Q3 :

$$f(x) = E[F(x;\xi)]$$

$$E[g(x;\xi)] \in \partial f(x)$$

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t g(x_t;\xi_t)) \quad t \in \mathbb{N}$$

(1)

$$\frac{1}{2} E_{\xi_1 \dots \xi_t}\left[\|x_t - x^*\|_2^2 + \eta_t^2 \|g(x_t;\xi_t)\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right] \quad (1)$$

$$\geq \frac{1}{2} E_{\xi_1 \dots \xi_t}\left[\|x_t - x^*\|_2^2 + \eta_t^2 \|g(x_t;\xi_t)\|_2^2 - \|x_t - \eta_t g(x_t;\xi_t) - x^*\|_2^2 \right]$$

$$= \frac{1}{2} E_{\xi_1 \dots \xi_t}\left[2\eta_t g(x_t;\xi_t)^T (x_t - x^*) \right]$$

$$= \eta_t E_{\xi_1 \dots \xi_{t-1}}\left[E_{\xi_t}[g(x_t;\xi_t)]^T (x_t - x^*) \right] \quad (x_t, x^* \text{ independent of } \xi_t)$$

$$\geq \eta_t E_{\xi_1 \dots \xi_{t-1}}\left[f(x_t) - f(x^*) \right] \quad (\text{convexity}, E_{\xi_t}[g(x_t;\xi_t)] \in \partial f(x))$$

$$= \eta_t E_{\xi_1 \dots \xi_{t-1}\xi_t}\left[f(x_t) - f(x^*) \right] \quad (x_t, x^* \text{ independent of } \xi_t)$$

Note that

$$\frac{1}{2} E_{\xi_1 \dots \xi_t}\left[\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right] + \frac{\eta_t^2}{2} E_{\xi_1 \dots \xi_t}\left[\|g(x_t;\xi_t)\|_2^2 \right]$$

$$= \frac{1}{2} E_{\xi_1 \dots \xi_t}\left[\|x_t - x^*\|_2^2 \right] - \frac{1}{2} E_{\xi_1 \dots \xi_t}\left[\|x_{t+1} - x^*\|_2^2 \right] + \frac{\eta_t^2}{2} L^2 \quad (\text{by assumption})$$

#

(2)

$$E\left[f(\bar{x}_{t_1:t_2}) - f(x^*)\right] \le \frac{1}{\sum_{\tau=t_1}^{t_2}\eta_\tau} \sum_{\tau=t_1}^{t_2}\eta_\tau E\left[f(x_t) - f(x^*)\right] \qquad \text{(Jensen inequality)}$$

$$\le \frac{\sum_{\tau=t_1}^{t_2}\left(E\left[\frac{1}{2}\|x_\tau - x^*\|_2^2\right] - E\left[\frac{1}{2}\|x_{\tau+1} - x^*\|_2^2\right] + \frac{1}{2}\eta_\tau^2 L^2\right)}{\sum_{\tau=t_1}^{t_2}\eta_\tau}$$

$$\le \frac{E\left[\frac{1}{2}\|x_{t_1} - x^*\|_2^2\right] - E\left[\frac{1}{2}\|x_{t_2} - x^*\|_2^2\right] + \frac{L^2}{2}\sum_{\tau=t_1}^{t_2}\eta_\tau^2}{\sum_{\tau=t_1}^{t_2}\eta_\tau}$$

$$\le \frac{E\left[\|x_{t_1} - x^*\|_2^2 - \|x_{t_2} - x^*\|_2^2\right] + L^2\sum_{\tau=t_1}^{t_2}\eta_\tau^2}{2\sum_{\tau=t_1}^{t_2}\eta_\tau}$$

$$\|x_{t_1} - x^*\|_2 \le \|x_{t_1} - x_1\|_2 + \|x^* - x_1\|_2$$

$$\le 2R$$

$$\le \frac{4R^2 + L^2\sum_{\tau=t_1}^{t_2}\eta_\tau^2}{2\sum_{\tau=t_1}^{t_2}\eta_\tau} \qquad \forall\, 1 \le t_1 \le t_2 \ \#$$

(3) notice that $\eta_t = R/L\sqrt{t}$

$$\frac{4R^2 + L^2\sum_{\tau=t_1}^{t_2}\eta_\tau^2}{2\sum_{\tau=t_1}^{t_2}\eta_\tau} = \frac{4R^2 + L^2\times\frac{R^2}{L^2}\sum_{\tau=t_1}^{t_2}\frac{1}{\tau}}{2\frac{R}{L}\sum_{\tau=t_1}^{t_2}\frac{1}{\sqrt{\tau}}} = \frac{RL\left(2 + \frac{1}{2}\sum_{\tau=t_1}^{t_2}\frac{1}{\tau}\right)}{\sum_{\tau=t_1}^{t_2}\frac{1}{\sqrt{\tau}}}$$

$$\le 1 \times t_2$$

$$\le \frac{2RL}{\frac{t_2 - t_1 + 1}{\sqrt{t_2}}} + \frac{RL}{2}\cdot\frac{\sum_{\tau=t_1}^{t_2}\frac{1}{\tau}}{\sum_{\tau=t_1}^{t_2}\frac{\sqrt{t_1}}{\sqrt{\tau}}} \times \frac{\frac{1}{t_1}}{\frac{1}{\sqrt{t_1}}}$$

$$\le \frac{RL}{\frac{t_2 - t_1 + 1}{\sqrt{t_2}}} + \frac{RL}{2}\frac{\sqrt{t_1}}{t_1} = \frac{RL}{\sqrt{t_2}}\left(2\frac{t_2}{t_2 - t_1 + 1} + \sqrt{\frac{t_2}{t_1}}\right) \ \#$$

