DSA5202 LEC2

Recap:

① $\hat{\omega} = \underset{\omega}{\operatorname{argmin}} f(\omega)$ $\qquad f(\omega) := \frac{1}{N}\sum_{i=1}^{N} L(y_i, h_\omega(x_i))$

② we need optimization to solve ① (approximately)

$\Downarrow$

stationary point $\leftarrow$ $\nabla f(\hat{\omega}) = 0$ (necessary condition)

③ convexity $\longrightarrow$ gurantee stationary point $\Leftrightarrow$ minimizer

(approximately $\Longrightarrow$ exactly)

$\boxed{\text{characterization}}$ of convexity:

$$\begin{cases} f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) & \forall x, y,\ \lambda \in [0,1] \\ f(y) \ge f(x) + \nabla f(x)^T (y-x) & \forall x, y \end{cases}$$

④ C-strongly convexity $\longrightarrow$ guarantee the unique minimizer

$\Rightarrow$ stationary point $\longrightarrow$ unique minimizer

$\downarrow$

$\boxed{\text{characterization}}$: $\begin{cases} f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{C}{2}\|y-x\|_2^2 \\ \langle \nabla f(x) - \nabla f(y), x-y \rangle \ge C\|y-x\|_2^2 \end{cases}$

$f(x)$ is C-strongly convex $\Leftrightarrow$ $g(x) = f(x) - \frac{C}{2}\|x\|_2^2$ is convex

$\boxed{\text{Today's lecture}}$

smallest eigenvalue

1. Positive Definite / Semi-Definite

① A is PD $\Leftrightarrow$ $x^T A x > 0$ for $\forall x \neq \underline{0}$ $\Leftrightarrow$ $\lambda_n > 0$

② A is PSD $\Leftrightarrow$ $x^T A x \geq 0$ for $\forall x \in \mathbb{R}^n$ $\Leftrightarrow$ $\lambda_n \geq 0$

③ $A \succeq cI \Leftrightarrow x^T A x \geq c\|x\|_2^2$ for $\forall x \in \mathbb{R}^n$

$\Leftrightarrow \lambda_n \geq C$

if f is well-conditioned, then H_f is symmetric

2. Hessian and convexity

$$[H_f(x)]_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

$\boxed{H_f : x \in \mathbb{R}^n \longrightarrow \mathbb{R}^{n \times n}}$

can characterize convexity & c-strongly convexity

a) $f \in C^2$ is convex on D

$\Leftrightarrow$ $H_f(x)$ is Positive Semi-Definite on all $x \in D$

[e.g.] $f(x) = \frac{1}{2} x^T A x$ $\Longrightarrow$ $\boxed{f(x) = \|x\|_2^2 \text{ is convex}}$

$\rightarrow \nabla f(x) = \frac{1}{2} A x + \frac{1}{2} A^T x$

$= A x$ (if A is symmetric)

$\rightarrow \nabla^2 f(x) = H_f(x) = A$

Therefore, $f(\cdot)$ is convex $\Leftrightarrow$ A is PSD

Proof: "$\Rightarrow$" f is convex

$\Rightarrow \forall x, y,\ f(y) \geq f(x) + \nabla f(x)^T (y - x)$

(Analyze: we want to show for $\forall x$, $z^T H_f(x) z \geq 0$, $\forall z$)

$\Rightarrow$ choose $y = x + \lambda z$: (fix x and randomly choose z)

then we have $f(x+\lambda z) \geq f(x) + \lambda \nabla f(x)^T z$

Also. from Taylor Expansion:

$$f(x+\lambda z) = f(x) + \nabla f(x)^T (\lambda z) + \frac{1}{2} (\lambda z)^T H_f(x) \cdot (\lambda z) + o(\lambda^2)$$

$$\geq f(x) + \nabla f(x)^T (\lambda z)$$

$\Rightarrow \lambda^2 \cdot z^T H_f(x) z + o(\lambda^2) \geq 0$

$\Rightarrow z^T H_f(x) z + o(1) \geq 0 \quad \forall \lambda \geq 0$ (let $\lambda \to 0$)

$\Rightarrow z^T H_f(x) z \geq 0$ holds for $\forall z \in \mathbb{R}^n$

"$\Leftarrow$" Now, $H_f(x)$ is PSD for $\forall x \in D$

$\Rightarrow z^T H_f(x) z \geq 0$ for $\forall z \in \mathbb{R}^n$, $\forall x \in D$

