible] L-[illegible]

Fig. 2.2 Illustration of Lipschitz smoothness. If a scalar function $F(x)$ is L-Lipschitz smooth, then for any point x, the function lies inside the shaded region shown in the picture, as specified by (2.2)



Smoothness

[illegible]

$$|\nabla f(x) - \nabla f(y)| \leq L||x - y||$$

[illegible] f [illegible] L-[illegible] ∇f [illegible] L-Lipschitz [illegible]

Gradient Decent & Mini-batch Stochastic Gradient Decent

Gradient Decent

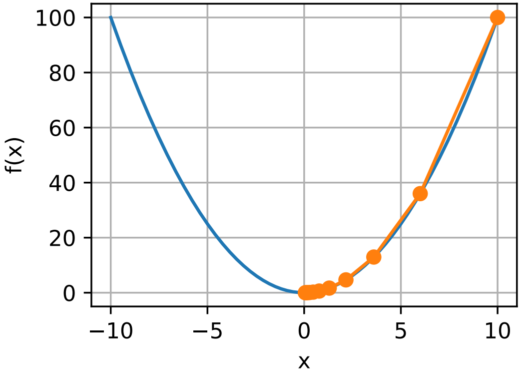
□□□□□□□□

$$w_{t+1} = w_t - \eta_t \nabla f(w_t)$$

□□□η_t □□□□□w_t □□□□$f(w_t)$ □□□□□□□ $\eta_t \leq \eta_{t-1} \leq \cdots \leq \eta_1$ □□□□□□

□□□□□□□□□□□□□□□

$$\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \|w_t - \eta_t \nabla f(w_t) - w^*\|^2 &(1)\\
&= \|w_t - w^*\|^2 - 2\eta_t \langle \nabla f(w_t), w_t - w^* \rangle + \eta_t^2 \|\nabla f(w_t)\|^2 &(2)
\end{aligned}$$



Gradient Decent

[illegible]

[illegible] $f(x)$ [illegible] x [illegible]

$$f(x+\epsilon) = f(x) + \epsilon f'(x) + o(\epsilon^2)$$

[illegible] $\eta > 0$[illegible] $\epsilon = -\eta f'(x)$[illegible]

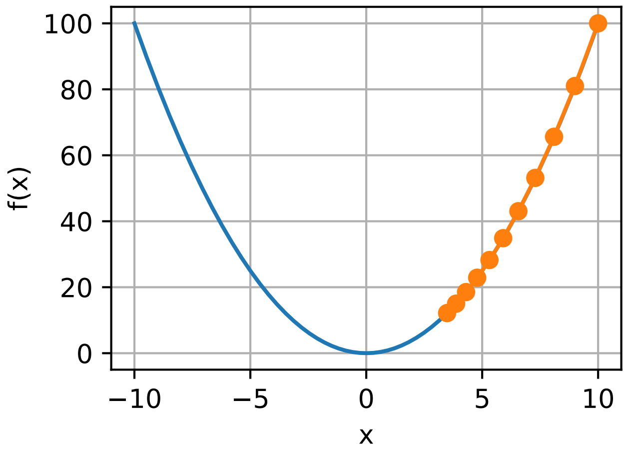
$$f(x - \eta f'(x)) = f(x) - \eta f'(x)^2 + o(\eta^2 f'(x)^2)$$

[illegible] $f'(x) \neq 0$[illegible] η [illegible] $o(\eta^2 f'(x)^2)$ [illegible]

$$f(x - \eta f'(x)) < f(x)$$

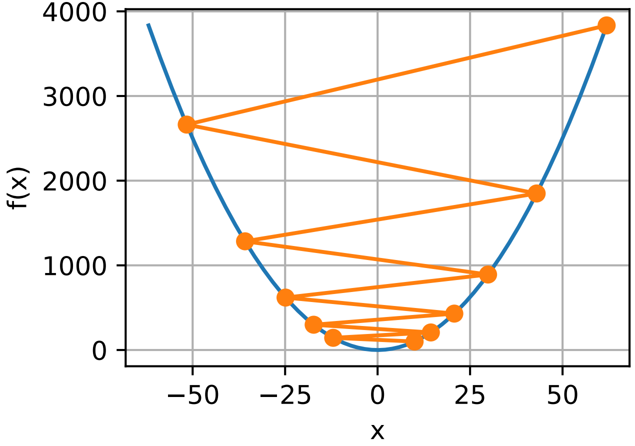
Gradient Decent (Learning Rate)

1.



