LEC3 DSA5202

Recap:

1. convexity $\longleftrightarrow$ $H_f(x) \succcurlyeq 0$ for $\forall x \in D$

   c-strongly convexity $\longleftrightarrow$ $H_f(x) \succcurlyeq cI$ for $\forall x \in D$

   (variants)

2. Newton's method (pure & line search)
   - pure $\to$ local convergence
   - line search $\to$ global convergence

3. Gradient Descent method $\longrightarrow$ some convergence result (fixed step length)

   (require convexity condition)

IDEA of these algos:

$$w^* = \underset{w}{\operatorname{argmin}}\, f(w) \xrightarrow[\text{condition}]{\text{necessary}} \nabla f(w^*) = 0$$

$\longrightarrow$ only for linear $f(\cdot)$, $\nabla f(w^*) = X^T(Xw^* - y) = 0$ is tractable

$$\Downarrow$$

$$w^* = (X^T X)^{-1} X^T y$$

$\longrightarrow$ for most $f(\cdot)$. WE WANT TO FIND A SET OF $\{w_i\}_{i=1}^{K}$

s.t $w_i \longrightarrow w^*$ $(i \to \infty)$ [iterative approach]

$$\begin{cases} \text{Newton Update: } & w_{k+1} = w_k - H_f(w_k)^{-1} \nabla f(w_k) \\ \text{GD Update: } & w_{k+1} = w_k - \alpha_k \nabla f(w_k) \end{cases}$$

Today's lecture

1. Issue of Gradient Descent

Population Risk Minimization

① Consider our high-level problem :

$z = (x, y)$

$F(w, z)$

intractable due to $\mathcal{D}$ $\Leftarrow$ $w^* = \underset{w}{\operatorname{argmin}} \ \underbrace{\boxed{\mathbb{E}_{(x,y)\sim\mathcal{D}} [\ell(h_w(x), y)]}}_{f(w)}$

e.g. $\begin{cases} \ell(z, z') = \|z - z'\|_2^2 \longrightarrow \boxed{\text{regression}} \\ \ell(z, z') = D_{KL}(z', z) = \sum_i z'(i) \log \frac{z'(i)}{z(i)} \longrightarrow \boxed{\text{classification}} \end{cases}$

Now, $\nabla f(w) = \nabla \mathbb{E}_{z\sim\mathcal{D}} [F(w, z)] \longrightarrow$ intractable

$\neq \mathbb{E}_{z\sim\mathcal{D}} [\nabla F(w, z)]$

② Surrogate : $\boxed{\text{PRM} \longrightarrow \text{ERM}}$ $\boxed{\text{LLN}}$

$\longrightarrow$ Empirical Risk Minimization

$$\hat{w} = \underset{w}{\operatorname{argmin}} \ \underbrace{\boxed{\frac{1}{n} \sum_{i=1}^{n} F(w, z_i)}}_{\hat{f}_n(w)} \xrightarrow{P} \mathbb{E}_{z\sim\mathcal{D}} [F(w, z)] = f(w)$$

Recap: $\underset{\text{population}}{f(w)} = \mathbb{E}_{z\sim\mathcal{D}} [F(w, z)]$

$\longrightarrow$ If we want to apply GD Framework,

then $w_{k+1} = w_k - \alpha_k \cdot \nabla \hat{f}_n(w_k)$

$$= w_k - \alpha_k \cdot \frac{1}{n} \underbrace{\sum_{i=1}^{n} \nabla F(w_k, z_i)}_{n\text{-summation}}$$

$\Rightarrow$ ☆ 1 update $\longleftrightarrow$ n calculation (gradient)

$\Downarrow$

inefficient when n is big (large dataset)

2. SGD $\longrightarrow$ Stochastic Gradient Descent

① original problem:

$$w^* = \arg\min_w f(w) := \mathbb{E}_{z\sim\mathcal{D}}[F(w, z)]$$

$$\longrightarrow \nabla f(w) = \nabla \mathbb{E}_{z\sim\mathcal{D}}[F(w, z)]$$

(under strong regularity of $F(\cdot,\cdot)$) $$= \mathbb{E}_{z\sim\mathcal{D}}[\nabla_w F(w, z)]$$

not always correct!

