1. Subgradients and Proximal Operators

a.i.

Special Case: $\partial f(x)$ is a singleton, it's obviously convex and closed.

General Case: (Stupied)

Convexity

let $g_1, g_2 \in \partial f(x)$ and $t \in [0,1]$ we have

$f(y) \geq f(x) + g_1^T (y-x)$, $t f(y) \geq t f(x) + t g_1^T (y-x)$

$f(y) \geq f(x) + g_2^T (y-x)$, $(1-t) f(y) \geq (1-t) f(x) + (1-t) g_2^T (y-x)$

add them up we have

$f(y) \geq f(x) + (t g_1 + (1-t) g_2)^T (y-x)$

which means $t g_1 + (1-t) g_2 \in \partial f(x)$.

$D_y = \{ g : f(y) \geq f(x) + g_1^T (y-x) \}$ is a closed half space,

$\partial f(x) = \bigcap_y D_y$, so $\partial f(x)$ is closed and convex.

a.ii.

when $x \neq 0$, it's obvious

Property of subdiffendal of norm

when $x = 0$, $\partial \|x\|_2 = \underset{\|z\|_2 \leq 1}{\text{argmax}}\ 0 = \{ z : \|z\|_2 \leq 1 \}$

a.iii.

let $\bar{y} = \frac{y}{\|y\|_q}$, $\|\bar{y}\|_q \le 1$

$$\|x\|_p = \max_{\|z\|_q \le 1} z^T x \ge \bar{y}^T x = x^T y / \|y\|_q$$

so

$$x^T y \le \|x\|_p \|y\|_q$$

a.iv.

let $z \in \partial \|x\|_p$ be given we have

$$\|y\|_p \ge \|x\|_p + z^T (y - x) \quad \text{for } \forall y$$

Taking $y = 0$ and $y = 2x$, we find

$$\begin{cases} 0 \ge \|x\|_p - z^T x \\ 2\|x\|_p \ge \|x\|_p + z^T x \end{cases} \Rightarrow \|x\|_p = z^T x$$

and because $\|x\|_p = \max_{\|z\|_q \le 1} z^T x$

$$\text{So } \partial \|x\|_p = \operatorname{argmax}_{\|z\|_q \le 1} z^T x \quad \checkmark$$

$$\|x\|_p = z^T x \le \|z\|_q \|x\|_p, \quad \|z\|_q \ge 1 \Rightarrow \|z\|_q = 1 \ ?$$

b.i.

$$\frac{\partial h(z)}{\partial z} = Az + b$$

$$z - x + tAz + tb = 0$$

$$\text{prox}_{h,t}(x) = (I + tA)^{-1}(x - tb)$$

b.ii

$$\left(\frac{\partial h(z)}{\partial z}\right)_i = -\frac{1}{z_i}$$

$$z_i - x_i - \frac{t}{z_i} = 0$$

$$z_i^2 - x_i z_i - t = 0$$

$$\left(\text{prox}_{h,t}(x)\right)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}$$

b.iii

$$\frac{\partial h(z)}{\partial z} = \frac{z}{\|z\|_2}$$

$$z - x + t\frac{z}{\|z\|_2} = 0$$

$$\text{prox}_{h,t}(x) = z_\star$$

b.iv.

$$\left(\text{prox}_{h,t}(x)\right)_i = \begin{cases} x_i, & \frac{1}{2}x_i^2 < t \\ 0, & \frac{1}{2}x_i^2 \geq t \end{cases}$$

2. Properties of Proximal Mappings and Subgradients

a.

let $v \in \partial f_i(x)$ we have $f_i(y) \geq f_i(x) + v^T(y-x)$ for $\forall y$

because $f_i(x) = f(x)$ and $f_i(y) \leq f(y)$ we have

$f(y) \geq f(x) + v^T(y-x)$ for $\forall y \Rightarrow v \in \partial f(x)$

$\partial f_i(x) \subseteq \partial f(x)$ for $i = 1 \cdots n$

So $\bigcup\limits_{i: f_i(x) = f(x)} \partial f_i(x) \subseteq \partial f(x)$

and because 1. $\partial f(x)$ is convex

2. $\mathrm{conv}(C)$ is the smallest convex set that includes C

$$\partial f(x) \supseteq \mathrm{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

b. fix x : $f(z) \geq f(x) + u^T(z-x)$ for $\forall z$

fix y : $f(z) \geq f(y) + v^T(z-y)$ for $\forall z$

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

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

add them up we got

$$(x-y)^T(u-v) \geq 0$$

c.

because $h(x,y) \in \partial_x g(x,y)$ we have
$g(z,y) \geq g(x,y) + \langle h(x,y), z-x \rangle$ for $\forall z$

$\int g(z,y)p(y)dy \geq \int g(x,y)p(y)dy + \int \langle \circ \rangle p(y)dy$

$f(z) \geq f(x) + \langle \int h(x,y)p(y)dy , z-x \rangle$

So $\int h(x,y)p(y)dy \in \partial f(x)$

d.

