"Optimization Homework"

1- Convexity:

1)- f convex, $f(w_1) \leq f(w)$ and $f(w_2) \leq f(w)$ for all $w \in \mathbb{R}^p$

so $f(\lambda w_1 + (1-\lambda) w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2)$

$$\leq \lambda f(w) + (1-\lambda) f(w)$$

$$\leq f(w)$$

2)- f μ-strongly convex, f only admits one minimizer?

Let's suppose that f has instead 2 minimizer.

so let w_1, w_2 be 2 local minimizers of f with $\begin{cases} f(w_1) \leq f(w_2) \\ w_1 \neq w_2 \end{cases}$

$$f(\lambda w_1 + (1-\lambda) w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2) - \frac{\lambda(1-\lambda)}{2} \mu \|w_1 - w_2\|^2$$

$$f(w_1) \leq f(w_2) \Rightarrow \lambda f(w_1) \leq \lambda f(w_2)$$

So: $\lambda f(w_1) + (1-\lambda) f(w_2) \leq \lambda f(w_2) + (1-\lambda) f(w_2)$

$$f(\lambda w_1 + (1-\lambda) w_2) \leq f(w_2) - \frac{\lambda(1-\lambda)}{2} \mu \|w_1 - w_2\|^2$$

$$f(\lambda w_1 + (1-\lambda) w_2) + \frac{\lambda(1-\lambda)}{2} \mu \|w_1 - w_2\|^2 \leq f(w_2)$$

However, w_2 is a local minimizer of f, yet $f(w_2) > f(w)$ which is a contradiction.

Finally, we can admit the unicity of w^* as a minimizer of f.

Scanned with CamScanner

1.3). f a convex differentiable function. $\nabla f(x^*) = 0 \Leftrightarrow f(x) \geq f^*(x)\ \forall x$

by definition: f convex and differentiable

$$\forall x, y: \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

$$\text{i.e:} \quad f(x) \geq f(x^*) + \langle \nabla f(x^*), \cancel{y} - x^* \rangle$$

$$\nabla f(x^*) = 0 \implies f(x) \geq f(x^*) \qquad \forall x \in \mathbb{R}^p$$

On the other hand,

if $f(x) \geq f(x^*)$

$\hookrightarrow \nabla f(x^*)^T (x - x^*) \leq 0$

so: if $x \geq x^* \Rightarrow \nabla f(x^*) \leq 0$

$x \leq x^* \Rightarrow \nabla f(x^*) \geq 0$ $\Big\}$ $\nabla f(x^*) = 0$

2/ Gradient descent:

1) we have f L-smooth: $f(y) \leq f(\omega) + \langle \nabla f(\omega), y - \omega \rangle + \frac{L}{2} \|y - \omega\|^2$

$$\langle \nabla f(\omega), y - \omega \rangle = -\eta \|\nabla f(\omega)\|^2$$

so: $\frac{L}{2} \|y - \omega\|^2 - \eta \|\nabla f(\omega)\|^2 = \eta(\eta \frac{L}{2} - 1) \|\nabla f(\omega)\|^2$.

$$\eta < 2/L \Rightarrow \eta \frac{L}{2} < 1 \Rightarrow (\eta \frac{L}{2} - 1) < 0$$

$$\eta > 0 \Rightarrow \eta(\eta \frac{L}{2} - 1) < 0.$$

D'où: $\langle \nabla f(\omega), y - \omega \rangle + \frac{L}{2} \|y - \omega\|^2 < 0$.

Finally: $f(y) \leq f(\omega)$.

2.2). Let's suppose $f(x) = x^2 \longrightarrow L = 2$

$\longrightarrow \eta > 1$

$$f(y) = (x - \eta 2x)^2$$
$$= (1 - 2\eta)^2 x^2$$

$\eta > 1 \Rightarrow \quad 1 - 2\eta < 1$

$(1 - 2\eta)^2 > 1$

$(1 - 2\eta)^2 x^2 > x^2 \iff f(y) > f(x)$

2.3). f μ- strongly convex.

we have: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$

$\hookrightarrow \min f(y) \geq \min(\quad\_\_\_\_\_\_\_\_\_\_\_\_ \quad)$

let ω^* be the minimizer of $f \iff f^* = f(\omega^*) \leq f(\omega) \quad \forall \omega \in \mathbb{R}^p$

also: $\nabla f(x) + \mu(y - x) = 0 \Rightarrow y = x - \frac{\nabla f(x)}{\mu}$

$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \geq f(x) - \frac{1}{\mu} \|\nabla f(x)\|^2 + \frac{1}{2\mu} \|\nabla f(x)\|^2$

So: $f(\omega^*) \geq f(\omega) - \frac{1}{2\mu} \|\nabla f(\omega)\|^2$

$f(\omega) - f^* \leq \frac{1}{2\mu} \|\nabla f(\omega)\|^2$.

3/. Logistic Regression:

1). $\phi(x) = \log(1 + e^{-x})$ and $g(x) = 1 + e^{-x}$

$$\phi'(x) = \frac{g'(x)}{g(x)} = \frac{-e^{-x}}{1 + e^{-x}}$$

$$\phi''(x) = \frac{g''(x) g(x) - g'(x)^2}{g^2(x)} = \frac{e^{-x}}{(1 + e^{-x})^2} > 0 \quad \forall x$$

$\hookrightarrow \phi$ est convex.