② ERM Surrogate:

$$\hat{w} = \arg\min_w \hat{f}_n(w) := \frac{1}{N}\sum_{i=1}^{N} F(w, z_i)$$

$$\longrightarrow \nabla \hat{f}_n(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla_w F(w, z_i)$$

$\Rightarrow$ $O(n)$ computational & storage cost per update

③ SGD $\longrightarrow$ using $\boxed{\begin{array}{c}\nabla_w F(w, z_I) \\ I \sim \text{uniform}[1, \dots, N]\end{array}} \approx \boxed{\nabla_w \hat{f}_n(w)}$

if we fix $z_1, \dots, z_n$, then this is a constant

$\Rightarrow$ only use one data point per update ($O(1)$)

consider the relationship:

a) $\underbrace{\nabla_w \hat{f}_n(w)}_{\text{R.V. with respect to } \{z_1, \dots, z_N\}} \longleftrightarrow \nabla_w f(w)$

$$\Rightarrow \mathbb{E}_z [\nabla_w \hat{f}_n(w)] = \mathbb{E}_{z_i} [\nabla_w F(w, z_i)]$$

$\star$

$$\text{(under regularity)} = \nabla_w \mathbb{E}_{z_i} [F(w, z_i)]$$

$$= \nabla_w f(w)$$

unbiased estimator

b) $\underbrace{\nabla_w \hat{f}_{SGD}(w) := \boxed{\nabla_w F(w, z_I)}}_{\text{R.V. with respect to } I \sim \text{uniform}([n])} \longleftrightarrow \nabla_w \hat{f}_n(w)$

stochastic gradient

$$\Rightarrow \mathbb{E}_I [\nabla_w \hat{f}_{SGD}(w)] = \mathbb{E}_I [\nabla_w F(w, z_I)]$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot \nabla_w F(w, z_i)$$

$$= \hat{f}_n(w)$$

$\Rightarrow$ **Stochastic gradient descent (SGD)** update:

$$w_{k+1} = w_k - \alpha_k \cdot \nabla_w F(w_k, z_{I_k})$$

where $I_k \sim \text{uniform}(1, 2, \dots, N)$ $\longleftrightarrow$ stochastic!

$\star$ for each update, we only require $O(1)$ computation!!!

Previously: $w_{k+1} = w_k - \alpha_k \cdot \frac{1}{N} \sum_{i=1}^{N} \nabla_w F(w_k, z_i)$

3. Convergence Result of SGD

Recap: under convexity regularity of $f(\cdot)$, GD method can achieve convergence to minimizer with sufficiently small step length.

Analysis:

$f(w)$ in optimization framework

$$\nabla F(w, z_I) = \nabla \boxed{\hat{f}_n(w)} + \xi \qquad \mathbb{E}_I[\xi] = 0$$

R.V. with respect to $I \sim \text{unif}(1, \dots, N)$

Assume : $\mathbb{E}_I[\xi^2] \leq \sigma^2$

Thm:

f is c-strongly convex, ∇f is L-Lipschitz.

then with $\boxed{\text{fixed step length } \alpha} \leq c/L^2$, we have Q-linear

$\boxed{\text{Linear Convergence}}$

$$\longrightarrow \mathbb{E}\|w_n - w^*\|_2^2 \leq (1-c\alpha)^n \mathbb{E}\|w_0 - w^*\|_2^2 + \boxed{\frac{\sigma^2 \alpha}{c}}$$