Gradient Decent (Learning Rate)

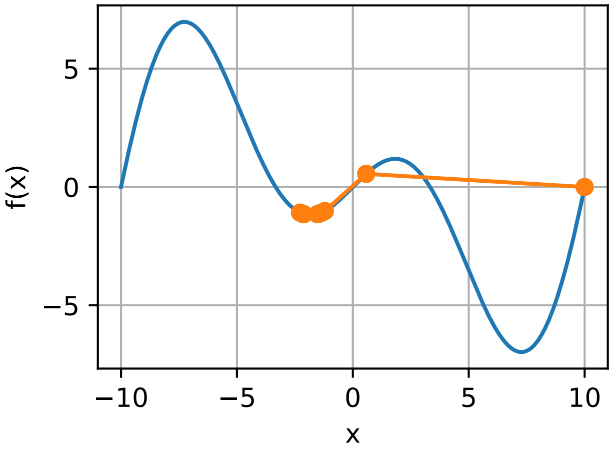
2. □□□□□□□□□□□□□□□□□□□□□□□□□□



Gradient Decent (Learning Rate)

3.

🤯

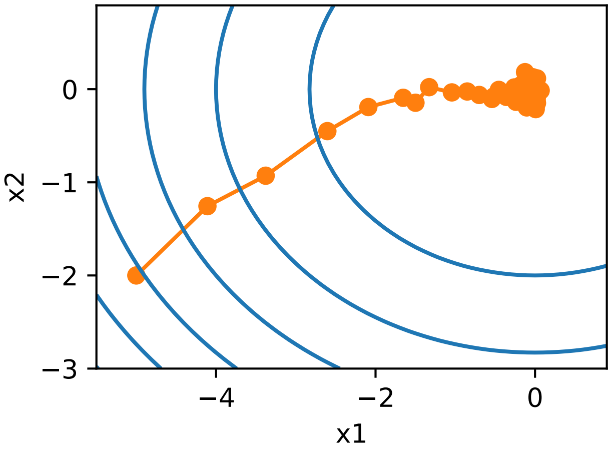


Stochastic Gradient Decent

[illegible] sample $(\mathbf{x_i}, y_i)$[illegible]

$$w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t)$$

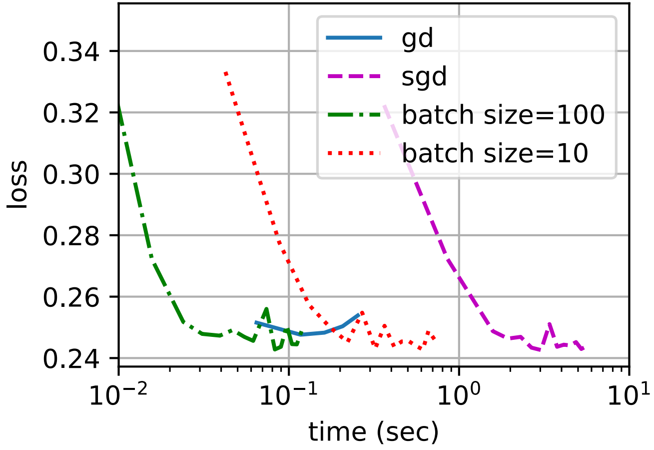
[illegible] $g(w_t) = \nabla f(w_t; \xi_t)$ [illegible]



Mini-batch Stochastic Gradient Decent

Mini-batch Stochastic Gradient Descent □□□□□□□□□□□□□□□□□ b □□□□□□□□□□□□□$g(w_t) = \frac{1}{b}\sum_{i=1}^{b} \nabla f(w_t; \xi_{t,i})$□Mini-batch Stochastic Gradient Descent □□□□

$$w_{t+1} = w_t - \frac{\eta_t}{b}\sum_{i=1}^{b} \nabla f(w_t; \xi_{t,i})$$



Mini-batch Stochastic Gradient Decent

Unbiased Estimate

mini-batch SGD

$$\mathbb{E}_{\xi}[g(w;\xi)] = \nabla f(w)$$

Gradient Bounded Variance Assumption

mini-batch sample $\nabla f(w;\xi)$

$$\mathrm{Var}(\nabla f(w;\xi)) \leq \sigma^2$$

Mini-batch Stochastic Gradient Decent

□□□□□□□□□□□□□□□□□□□□□□□□□

$$\mathrm{Var}(g(w;\xi)) = \mathbb{E}_\xi[||g(w;\xi)||^2] - ||\mathbb{E}_\xi[g(w;\xi)]||^2 \le \frac{\sigma^2}{b}$$

