1.1 Convexity

1. $f(w) = \frac{1}{2}(Xw-y)^T V(Xw-y) + \frac{\lambda}{2}\|w\|^2$

$= \frac{1}{2}(w^T X^T V X w - y^T V X w - w^T X^T V y + y^T V y) + \frac{\lambda}{2} w^T I w$

$\therefore \nabla^2 f = X^T V X + \lambda I$

- for $w \in \mathbb{R}^{d \times 1}$, $w^T X^T V X w = \|V^{\frac{1}{2}} X w\|_2^2 \geq 0$, $w^T \lambda I^T w = \lambda \|w\|_2^2 > 0$

$\therefore w^T \cdot \nabla^2 f \cdot w > 0$ $\therefore \nabla^2 f = X^T V X + \lambda I$ is positive semidefinite $\therefore f$ is convex

2. $f(w) = -y^T X w + 1^T V$, $V_i = \exp(w^T x^i)$

$= -y^T X w + \sum_{i=1}^{n} \exp(w^T x^i)$

$\nabla^2(-y^T X w) = 0$, Let $g(w) = \sum_{i=1}^{n} \exp(w^T x^i)$

$\therefore \forall j, k \in \{1, 2 \dots d\}$,

$\frac{\partial g}{\partial w_j} = \sum_{i=1}^{n} \exp(w^T x^i) \cdot x_j^i$

$\therefore \frac{\partial^2 g}{\partial w_k \partial w_j} = \sum_{i=1}^{n} \exp(w^T x^i) \cdot x_j^i \cdot x_k^i$. $\therefore \nabla^2 f = \nabla^2 g = X^T \begin{pmatrix} e^{w^T x^1} & 0 & \dots & 0 \\ 0 & e^{w^T x^2} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & e^{w^T x^n} \end{pmatrix} \cdot X$

- let V be the diagonal matrix of size $n \times n$, with $V_{ii} = e^{w^T x^i}$. Similar to question 1, since each $V_{ii} = e^{w^T x^i} > 0$, $\nabla^2 f = X^T V X$ is positive semi-definite $\therefore f$ is convex

3. $f(w) = \max_{j \in \{1 \dots d\}} L_j |w_j|$

- for $v, w \in \mathbb{R}^d$, we examine $f((1-t)v + tw)$ for $t \in [0,1]$
- $v = (v_1, \dots, v_d)$, $w = (w_1, \dots, w_d)$
- $\forall j \in \{1 \dots d\}$,

$L_j |(1-t)v_j + t w_j|$

$\leq L_j [(1-t)|v_j| + t|w_j|]$ (triangle inequality + property of norm)

$= L_j (1-t)|v_j| + t \cdot L_j |w_j|$

$\therefore \max_{j \in \{1 \dots d\}} L_j |(1-t)v_j + t w_j| \leq \max_{j \in \{1 \dots d\}} L_j (1-t)|v_j| + t \cdot L_j |w_j|$

$\leq (1-t) \max_{j \in \{1 \dots d\}} L_j |v_j| + t \cdot \max_{k \in \{1 \dots d\}} L_k |w_k|$

$\therefore f((1-t)v + tw) \leq (1-t) f(v) + t f(w)$ $\forall t \in [0,1]$ $\therefore f$ is convex

4. $f(w) = \|Xw-y\|_p + \lambda\|Aw\|_q$.

- function $\|\cdot\|_p$ ~~are~~ (norm) are convex for $1 \le p \le \infty$. $Xw-y$, Aw are affine maps
- convex function combine with affine map is convex : $\|Xw-y\|_p$, & $\|Aw\|_q$ are convex
- multiplication by positive constant ~~is~~ keeps convexity : $\lambda\|Aw\|_q$ is convex
- ~~a~~ sum of convex functions are convex $\therefore f(w) = \|Xw-y\|_p + \lambda\|Aw\|_q$ is convex

5. $f(w) = \sum_{i=1}^{n} \max\{0, |w^T x_i - y_i| - \varepsilon\} + \frac{\lambda}{2}\|w\|_2^2$,