Subgradient optimality condition

$-\frac{1}{t}(x-u) + \partial h(u) = 0$

$h(y) \geq h(u) + \partial h(u)^T(y-u)$ $\forall y$

$h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u)$ $\forall y$

e. let $u_x = \text{prox}_t(x)$, $u_y = \text{prox}_t(y)$ then

$\partial h(u_x) = \frac{1}{t}(x-u_x)$ $\partial h(u_y) = \frac{1}{t}(y-u_y)$

because

$(u_x-u_y)^T(\partial h(u_x) - \partial h(u_y)) \geq 0$ (Question b)

$(u_x-u_y)^T(x-y-(u_x-u_y)) \geq 0$

$$\|u_x - u_y\|_2^2 \le (u_x - u_y)^T (x-y) \le \|u_x - u_y\|_2 \|x-y\|_2$$

so

$$\|u_x - u_y\|_2 \le \|x-y\|_2$$

3. Convergence Rate for Proximal Gradient Descent

a.

Because Q2 part (d)

$$h(y) \geq h(x^i) + \frac{1}{t}(x^{i-1} - t_i \nabla g(x^{i-1}) - x^i)^T (y - x^i)$$

$$= h(x^i) + s^T (y - x^i) \quad \forall y$$

So s is a subgrdient of h at x^i

b.

We want to know that

$$f(x^i) \leq f(z) + G_t(x^{i-1})^T (x^i - z + t G_t(x^{i-1})) - \frac{t}{2} \|G_t(x^{i-1})\|_2^2$$

$$= f(z) + G_t(x^{i-1})^T (x^i - z) + \frac{t}{2} \|G_t(x^{i-1})\|_2^2$$

in part a, we have

$$h(x^i) \leq h(z) + \left(G_t(x^{i-1}) - \nabla g(x^{i-1})\right)(x^i - z)$$

and we are given that

$$g(x^i) \leq g(x^{i-1}) - t \nabla g(x^{i-1})^T G_t(x^{i-1}) + \frac{t}{2} \|G_t(x^{i-1})\|_2^2$$

$$\leq g(z) + \nabla g(x^{i-1})^T (x^i - z) + \frac{t}{2} \|G_t(x^{i-1})\|_2^2$$

$$g(z) \geq \nabla g(x^{i-1})^T (z - x^{i-1}) + g(x^{i-1})$$

add them up we have

$$f(x^i) \leq f(z) + G_t(x^{i-1})^T(x^i - z) + \frac{t}{2}\|G_t(x^{i-1})\|_2^2$$

that's it.

c. just let z in part b equals x^{i-1}, you will see it.

d. let z in part b equals to x^*

$$f(x^i) - f(x^*) \leq \frac{1}{t}(x^{i-1} - x^i)^T(x^{i-1} - x^*) - \frac{1}{2t}\|x^{i-1} - x^i\|_2^2$$

$$= \frac{1}{2t}(\|x^{i-1}\|_2^2 - 2\langle x^{i-1}, x^*\rangle - 2\langle x^i, x^*\rangle - \|x^i\|_2^2)$$

$$= \frac{1}{2t}(\|x^{i-1} - x^*\|_2^2 - \|x^i - x^*\|_2^2)$$

e. $\sum_{i=1}^{K}\left(f(x^i) - f(x^*)\right) = \frac{1}{2t}(\|x^0 - x^*\|_2^2 - \|x^K - x^*\|_2^2)$

in part c, we know that $\{f(x^{(i)})\}$ is nonincreasing

which means $\sum_{i=1}^{K} f(x^i) \leq K f(x^k)$

so $f(x^K) - f(x^*) \leq \frac{1}{2Kt}(\|x^0 - x^*\|_2^2 - \|x^K - x^*\|_2^2)$

$$\leq \frac{1}{2Kt}\|x^0 - x^*\|_2^2$$

d. ?