$\Rightarrow$ for arbitrary $y \in \mathbb{R}^n$. $y = x + z$ ($z = y - x$)

then $f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} z^T H_f(\hat{x}) z$

$\geq f(x) + \nabla f(x)^T (y-x)$

$\Rightarrow f(\cdot)$ is a convex function

#

$$\boxed{\begin{array}{l} \text{b) } f(x) \text{ is a } c\text{-strongly convex function on } D \\ \Leftrightarrow H_f(x) \succeq cI \text{ at each } x \in D \end{array}}$$

Pf Sketch. $f(x)$ is c-strongly convex

$\Leftrightarrow g(x) = f(x) - \frac{c}{2}\|x\|_2^2$ is convex

$\Leftrightarrow H_g(x) \succeq 0$ for $\forall x \in D$

$\Leftrightarrow H_f(x) - cI \succeq 0$ for $\forall x \in D$

$\Leftrightarrow H_f(x) \succeq cI$ for $\forall x \in D$

[e.g.] Compute Hessian !!!

① $f(x) = e^x \quad x \in \mathbb{R}$

convex but not c-strongly convex for arbitrary c

$f''(x) = e^x \to 0$ when $x \to -\infty$

$\Rightarrow$ there does not exist $c > 0$ s.t $e^x \geq c$ for all $x \in \mathbb{R}$!

$\Rightarrow$ $f(\cdot)$ is not c-strongly convex on $\mathbb{R}$

② $f(x) = e^x \quad x \in \mathbb{R}^+$

$\to$ c-strongly convex for $c \leq 1$

③ $f(x) = \|x\|_2^2 \to$ c-strongly convex for $c \leq 2$

$\to H_f(x) = 2I$

④ $f(x) = \|Ax - b\|_2^2 \to$ convex

$$\begin{cases} \nabla f(x) = 2A^T(Ax - b) \\ H_f(x) = 2A^TA \succeq 0 \end{cases}$$

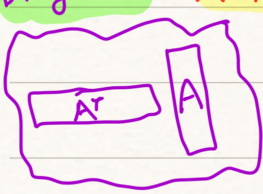
Rmk: 1. in e.g. 4, $f(x)$ is at least convex.

2. in order to further determine whether $f(x)$ is c-strongly convex or not, check the smallest eigenvalue of matrix A^TA ($\lambda_n \geq c$ or not) $\to$ Implementation

$\Updownarrow$

whether A is full column rank matrix

Diagram



3. $\boxed{\text{Algos}}$ to find w^* such that $\nabla f(w^*) = 0$

Rmk: actually there are 2 perspectives

a) w^* s.t $\nabla f(w^*) = 0$

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

here, we focus on this

idea: use $\boxed{w^{(k+1)}}$ to approximate w^*

$\Downarrow$

iterative algorithm

a) stop criterion $\boxed{w_1, w_2, \ldots\ldots, w_k \longrightarrow w^*}$

① stop at $k = T$, T is pre-determined $\Longrightarrow$ $\boxed{\text{deep learning}}$

② stop at $\|\nabla f(w_k)\| \leq$ tolerance

tolerance $\Longrightarrow$ small precision requirement

③ stop at $\|w_{k+1} - w_k\| \leq$ tolerance

b) Question: How many iterations are required to satisfy

$$\|w_k - w^*\| \leq \varepsilon$$

$\downarrow$

$\boxed{\text{convergence rate}}$

c) update cost $w_{k+1} = w_k + \alpha_k p_k$

p_k
- $\nearrow$ first order information $\longrightarrow \nabla f(w_k)$
- $\searrow$ second order information $\longrightarrow H_f(w_k)$

d) scalability & parallelization

4. Newton's Method

$\boxed{\text{formulation}}$

$$\begin{cases} ① \ \nabla f(w) = 0 \ \approx\!\!> \ \nabla f(w_k) + H_f(w_k)\cdot(w - w_k) = 0 \\ ② \ w^* = \underset{w}{\operatorname{argmin}} f(w) \ \approx\!\!> \ w = \underset{w}{\operatorname{argmin}} \left\{ \nabla f(w_k)^T (w - w_k) + \frac{1}{2}(w - w_k)^T H_f(w_k)(w - w_k) \right\} \end{cases}$$

$\Rightarrow$ $\boxed{w_{k+1} = w_k - H_f(w_k)^{-1} \nabla f(w_k)}$ $\rightarrow$ Newton's method

$\longrightarrow$ **Theorem**: (Local quadratic convergence rate)

(no need of convexity)