Q1 :

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix} \qquad y \approx X\hat{\beta}$$

$$\text{Lasso} = \beta^* \in \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\tfrac{1}{2}\|y - X\beta\|_2^2}_{f(\beta)} \;\middle|\; \underbrace{\beta \in \mathbb{R}^p, \|\beta\|_1 \le C}_{\Omega} \right\}$$

$$\nabla f(\beta) = -(y - X\beta)^T X$$

Consider Frank-wolf , $\beta_0 = 0 \in \mathbb{R}^p$

$$v_0 = \underset{v \in \Omega}{\operatorname{argmin}} \{ \langle -X^T y , v \rangle \mid v \in \Omega \}$$

$$= C \operatorname{sign}(x^T y_{i_0}) e_{i_0}$$

where i_0 s.t $|x^T y^{(i_0)}| = \|X^T y\|_\infty$

$$= \underset{j}{\operatorname{argmax}} \{ |\langle y, x_j \rangle| \mid j \in \{1, 2 \cdots p\} \}$$

Remark : $i_0 = j_0$. from forward stagewise regression and , we always choose smallest index to break the tie.

For forward stagewise regression we use $\tilde{\gamma}_t$ instead of γ_t for Frank-wolfe, where $\gamma_t = \frac{1}{C}\tilde{\gamma}_t \; \forall t$

(Frank-Wolfe) $\beta_1 = \gamma_t v_0 = \gamma_t C \operatorname{sign}(X^T y^{(i_0)}) e_{i_0} = \tilde{\beta}_1$ (stagewise regression)

$$r_1 = y - \tilde{\gamma}_t [\operatorname{sign}(\langle y, x_{j_0} \rangle) x_{j_0}]$$

$$= y - c\gamma_t \operatorname{sign}(\langle y, x_{j_0} \rangle) x_{j_0} = y - X\beta_1$$

Mathematical induction:

$(t=0)$ verified above

Suppose that $\forall\ t \in \mathbb{N},\ t \le n$, we have

$$i_t = j_t,\ \beta_{t+1} = \tilde{\beta}_{t+1} \text{ and } r_{t+1} = y - X\beta_{t+1},$$

Consider $t = n+1$

$$U_{n+1} = \underset{v}{\operatorname{argmin}} \{\langle -X^T(y - X\beta_{n+1}), v \rangle \mid v \in \Omega\}$$

$$= \underset{v}{\operatorname{argmin}} \{\langle -X^T(r_{n+1}), v \rangle \mid v \in \Omega\}$$

$$= c \operatorname{sign}(r_{n+1}^T x_{i_{n+1}}) e_{i_{n+1}}$$

where i_{n+1} s.t. $|r_{n+1}^T x^{(i_{n+1})}| = \|r_{n+1} X\|_\infty$

that is we can have

$$i_{n+1} = j_{n+1} \in \underset{j}{\operatorname{argmax}} \{|\langle r_{n+1}, x_j \rangle| \mid j \in \{1 \dots p\}\}$$

$$\beta_{n+2} = (1 - \gamma_{n+1})\beta_{n+1} + U_{n+1}\gamma_{n+1}$$

$$= \begin{cases} (1 - \gamma_{n+1})\beta_{n+1}^{(j)} + \gamma c \operatorname{sign}(\langle r_{n+1}, x_{i_{n+1}} \rangle) & \text{if } j = i_{n+1} \\ (1 - \gamma_{n+1})\beta_{n+1}^{(j)} \end{cases}$$

$$= \begin{cases} (1-\frac{\tilde{\zeta}_{n+1}}{C})\tilde{\beta}_{n+1}(j) + \tilde{\zeta}\ \mathrm{sign}(\langle r_{n+1}, x_{j_{n+1}}\rangle) & \text{if } j = j_{n+1} \\ (1-\frac{\tilde{\zeta}_{n+1}}{C})\tilde{\beta}_{n+1}(j) \end{cases}$$

$$= \tilde{\beta}_{n+2}$$

$$r_{n+2} = r_{n+1} - \tilde{\zeta}_{n+1}\left[\mathrm{sign}(\langle r_{n+1}, x_{j_{n+1}}\rangle) x_{j_{n+1}} + \frac{1}{C}(r_{n+1} - y)\right]$$

$$= r_{n+1} - \zeta_{n+1} X v_{n+1} + \zeta_{n+1} X \beta_{n+1}$$

$$= y - X\beta_{n+1} + \zeta_{n+1} X \beta_{n+1} - \zeta_{n+1} X v_{n+1}$$

$$= y - X\left(\zeta_{n+1} v_{n+1} + (1-\zeta_{n+1})\beta_{n+1}\right)$$

$$= y - X\beta_{n+2}$$

The iteration rule of the modified forward stage regression is equivalent to that of the Frank-wolfe algorithm applied to the Lasso problem #.