4. Proximal Gradient Descent for Group Lasso

a.

$$\mathrm{prox}_{h,t}(x) = \underset{z}{\mathrm{argmin}} \ \frac{1}{2}\|z - x\|_2^2 + t\lambda \sum_{j=1}^{J} w_j \|z_{(j)}\|_2$$

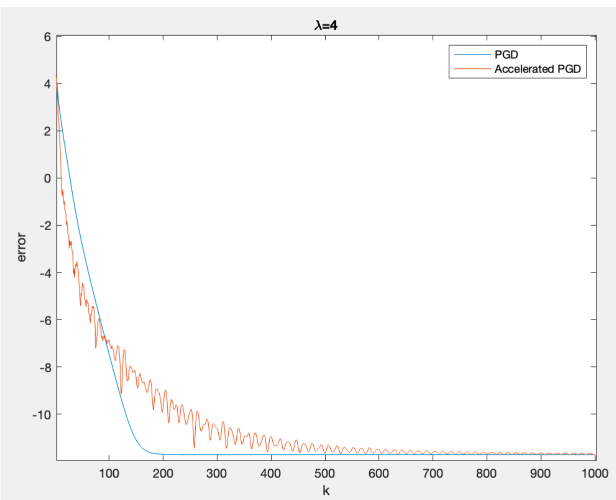
$$\mathrm{prox}_{h,t}(x) = (z_0, z_{(1)} \cdots z_{(J)})$$

$$z_0 = x_0 \ , \quad z_i = \left(1 - \frac{t\lambda w_j}{\|x_{(j)}\|_2}\right)_+ x_i \qquad , \ (x)_+ = \max\{0, x\}$$

$i = 1, \cdots, p \qquad j =$ correspond group of i

b. (i) $$\frac{\partial g}{\partial \beta} = -2X^T(y - X\beta)$$

(ii)



(iii)

Columns 1 through 9

3.0183	0	0	0	0	0	0	0.3034	−0.0542
3.0183	0	0	0	0	0	0	0.3034	−0.0542

Columns 10 through 17

−0.2888	−0.2250	0.0343	−0.2748	−0.4581	0.0230	0.0026	−0.0092
−0.2888	−0.2251	0.0343	−0.2747	−0.4581	0.0230	0.0026	−0.0092

row1: PGD

row2: Accelerated PGD

(iv)

Columns 1 through 9

3.0414	−0.0065	1.2759	0.7411	1.5328	0.0526	1.1450	0.3028	−0.1299
3.0422	−0.0301	1.3128	0.7611	1.6093	0.0488	1.1799	0.3007	−0.1351

Columns 10 through 17

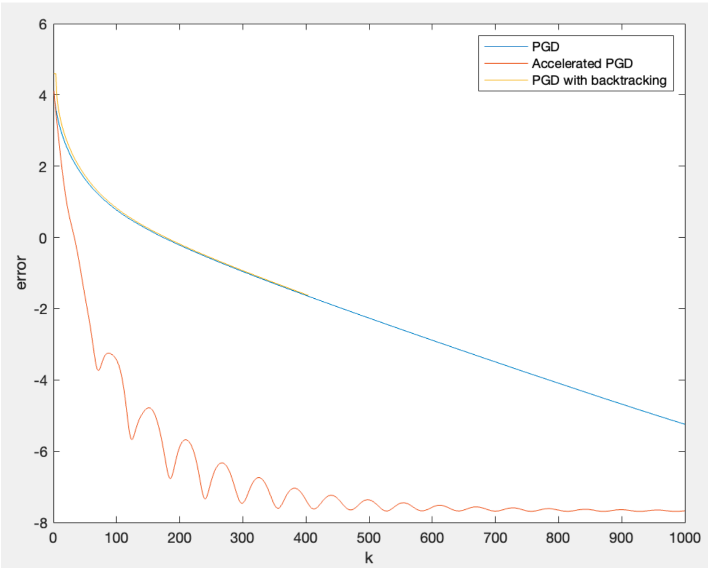
−0.2865	−0.2974	0.1823	−0.5285	−0.4820	0.0871	0.0211	−0.1368
−0.2851	−0.2956	0.1865	−0.5332	−0.4802	0.0878	0.0217	−0.1383

c.

(i) $$\frac{\partial g}{\partial \beta} = -X^T y + \sum_{i=1}^{n} \frac{\exp\{(X\beta)_i\}}{1+\exp\{(X\beta)_i\}} X_i$$

in which X_i^T is the first row of X.

(ii)



(iii) prediction accuracy: 77.46%

have a look at solution, movie genre corresponding to 0 element is not important.