□□

$$\mathbb{E}_\xi[||g(w;\xi)||^2] \le ||\nabla f(w)||^2 + \frac{\sigma^2}{b}$$

Newton's Method

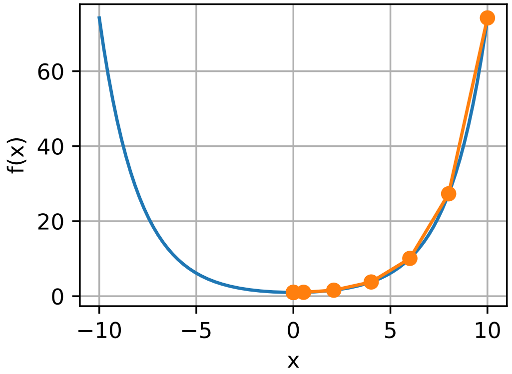
[illegible]

$$f(x+\epsilon) = f(x) + \epsilon \nabla f(x) + \frac{1}{2}\epsilon^2 \nabla^2 f(x) + o(\epsilon^3)$$

[illegible] $\mathbf{H} = \nabla^2 f(x)$ [illegible] Hessian [illegible] $n \times n$ [illegible] i [illegible] j [illegible] $\frac{\partial^2 f}{\partial x_i \partial x_j}$[illegible]

[illegible] $\nabla f(x) = \frac{f(x+\epsilon)-f(x)}{\epsilon}$ [illegible] 0[illegible] ϵ [illegible]

$$\nabla f(x) + \mathbf{H}\epsilon = 0 \Rightarrow \epsilon = -\mathbf{H}^{-1}\nabla f(x)$$



Newton’s Method

[illegible] Hessian [illegible]

[illegible]

Convergence Analysis

[illegible] & GD[illegible]mini-batch SGD[illegible] & mini-batch SGD[illegible]

Convergence Definition

[illegible]

1. [illegible]$\mathbb{E}f(w_T) - f^* \le \epsilon(T)$[illegible] f^* [illegible]

2. [illegible]$\mathbb{E}||w_T - w^*||^2 \le \epsilon(T)$[illegible] w^* [illegible]

[illegible]$\epsilon(T) \to 0$[illegible] $\log \epsilon(T)$ [illegible]

- 1[illegible] $\log \epsilon(T)$ [illegible] $-T$ [illegible]
- 2[illegible] $\log \epsilon(T)$ [illegible] $-T$ [illegible]
- 3[illegible] $\log \epsilon(T)$ [illegible] $-T$ [illegible] $\log \log \epsilon(T)$ [illegible] $-T$ [illegible]

[illegible] 0 [illegible]

3. [illegible]$\min_{t=1,\cdots,T} \mathbb{E}||\nabla f(w_T)||^2 \le \epsilon(T)$ [illegible] $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}||\nabla f(w_t)||^2$ [illegible] 0

Convergence Analysis (μ-strongly convex and L-smooth & GD)

□□□□□□ f □ R^d □□μ-□□□□□□□ L-□□□□□□ $\eta \leq \frac{1}{L}$ □□□□□ w_0 □□□□ t □□□□$f(w_T)$ □ bounded □□□

$$f\left(w_t\right) - f^* \leq (1 - \eta\mu)^t \left(f\left(w_0\right) - f^*\right)$$

Proof:

□□□□□□□

$$\begin{aligned}
f\left(w_{t+1}\right) - f\left(w_t\right) &= f\left(w_t - \eta \nabla f\left(w_t\right)\right) - f\left(w_t\right) \\
&\leq \nabla f\left(w_t\right)^{\top} \left(w_t - \eta \nabla f\left(w_t\right) - w_t\right) + \frac{L}{2} \left\|w_t - \eta \nabla f\left(w_t\right) - w_t\right\|^2 \\
&= -\eta \left\|\nabla f\left(w_t\right)\right\|^2 + \frac{L}{2} \eta^2 \left\|\nabla f\left(w_t\right)\right\|^2 \\
&= \left(\frac{L}{2} \eta^2 - \eta\right) \left\|\nabla f\left(w_t\right)\right\|^2
\end{aligned}$$