- norm squared is convex, and multiplication of positive constant is convex $\therefore \frac{\lambda}{2}\|w\|_2^2$ is convex
- $w^T x_i - y_i$ is ~~affine~~, and absolute value is convex. Composition of convex function with ~~linear~~ affine map is convex $\therefore |w^T x_i - y_i|$ is convex
- subtracting a constant (ε) remains convex, and 0 is convex, maximum of 2 functions is convex $\therefore \max\{0, |w^T x_i - y_i| - \varepsilon\}$ is convex
- sum of convex function is convex $\therefore f(w) = \sum_{i=1}^{n} \max\{0, |w^T x_i - y_i| - \varepsilon\} + \frac{\lambda}{2}\|w\|_2^2$ is convex

6. note that $C = \{w \in \mathbb{R}^d \mid Aw \le b\}$ is a convex set. We use the convention that $\infty \cdot 0 = 0$.

- We show that f is convex by cases. Let $v, w \in \mathbb{R}^d$

Case 1: $v, w \in C$,

$\therefore \forall t \in [0,1]$, $(1-t)v + tw \in C$

$\therefore f((1-t)v+tw) = 0 \le (1-t)\cdot 0 + t\cdot 0 = (1-t)f(v) + tf(w)$

Case 2: one of v, w is in C^c. Suppose, wlog, $v \in C$, $w \in C^c$.

- $\forall t \in (0,1)$ $f((1-t)v+tw) \le \infty = (1-t)f(v) + tf(w)$ $(\because f(w) = \infty,\ t>0 \Rightarrow tf(w) = \infty)$.

Case 3: $v, w \in C^c$

$\therefore \forall t \in (0,1)$, $f((1-t)v+tw) \le \infty = (1-t)f(v) + tf(w)$ $(f(v) = f(w) = \infty)$

$\therefore f$ is convex.

1.2 Convergence of Gradient Descent

1. $\min_{k\in\{1,\dots,t\}} \|\nabla f(w^k)\|^2 \le \frac{2L(f(w^0)-f^*)}{t^{\frac{4}{3}}} \le \varepsilon$

$\therefore \frac{2L(f(w^0)-f^*)}{\varepsilon} \le t^{\frac{4}{3}} \quad \therefore \frac{[2L(f(w^0)-f^*)]^{\frac{3}{4}}}{\varepsilon^{\frac{3}{4}}} \le t, \quad \boxed{t = O\left(\frac{1}{\varepsilon^{\frac{3}{4}}}\right) \text{ iterations}}$

2. Note that by the algorithm, $L^0 \le L^1 \le \dots \le L^k$

- we claim that $L^k \le 2L$.
- suppose, for a contradiction, that $L^k > 2L$. Then $\exists$ a minimum $t \in \{1,\dots,k\}$, $L^t > 2L$
  $\therefore L^{t-1} = \frac{L^t}{2^i}$ for some $i \in \mathbb{N}$, and $L^{t-1} \le 2L$.

There are two cases:

Case 1: $i = 1$. Then $L^{t-1} = \frac{L^t}{2} > L$. But f is L-Lipschitz $\Rightarrow L^{t-1}$-Lipschitz,
$\forall w \in \mathbb{R}^d$, $f\left(w - \frac{1}{L^{t-1}}\nabla f(w)\right) \le f(w) - \frac{1}{2L^{t-1}}\|\nabla f(w)\|^2$ (*)
$\therefore$ there is no need to double L^{t-1} for w^t $\therefore L^t = L^{t-1}$, contradiction.

Case 2: $i \ge 2$.

- if $L \le L^{t-1} \le 2L$, then similar as above, there is no need to double L^{t-1} for w^t $\therefore L^t = L^{t-1}$, contradiction,
- if $L^{t-1} < L$, then we may need to double L^{t-1} multiply times (to L^t) for the inequality to hold for w^t. But $2^{i-1}L^{t-1} = \frac{L^t}{2} > L$, and f is $2^{i-1}L^{t-1}$-Lipschitz thus the algorithm will produce an $L^t \le 2^{i-1}L^{t-1}$, but $L^t = 2^i \cdot L^{t-1}$, contradiction.