$\boxed{w^* = \underset{w}{\operatorname{argmin}} f(w)}$

$\boxed{\text{Trade off}}$

noise

intuitively, this term comes from all the randomness of sampling for each update

Note: $f(w) = \frac{1}{N} \sum_{i=1}^{N} F(w, z_i)$

$\boxed{\text{Pf sketch}}$: ① $w_n - w^* = w_{n-1} - w^* - \alpha \nabla_w F(w_{n-1}, z_{I_{n-1}})$

$$= w_{n-1} - w^* - \alpha \nabla_w f(w_{n-1}) - \alpha \xi_{n-1}$$

② $\mathbb{E}_{I_{n-1}} \|w_n - w^*\|_2^2$ (conditioned on w_{n-1})

$$= \boxed{\|w_{n-1} - w^* - \alpha \nabla_w f(w_{n-1})\|_2^2} \rightarrow \text{previous result}$$

$$+ \boxed{\mathbb{E}_{I_{n-1}}[\xi_{n-1}] \cdot *} \to 0$$

$$+ \boxed{\alpha^2 \mathbb{E}_{I_{n-1}}[\xi_{n-1}^2]} \leq \alpha^2 \sigma^2$$

$$① + ② \Rightarrow \mathbb{E}_{I_{n-1}} \|w_n - w^*\|_2^2 \leq (1-\alpha c)\|w_{n-1} - w^*\|_2^2 + \alpha^2\sigma^2$$

$$\Rightarrow \mathbb{E}_{I_{n-2}, I_{n-1}}[\|w_n - w^*\|_2^2 \mid w_{n-2}]$$

$$\leq (1-\alpha c)^2 \|w_{n-2} - w^*\|_2^2 + [(1-\alpha c) + 1]\alpha^2\sigma^2$$

$$\Rightarrow \cdots \quad \mathbb{E}_{I_0, \ldots I_{n-1}}[\|w_n - w^*\|_2^2]$$

$$= \mathbb{E}_{I_0}[\mathbb{E}_{I_1, \ldots, I_{n-1}}[\|w_n - w^*\|_2^2 \mid w_0]]$$

$$\leq (1-\alpha c)^n \|w_0 - w^*\|_2^2 + [1 + \cdots + (1-\alpha c)^{n-1}]\alpha^2\sigma^2$$

$$= (1-\alpha c)^n \|w_0 - w^*\|_2^2 + \frac{1-(1-\alpha c)^n}{\alpha c} \cdot \alpha^2\sigma^2$$

$$= (1-\alpha c)^n \|w_0 - w^*\|_2^2 + \frac{\alpha}{c} \cdot \sigma^2$$

#

Rmk: There is no guarantee that $w_n \longrightarrow w^*$ as $n \to +\infty$

since we only have $\mathbb{E}_I \|w_n - w^*\|_2^2 \leq \frac{\sigma^2}{c} \cdot \alpha$ as $n \to +\infty$

4. Convergence Result of SGD with different step length

not fixed step length

Notation

$$\rightarrow \begin{cases} S_{1,n} = \sum_{k=1}^{n} \alpha_k \\ S_{2,n} = \sum_{k=1}^{n} \alpha_k^2 \end{cases}$$

Theorem:

$\rightarrow$ if f is c-strongly convex & ∇f is L-Lip.

then with step length $\alpha_k \le c/L^2$, we have:

$\boxed{\text{no need to be constant}}$

$$\Rightarrow \min_{k \in [n]} \mathbb{E}_I \|w_k - w^*\|_2^2 \le \frac{\mathbb{E}_I \|w_0 - w^*\|_2^2 + 6^2 S_{2,n}}{c\, S_{1,n}}$$

ISSUE: convergence might be slow!