[illegible] η [illegible]

$$2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

[illegible]

$$\begin{aligned} f\left(w_{t+1}\right) - f\left(w_t\right) &\leq \eta\left(1 - \frac{L}{2}\eta\right)\left(-\left\|\nabla f\left(w_t\right)\right\|^2\right) \\ &\leq \eta\left(1 - \frac{L}{2}\eta\right)\left(2\mu\left(f\left(w_t\right) - f^*\right)\right) \end{aligned}$$

[illegible] $\eta \leq \frac{1}{L}$ [illegible] $(1 - \frac{L}{2}\eta) \geq \frac{1}{2}$ [illegible]

$$f\left(w_{t+1}\right) - f\left(w_t\right) \leq -\eta\mu(f\left(w_t\right) - f^*)$$

[illegible] $f(x)$ [illegible] $+ - f^*$ [illegible]

$$\begin{aligned} f\left(w_{t+1}\right) - f^* + f^* - f\left(w_t\right) \leq -\eta\mu(f\left(w_t\right) - f^*) \\ \Rightarrow f\left(w_{t+1}\right) - f^* \leq (1 - \eta\mu)(f\left(w_t\right) - f^*) \end{aligned}$$

□□□□□□□

$$
\begin{aligned}
f\left(w_{t+1}\right)-f^{*} &\leq(1-\eta \mu)\left(f\left(w_{t}\right)-f^{*}\right) \\
&\leq(1-\eta \mu)(1-\eta \mu)\left(f\left(w_{t-1}\right)-f^{*}\right) \\
&\leq \cdots \\
&\leq(1-\eta \mu) \cdots(1-\eta \mu)\left(f\left(w_{1}\right)-f^{*}\right) \\
&\leq(1-\eta \mu)^{t+1}\left(f\left(w_{0}\right)-f^{*}\right)
\end{aligned}
$$

□□□□□□□□ boundary□□□□□□□□

□□□□□□□□□□ $f(w_t) - f^* \leq \epsilon$□□□

$$
\begin{gathered}
(1-\eta \mu)^{t}(f(w_0)-f^{*}) \leq \epsilon \\
\Rightarrow t \log (1-\eta \mu)+\log (f(w_0)-f^{*}) \leq \log \epsilon \\
\Rightarrow t \log \frac{1}{1-\eta \mu}-\log (f(w_0)-f^{*}) \geq \log \frac{1}{\epsilon} \\
\Rightarrow t=O(\log \frac{1}{\epsilon})
\end{gathered}
$$

□□□□□□□□□ GD □□□□□□□□□

Convergence Analysis (μ-strongly convex and L-smooth & mini-batch SGD)

f R^d μ- L- $\eta \leq \frac{1}{L}$ w_0 t mini-batch SGD $\mathbb{E}[f(w_t)]$ bounded

$$\mathbb{E}[f(w_t)] - f(w^*) - \frac{\eta L \sigma^2}{2\mu b} \leq (1-\eta\mu)^t \left(\mathbb{E}[f(w_0)] - f(w^*) - \frac{\eta L \sigma^2}{2\mu b} \right)$$

Proof:

$$\begin{aligned} f(w_{t+1}) - f(w_t) &\leq \nabla f(w_t)^\top (w_{t+1} - w_t) + \frac{L}{2}||w_{t+1} - w_t||^2 & (3) \\ &\leq -\eta \nabla f(w_t)^\top g(w_t;\xi_t) + \frac{L}{2}|| -\eta g(w_t;\xi_t)||^2 & (4) \\ &\leq -\eta \nabla f(w_t)^\top g(w_t;\xi_t) + \frac{L\eta^2}{2}||g(w_t;\xi_t)||^2 & (5) \end{aligned}$$

[illegible]

$$\mathbb{E}[f(w_{t+1}) - f(w_t)] \leq -\eta \mathbb{E}[\nabla f(w_t)^\top g(w_t;\xi_t)] + \frac{L\eta^2}{2}\mathbb{E}[||g(w_t;\xi_t)||^2] \tag{6}$$

$$\mathbb{E}[f(w_{t+1})] - f(w_t) \leq -\eta||\nabla f(w_t)||^2 + \frac{L\eta^2}{2}||\nabla f(w_t)||^2 + \frac{L\eta^2\sigma^2}{2b} \tag{7}$$

$$\leq (\eta - \frac{L\eta^2}{2})(-||\nabla f(w_t)||^2) + \frac{L\eta^2\sigma^2}{2b} \tag{8}$$