- $\forall w \in \mathbb{R}^d$, $\exists \mu > 0$, $\frac{1}{2}\|\nabla f(w)\|^2 \ge \mu(f(w) - f^*(w^*))$

$\forall k \in \mathbb{N}$: $f(w^k) \le f(w^{k-1}) - \frac{1}{2L^{k-1}}\|\nabla f(w^{k-1})\|^2 \le f(w^{k-1}) - \frac{\mu}{L^{k-1}}(f(w^{k-1}) - f(w^*))$

$\therefore f(w^k) \le \left(1 - \frac{\mu}{L^{k-1}}\right) f(w^{k-1}) + \frac{\mu}{L^{k-1}} f(w^*) \Rightarrow f(w^k) - f(w^*) \le \left(1 - \frac{\mu}{L^{k-1}}\right)(f(w^{k-1}) - f(w^*))$

$$\therefore f(w^k) - f(w^*) \le \left[\prod_{j=0}^{k-1}\left(1 - \frac{\mu}{L^j}\right)\right] \cdot (f(w^0) - f(w^*))$$

$\because \forall j \in \{0,1,\dots,k-1\}$, $L^j \le L^k \le 2L$ $\therefore \frac{\mu}{L^j} \ge \frac{\mu}{2L}$, $\therefore 1 - \frac{\mu}{L^j} \le 1 - \frac{\mu}{2L}$.

$$\therefore f(w^k) - f(w^*) \le \left[\prod_{j=0}^{k-1}\left(1 - \frac{\mu}{2L}\right)\right](f(w^0) - f(w^*))$$

$$= \left(1 - \frac{\mu}{2L}\right)^k (f(w^0) - f(w^*)).$$

3. $L^k = \frac{1}{2}L^{k-1}$.

- Let $C^k = \{w \in \mathbb{R}^d \mid f(w) - \leq f(w^{k-1})\}$. Then if f restricted to the set C^k is Lipschitz continuous with a smaller L' (as compared to L where $f: \mathbb{R}^d \to \mathbb{R}$), then decreasing L^k will increase the step size.
- This is effective when the ~~stoc~~ gradient descend algorithm is approaching the global minimum of the function, and we want to increase the step size for the algorithm to converge to the ~~min~~global minimum faster.

4. $f(v) = f(w) + \nabla f(w)^T(v-w) + \frac{1}{2}(v-w)^T \nabla^2 f(u)(v-w)$ for some u between v and w

- by strong convexity, $d^T \nabla^2 f(u) d \geq \mu \|d\|^2$

$\therefore f(v) \geq f(w) + \nabla f(w)^T(v-w) + \frac{\mu}{2}\|v-w\|^2$

- let $w = w^*$ $\therefore \nabla f(w^*) = 0$

$\therefore f(v) \geq f(w^*) + \frac{\mu}{2}\|v-w^*\|^2$ (*)

- in class we have shown that $f(w^k) - f(w^*) \leq (1-\frac{\mu}{L})^k (f(w^0) - f(w^*))$, $\rho = 1 - \frac{\mu}{L}$
- from (*), we have that $f(w^k) - f(w^*) \geq \frac{\mu}{2}\|w^k - w^*\|^2$

$\therefore \frac{\mu}{2}\|w^k - w^*\|^2 \leq (1-\frac{\mu}{L})^k (f(w^0) - f(w^*))$

$$\|w^k - w^*\| \leq \left(\frac{2}{\mu}\right)^{\frac{1}{2}} (1-\tfrac{\mu}{L})^{\frac{k}{2}} (f(w^0) - f(w^*))^{\frac{1}{2}}$$

- thus $\|w^k - w^*\|$ has a convergence rate of $O((1-\frac{\mu}{L})^{\frac{k}{2}}) = O(\rho^{\frac{k}{2}})$.