Q2 : $f^* = \min_{X} \{ f(X) \mid X \in \mathbb{R}^{d \times d}, X \succeq 0, \mathrm{Tr}(X) = 1 \}$, f L-smooth with Frobenius $\to \mathcal{X}$

consider using Frank-wolfe algorithm with

by assumption, the curvature

$$C_f \le L \max_{X,Y} \{ \|X-Y\|_F^2 \mid X, Y \in \mathcal{X} \}$$

$$\|X-Y\|_F^2 = \mathrm{trace}((X-Y)^T(X-Y))$$
$$= \mathrm{trace}(X^TX) - 2\,\mathrm{trace}(X^TY) + \mathrm{trace}(Y^TY)$$

note that: $\mathrm{trace}(X) = \sum_{i=1}^{d} \lambda_d$, where X is psd. $\lambda_d \ge 0$, λ_d eigenvalue of X.

$\mathrm{trace}(X^TX) = \sum_{i=1}^{d} \lambda_d^2 \le 1$ (each $\lambda_d^2 \le \lambda_d$).

$\mathrm{trace}(Y^TY) \le 1$

notice that $\mathrm{tr}(X,Y)$ defines inner-product $\langle X, Y \rangle$,

by cauchy $\mathrm{tr}(X^TY) \le \sqrt{\mathrm{tr}(X^TX)}\sqrt{\mathrm{tr}(Y^TY)} \le 1$.

It shows that $C_f \le 4L$

C_f exists.

Set $X_0 = u_{-1}u_{-1}^T$, where $u_{-1} = e_1 \in \mathbb{R}^d$.

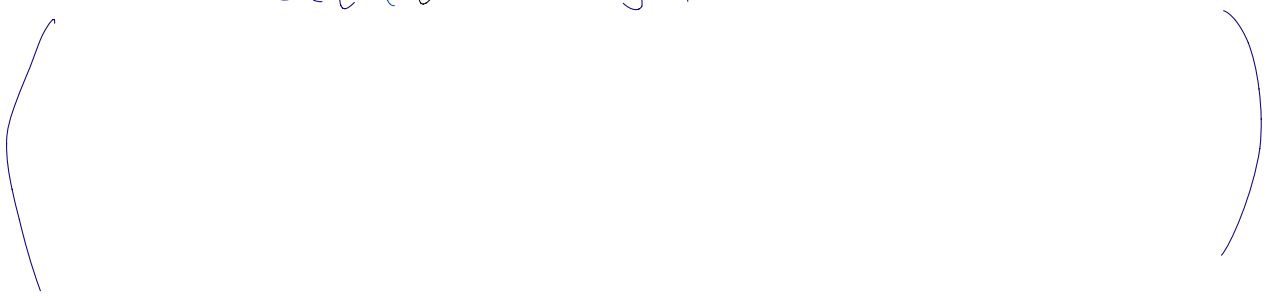
$X_0 \in \mathcal{X}$, $\gamma_t = \frac{2}{t+2}$

using Frank-wolfe:

for $t = 0, 1 \ldots T$ do.

$V_t = \arg\min_{V} \{\langle \nabla f(X_t), V \rangle \mid V \in \mathcal{X}\}$

$= u_t u_t^T$ (by proposition 8-19



$X_{t+1} = (1-\gamma_t) X_t + \gamma_t V_t$

end

by convergence guarantee of Frank-wolfe,

take $T > \frac{2C_f}{\varepsilon} - 2$.

then $f(X_T) - f^* \leq \varepsilon$, X_T is a ε-approximation solution, since $\text{rank}(A+B) \leq \text{rank}(A) + \text{rank}(B)$,

$\text{rank}(X_T) \leq T$, $\text{rank}(X_T) = O(\frac{1}{\varepsilon})$ #

Remark:

We don't have any constraint to force $\nabla f(x)$ P.S.D $\forall x \in \mathcal{X}$,

so it is not true that $V_t = U_t U_t^T$.

Suppose we have the linear minimization oracle, let $\tilde{V}_t$ be the output, $\tilde{V}_t \in \mathcal{X}$, $\text{trace}(\tilde{V}_t) = 1$.

$\tilde{V}_t$ is now P.S.D.

$$\tilde{V}_t = \sum \lambda_k b_k b_k^T \quad \text{(by eigen-decomposition)}$$

then $\langle \nabla f(x_t), \tilde{V}_t \rangle = \langle \nabla f(x_t), b_k b_k^T \rangle$

$\forall k = 1, 2 \ldots \text{rank}(\tilde{V}_t)$

we can choose our V_t as $b_k b_k^T$ for some k #.