Suppose $\|w_0 - w^*\|$ is sufficiently small, $H_f(w)^{-1}$ & $\nabla f(w)$ are L-Lipschitz, then there exists M s.t

$$\|w_{k+1} - w^*\| \leq M \|w_k - w^*\|^2$$

$\boxed{\nabla f(w^*) = 0}$ $\rightarrow$ stationary point

[e.g.] $w_0 = 1.1$ $w_1 = 1.01$, $w_2 = 1.0001$ .... stationary point

[Pf Sketch]: $\begin{cases} w_{k+1} = w_k - H_f(w_k)^{-1} \cdot \nabla f(w_k) \\ w^* \text{ satisfy that } \nabla f(w^*) = 0 \end{cases}$

$$\Leftrightarrow \nabla f(w_k) + H_f(\bar{w})(w^* - w_k) = 0$$

$\boxed{\bar{w} \in (w_k, w^*)}$ $\Rightarrow w^* = w_k - H_f(\bar{w})^{-1} \nabla f(w_k)$

Thus, $w_{k+1} - w^* = (H_f(\bar{w})^{-1} - H_f(w_k)^{-1}) \cdot \nabla f(w_k)$

$$= \underbrace{(H_f(\bar{w})^{-1} - H_f(w_k)^{-1})}_{\text{L-lip. of } H_f(w)} \cdot \underbrace{(\nabla f(w_k) - \nabla f(w^*))}_{\text{L-lip. of } \nabla_f(w)}$$

Rmk: ① quadratic local convergence rate

② converge to w^*, which is the stationary point ($\boxed{\nabla f(w^*) = 0}$)

(not minimizer)

step size issue { ③ no global convergence guarantee

$\quad\hookrightarrow$ line search rule modification (globalize)

Hessian matrix issue { ④ if $H_f(w_k)$ is not PD, then Newton's direction

$p_k = -H_f(w_k)^{-1} \nabla f(w_k)$ may not be descent direction

⑤ invert / record Hessian (attain $H_f(\cdot)^{-1}$) Is expensive

(n^2) space

Possible Solution: $\boxed{\text{quasi-newton}}$

a) approximate $H_f(w_k)^{-1}$ via $\nabla f(w_k)$ [BFGS]

b) solve linear equation $H_f(w_k)\, p_k = -\nabla f(w_k)$ via iterative methods

5. Gradient Descent

① idea: $w_{k+1} = w_k + \alpha_k p_k$

$\Downarrow$

$$f(w_{k+1}) \approx f(w_k) + \alpha_k \nabla f(w_k)^T p_k + o(\alpha_k)$$

$$p_k^{GD} = -\nabla f(w_k) \Rightarrow \text{steepest descent direc.}$$

Rmk: ① if α_k is sufficiently small, $f(w_{k+1}) \leq f(w_k)$

$\downarrow$

$\boxed{\text{descent direction}}$

② Another formulation:

$$p_k^{GD} = \arg\min_p \underbrace{f(w_k) + \nabla f(w_k)^T p}_{\text{approximation term}} + \underbrace{\tfrac{1}{2}\|p\|_2^2}_{\text{proximal term}}$$

② Determine step length α_k

- exact line search: $\alpha_k = \arg\min_\alpha f(w_k + \alpha p_k)$
- inexact line search: backtracking $\xrightarrow{\text{similar}}$ exponential decay / cosine schedule
- fixed step size $\alpha_k \equiv \alpha \longrightarrow \boxed{\text{most popular}}$

③ Thm: $\boxed{\text{convex}}$ $w^* = \arg\min f(w)$

if $f(w) \geq 0$ & ∇f is L-Lipschitz, then fixed step size GD with $\alpha_k \equiv \alpha \leq \frac{1}{L}$ satisfy: $\boxed{\text{sub-linear}}$

$$0 \leq f(w_n) - f(w^*) \leq \frac{1}{n} \frac{\|w_0 - w^*\|_2^2}{2\alpha}$$

④ Thm: (much better than previous result due to strongly convexity)

if f is c-strongly convex and $\alpha \le \frac{c}{L^2}$, then Linear converg

$w^* = \arg\min f(w)$

$$0 \le f(w_n) - f(w^*) \le (1 - c\alpha)^n \, L \, \| w_0 - w^* \|_2^2$$

"exponential convergence"

Rmk: in application, if our $f(\cdot)$ is strongly convex, then we can choose our learning rate small but not too small to achieve a good convergence rate.

satisfy $\alpha \le \frac{c}{L^2}$

make $1 - c\alpha \ll 1$