1.3 Beyond Gradient Descent

1. $\frac{1}{2}\|v - w^{k+\frac{1}{2}}\|^2 + \alpha_k r(v)$

$= \frac{1}{2}\|v - w^k + \alpha_k \nabla f(w^k)\|^2 + \alpha_k r(v)$

$= \frac{1}{2}(v^T - w^{kT} + \alpha_k \nabla f(w^k)^T)(v - w^k + \alpha_k \nabla f(w^k)) + \alpha_k r(v)$

$= \frac{1}{2}(v^T - w^{kT})(v - w^k) + \alpha_k \nabla f(w^k)^T (v - w^k) + \frac{1}{2}\alpha_k^2 \|\nabla f(w^k)\|^2 + \alpha_k r(v)$

$= \alpha_k \left(\frac{1}{2\alpha_k}\|v - w^k\|^2 + \nabla f(w^k)^T (v - w^k) + r(v)\right) + \frac{1}{2}\alpha_k^2 \|\nabla f(w^k)\|^2$

- since multiplying by a constant (α_k) and adding a constant ($\frac{1}{2}\alpha_k^2 \|\nabla f(w^k)\|^2$ does not depend on v) does not change the minimizer,

$$w^{k+1} \in \operatorname*{argmin}_{v \in \mathbb{R}^d} \left\{\frac{1}{2\alpha_k}\|v - w^k\|^2 + \nabla f(w^k)^T (v - w^k) + r(v) + f(w^k)\right\}.$$

2. $f(W) = \sum_{i=1}^{n} \sum_{c \neq y^i} [1 - w_{y^i}^T x^i + w_c^T x^i]^+ + \frac{\lambda}{2}\|W\|_F^2$.

Notation: w_e^t: t^{th} weight for class e ($t \in \mathbb{N}$, $1 \leq t \leq d$).

- $1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$, $\delta_{c,e} = \begin{cases} 1 & \text{if } c = e \\ 0 & \text{if } c \neq e \end{cases}$
- $\theta_{i,c}^{e,t} \in [0,1]$ used to describe subgradient when normal gradient does not exist.

$$\therefore \frac{\partial f}{\partial w_e^t} = \sum_{i=1}^{n} \sum_{c \neq y^i} \Big[1_{\{z>0\}}(1 - w_{y^i}^T x^i + w_c^T x^i) \cdot (-x_t^i \cdot \delta_{e,y^i} + x_t^i \delta_{e,c}) + 1_{\{z=0\}}(1 - w_{y^i}^T x^i + w_c^T x^i) \cdot \theta_{i,c}^{e,t} \cdot (-x_t^i \cdot \delta_{e,y^i} + x_t^i \delta_{e,c}) \Big] + \lambda w_e^t.$$

3. $f(v) \le f(u) + \nabla f(u)^T (v-u) + \frac{L_\infty}{2}\|v-u\|_\infty^2$.

$\therefore f(w^{k+1}) \le f(w^k) - \nabla f(w^k)^T \cdot \frac{\|\nabla f(w^k)\|_1}{L_\infty} \cdot \text{sign}(\nabla f(w^k)) + \frac{L_\infty}{2} \| \frac{\|\nabla f(w^k)\|_1}{L_\infty} \cdot \text{sign}(\nabla f(w^k)) \|_\infty^2$

$= f(w^k) - \frac{\|\nabla f(w^k)\|_1}{L_\infty} \nabla f(w^k)^T \cdot \text{sign}(\nabla f(w^k)) + \frac{\|\nabla f(w^k)\|^2}{2L_\infty} \underbrace{\| \text{sign}(\nabla f(w^k)) \|_\infty^2}_{\le 1}$.

$\le f(w^k) - \frac{\|\nabla f(w^k)\|_1}{L_\infty} \|\nabla f(w^k)\|_1 + \frac{\|\nabla f(w^k)\|^2}{2L_\infty} \cdot 1$

$= f(w^k) - \frac{\|\nabla f(w^k)\|_1^2}{2L_\infty}$.

$\therefore \|\nabla f(w^k)\|_1^2 \le 2L_\infty (f(w^k) - f(w^{k+1}))$

$\therefore t \cdot \min_{k=0,\dots,t-1} \|\nabla f(w^k)\|_1^2 \le 2L_\infty (f(w^0) - f(w^t))$

If $\|\nabla f(w^k)\|_1^2 \le \varepsilon$ to be guaranteed, then $\frac{2L_\infty (f(w^0) - f(w^t))}{t} \le \varepsilon$.

$\therefore \frac{2L_\infty (f(w^0) - f(w^t))}{\varepsilon} \le t$ $\therefore \|\nabla f(w^k)\|_1^2 \le \varepsilon$ after $O(\frac{1}{\varepsilon})$ iterations.