☆ **Rmk**: based on this thm, we would like to choose $\{\alpha_k\}_{k=1}^{n}$

such that $\begin{cases} S_{1,\infty} = \sum_{k=1}^{\infty} \alpha_k = \infty \\ S_{2,\infty} = \sum_{k=1}^{\infty} \alpha_k^2 < \infty \end{cases}$ to guarantee convergence

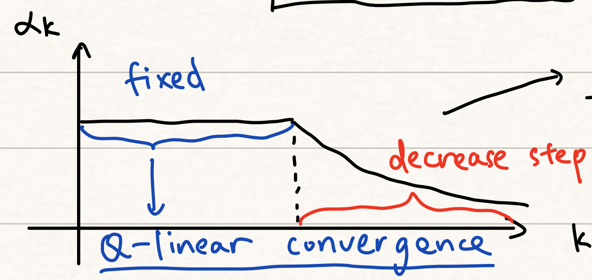
get rid of noise!

$$\lim_{n\to\infty} \min_{k \in [n]} \mathbb{E}_I \|w_k - w^*\|_2^2 = 0$$

$\rightarrow$ choice: $\alpha_k = \alpha_0 \dfrac{1}{k^t}$ $\quad t \in (1/2, 1]$ $\rightarrow$ $\boxed{\text{thm guarantee}}$

$\boxed{\alpha_k = \alpha_0 \cdot \frac{1}{k}}$

$\rightarrow$ choice 2:



Fixed + decrease schedule

decrease step length $\Rightarrow$ guarantee convergence

5. Mini-batch SGD

① $\nabla_w F(w, z_I) = \nabla_w \hat{f}_n(w) + \xi \qquad \mathbb{E}_I[\xi] = 0$

$$\boxed{\mathbb{E}_I[\xi^2] \longleftrightarrow \sigma^2}$$

previously, we have the result:

$\Downarrow$

$$\mathbb{E}_I \|w_n - w^*\|_2^2 \le (1 - \alpha \cdot c)^n \, \mathbb{E}_I \|w_0 - w^*\|_2^2 + \frac{\sigma^2}{c} \cdot \alpha$$

→ to improve performance ($w_n \to w^*$), we can decrease σ^2

→ idea: previously, we use $\nabla_w F(w, z_I) \approx \nabla_w \hat{f}_n(w)$

$\downarrow$ Bagging !!!

why not use $\frac{1}{B} \sum_{b=1}^{B} \nabla_w F(w, z_{I_b}) \approx \nabla_w \hat{f}_n(w)$

$\downarrow$

$\begin{cases} \text{same expectation as } \nabla_w F(w, z_I) \\ \text{but with } \frac{1}{B} \text{ variance } \left(\frac{\sigma^2}{B}\right) \end{cases}$

② mini-batch SGD: $\boxed{1 \ll B \ll N}$ $\begin{cases} B \ll N: \text{computation} \\ 1 \ll B: \text{reduce var.} \end{cases}$

$\downarrow$

$$w_{k+1} = w_k - \alpha_k \cdot \frac{1}{B} \sum_{b=1}^{B} \nabla_w F(w, z_{I_k^b})$$

$$\boxed{I_k^1, \ldots\ldots, I_k^B \sim \text{Uniform}(1, 2, \ldots, N)}$$

6. Momentum GD $\longrightarrow$ utilize momentum information in previous step

Framework:

Goal: $\hat{\omega} = \arg\min_{\omega} f(\omega)$

Update: $\omega_{k+1} = \omega_k - \alpha_k \cdot m_k$ $\qquad \boxed{\beta \in (0,1)}$

where $m_k = \beta \cdot \underbrace{m_{k-1}}_{\boxed{\text{previous state}}} + (1-\beta) \cdot \nabla_\omega f(\omega)$

Rmk:

① common choice of $\beta \longrightarrow 0.9$

② variants of momentum SGD $\longrightarrow$ $\begin{cases} \text{ADAM} \\ \text{AdaGrad} \end{cases}$

③ converge faster since it can ☆ accumulate "speed"

④ can help to escape bad local minima

⑤ Momentum SGD can be viewed as:

$\rightarrow$ increase batch size (since it takes previous update into consideration) to reduce variance