[illegible] GD [illegible] $\eta < \frac{1}{L}$ [illegible]

$$\mathbb{E}[f(w_{t+1})] - f(w_t) \leq \frac{\eta}{2}(-||\nabla f(w_t)||^2) + \frac{L\eta^2\sigma^2}{2b} \tag{9}$$

$$\leq \frac{\eta}{2}(-2\mu(f(w_t) - f^*)) + \frac{L\eta^2\sigma^2}{2b} \tag{10}$$

$$\leq -\eta\mu(\mathbb{E}[f(w_t)] - f^*) + \frac{L\eta^2\sigma^2}{2b} \tag{11}$$

□□□ $+f^* - f^*$□□□

$$\mathbb{E}[f(w_{t+1})] - f^* \leq (1-\eta\mu)(\mathbb{E}[f(w_t)] - f^*) + \frac{L\eta^2\sigma^2}{2b}$$

□□□□□□□□□□□□□□□□□□□□□ x□□□

$$\mathbb{E}[f(w_{t+1})] - f^* + x \leq (1-\eta\mu)(\mathbb{E}[f(w_t)] - f^* + x)$$

□□□

$$(1-\eta\mu)x - x = \frac{L\eta^2\sigma^2}{2b}$$
$$x = -\frac{L\eta\sigma^2}{2\mu b}$$

□

$$\mathbb{E}[f(w_{t+1})] - f^* - \frac{L\eta\sigma^2}{2\mu b} \leq (1-\eta\mu)(\mathbb{E}[f(w_t)] - f^* - \frac{L\eta\sigma^2}{2\mu b})$$

□□

$$\mathbb{E}[f(w_t)] - f(w^*) - \frac{\eta L \sigma^2}{2\mu b} \leq (1 - \eta\mu)^t \left(\mathbb{E}[f(w_0)] - f(w^*) - \frac{\eta L \sigma^2}{2\mu b} \right)$$

□□□

□□

Convergence Analysis (non-convex & mini-batch SGD)

[illegible] f [illegible] R^d [illegible] L-[illegible] $\eta \leq \frac{1}{L}$ [illegible] w_0 [illegible] t [illegible] mini-batch SGD [illegible]f [illegible] bounded [illegible]

$$\mathbb{E}[\frac{1}{t}\sum_{i=1}^{t}||\nabla f(w_i)||^2] \leq \frac{L\sigma^2}{b} + \frac{2(f(w_0) - f(w_{\text{inf}}))}{t\eta}$$

Proof:

[illegible] L-[illegible] $\eta < \frac{1}{L}$ [illegible]

$$\mathbb{E}[f(w_{t+1})] - f(w_t) \leq -\eta||\nabla f(w_t)||^2 + \frac{L\eta^2\sigma^2}{2b}$$

[illegible] t [illegible]

$$\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}[f(w_{i+1})]-f(w_i) \leq -\frac{\eta}{2t}\sum_{i=1}^{t}||\nabla f(w_i)||^2+\frac{L\eta^2\sigma^2}{2b} \tag{12}$$

$$\frac{1}{t}\sum_{i=1}^{t}||\nabla f(w_i)||^2 \leq -\frac{2\mathbb{E}[f(w_{t+1})-f(w_0)]}{\eta t}+\frac{L\eta\sigma^2}{2b} \tag{13}$$

$$\leq \frac{2(f(w_0)-f(w_{\inf}))}{\eta t}+\frac{L\eta\sigma^2}{2b} \tag{14}$$

[illegible]

[illegible] $t=O(\frac{1}{\epsilon})$ [illegible]

Other Convergence Analysis

Other Convergence Analysis

Condition	GD	SGD
Convex	$O(\frac{1}{\sqrt{T}})$	$O(\frac{1}{\sqrt{T}})$
+ Lipschitz	$O(\frac{1}{T})$	$O(\frac{1}{\sqrt{T}})$
+ Strongly Convex	$O(c^T)$	$O(\frac{1}{T})$

> 10-725/36-725: Convex Optimization(Fall 2018), Lecture 24: November 26, Ryan Tibshirani

Distributed Optimization

- Distributed Synchrounous SGD
- Distributed Asynchrounous SGD
- Federated Learning

Reference

1. 《Convex Optimization》, Stephen Boyd
2. 《Optimization Algorithm for Distributed Machine Learning》, Gauri Joshi
3. 《Dive into Deep Learning》, D2L.ai
4. 《Optimization Methods for Large-Scale Machine Learning》, Léon Bottou

Learning more about optimization and convergence analysis:

https://blog.bj-yan.top/tags/convergence-analysis/