Scanned with CamScanner

3.1). pose $x \in \mathbb{R}$. $e^{-x} \ll 1$ and $\frac{e^{-x}+e^{x}}{2} \geqslant 1$

We have: $\phi''(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{2+e^{-x}+e^{x}}$

$2+e^{-x}+e^{x} \geqslant 4 \Rightarrow \phi''(x) \leqslant 1/4$

3.2). $f(w) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_i \langle w, x_i \rangle)$

$\nabla^2 f(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \phi(y_i \langle w, x_i \rangle$

$= \frac{1}{n} \mathcal{H}$ $\qquad$ $\mathcal{H}$: Hessian matrix of $W(\phi(x))$

ϕ is convex $\Rightarrow \mathcal{H}$ semi-definite positive.

$\Rightarrow \mathcal{H}$ symetric.

$\Rightarrow \mathcal{H}$ diagonalisable

$\mathcal{H} = P^T D P$ with $D(w)_{ii} = \phi''(y_i \langle w, x_i \rangle)$

3.3). We have $\phi''(x) > 0$ so $\phi''(y_i \langle w, x_i \rangle) > 0$.

So all the eignvalues of $\nabla^2 f$ (as the ones of D) are positive

Thus the Hessian of f is semi-definite positive.

Which give f is convex.

Moreover $\max(\lambda) \leqslant \frac{1}{4}$ since $\phi''(x) \leqslant 1/4$ $\forall$

$\downarrow$

we simplify the sum with $\frac{1}{n}$

With λ : eignvalue of the Hessian of f

As a result: $L = 1/4$

4)- Proximal Operations:

1). $\text{prox}_{u|\cdot|}(x) = \text{argmin } \frac{1}{2}\|y - x\|^2 + u\,|y|$

$$= \text{argmin} \begin{cases} uy + \frac{1}{2}\|y-x\|^2 & y>0 \\ -uy + \frac{1}{2}\|y-x\|^2 & y<0 \end{cases}$$

$\nabla(uy + \frac{1}{2}\|y-x\|^2) = u + y - x$ if $x - u > 0$

so $y^* = x - u$

the 2nd case: $y^* = x + u$ if $x + u < 0$

while if $|x| < u$, the only feasible point is:

the point of non-differentiability of the function

$\hookrightarrow \text{prox}_{u|\cdot|}(x) = 0$

2) $\text{prox}_{u\|\cdot\|_2^2}(x) = \text{argmin } \frac{1}{2}\|y-x\|^2 + u\,\|y\|^2$

$$\nabla\left(\frac{1}{2}\|y-x\|^2 + u\|y\|^2\right) = y - x + uy$$

$$y^* = \frac{x}{1+u}$$

3)- $\text{prox}_{I_C}(x) = \text{argmin } \frac{1}{2}\|y - x\|^2 + I_C(x)$

$$= \text{argmin} \begin{cases} \frac{1}{2}\|y-x\|^2 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

while: $\underset{x\in C}{\text{argmin}}\ \frac{1}{2}\|y - x\|^2$: orthogonal projection on C

Scanned with CamScanner

4.4). $\min_{w \in C} f(w) \iff \min_{w \in \mathbb{R}^p} \underbrace{f(w) + I_C(w)}$

. we have f L-smooth.

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

- To solve $\min_{w \in \mathbb{R}^p} f(w) + I_C(w)$ we proceed as follows

$$0 \in \partial(f w + I_C(w)) = \nabla f(w^*) + \partial I_C(w^*)$$

i.e: $-\nabla f(w^*) \in \partial I_C(w^*)$

$$w^* + \nabla f(w^*) \in w^* - \partial I_C(w^*)$$

so: $w^* \in w^* - \nabla f(w^*) - \partial I_C(w^*)$

$$w^* = \mathrm{prox}_{I_C}(w^* - \nabla f(w^*))$$

so the proximal gradient algo update will be:

$$w^{k+1} = \mathrm{prox}_{I_C}(w^k - \nabla f(w^k))$$

Meaning, each we are projecting "$w^k - \nabla f(w^k)$) on C to get w^{k+1}

which means that we get closer by the projection of w^k in the opposite direction of $\nabla f(w^k)$.

Scanned with CamScanner

5°/- Quasi-Newton methods:

$$x_{n+1} = x_n - \alpha A \nabla f(x_n)$$

$$x = Cy \cdot \quad \text{and} \quad g(y) = f(Cy)$$

1)- $\nabla g(y) = C^T \nabla f(Cy)$

$\nabla^2 g(y) = C^T \nabla^2 f(Cy) C$.

2). • We have $y_{n+1} = y_n - \alpha \nabla g(y_n)$

$$C y_{n+1} = C y_n - \alpha C \nabla g(y_n)$$

$$x_{n+1} = x_n - \alpha C C^T \nabla f(x_n)$$

• By identification we have: $A = CC^T$

so $C \in \mathcal{A} := \{ B \in \mathbb{R}^{p \times p}$, B invertible and BB^T positive definite symetric$\}$