In the name of God

Convex optimization project

Ardalan Gerami 99102112

Mahyar Ghazanfari 98102057

Q1.

we know :

$$X^{(t+1)} = X^{(t)} - \eta \nabla f\left(X^{(t)}\right), X^{(1)} = 0$$

and also :

$$\nabla f\left(X^{(t)}\right) = AX^{(t)} - b$$

so we can write →

$$X^{(t+1)} = X^{(t)} - \eta \mathrm{A} X^{(t)} + \eta b = (I - \eta A)X^{(t)} + \eta b$$

now from :

$$X^* = \underset{X}{\operatorname{argmin}} \{f(X)\} \quad \rightarrow \quad AX^* - b = 0$$

$$\rightarrow X^{(t+1)} - X^* = X^{(t)} - \eta\left(\mathrm{A}X^{(t)} - b\right) - X^* = X^{(t)} - \eta\left(\mathrm{A}X^{(t)} - b\right) - X^* + \eta(\mathrm{A}X^* - b)$$

$$\rightarrow X^{(t+1)} - X^* = (I - \eta A)\left(X^{(t)} - X^*\right) = (I - \eta A) \ldots . (I - \eta A)\left(X^{(1)} - X^*\right)$$

$$\rightarrow X^{(t+1)} - X^* = -(I - \eta A)^t X^*, \left\|X^{(t+1)} - X^*\right\| \leq \|(I - \eta A)^t\| \|X^*\|$$

$$\rightarrow \left\|X^{(t+1)} - X^*\right\| \leq \|(I - \eta A)\|^t \|X^*\|$$

on the other hand if we want to have $\lim_{T \to \infty} \|X^T - X^*\| = 0$

we should consider f *as a* M_{smooth} *and* $m_{strongly}$ *convex* :

$$mI \leq \nabla^2 f(X) \leq MI \rightarrow \quad \frac{m}{M} I \leq \frac{A}{M} \leq I$$

also :

$$\left\| I - \frac{A}{\lambda_{\max(A)}} \right\|_2 = 1 - \frac{\lambda_{\min(A)}}{\lambda_{\max(A)}}$$

$$\rightarrow \left\|X^{(t+1)} - X^*\right\| \leq \left(1 - \frac{m}{M}\right)^t \|X^*\| \leq e^{-\frac{m}{M}t} \|X^*\|, \text{now if } t \to \infty : \left\|X^{(t+1)} - X^*\right\| \leq 0$$

so

$$\lim_{T \to \infty} \|X^T - X^*\| = 0$$

Q2.

we want to prove

$$X^{(t+1)} = X^{(t)} - \eta \nabla f\left(X^{(t)}\right) = arg\min_{X}\{ f\left(X^{(t)}\right) + \langle \nabla f\left(X^{(t)}\right), X - X^{(t)} \rangle + \frac{1}{2\eta}\left\|X - X^{(t)}\right\|_2^2\}$$

let's calculate the gradient of

$$h(f) = f\left(X^{(t)}\right) + \langle \nabla f\left(X^{(t)}\right), X - X^{(t)} \rangle + \frac{1}{2\eta}\left\|X - X^{(t)}\right\|_2^2$$

$$\rightarrow \nabla_X h(f) = \nabla_X \left(f\left(X^{(t)}\right) + \langle \nabla f\left(X^{(t)}\right), X - X^{(t)} \rangle + \frac{1}{2\eta}\left\|X - X^{(t)}\right\|_2^2 \right)$$

$$\nabla_X h(f) = \nabla_X \langle \nabla f\left(X^{(t)}\right), X - X^{(t)} \rangle + \nabla_X \frac{1}{2\eta}\left\|X - X^{(t)}\right\|_2^2$$

$$\nabla_X h(f) = \nabla f\left(X^{(t)}\right) + \frac{1}{2\eta}\left(2X - 2X^{(t)}\right) = 0$$

$$\nabla f\left(X^{(t)}\right) + \frac{\left(X - X^{(t)}\right)}{\eta} = 0 \rightarrow X = X^{(t)} - \eta \nabla f\left(X^{(t)}\right) \; and \; X^{(t+1)} = X$$

so

$$X^{(t+1)} = arg\min_{X}\{ f\left(X^{(t)}\right) + \langle \nabla f\left(X^{(t)}\right), X - X^{(t)} \rangle + \frac{1}{2\eta}\left\|X - X^{(t)}\right\|_2^2\}$$

Q3.

$$\left\|X^{(t+1)} - X^*\right\|_2^2 = \left\|X^{(t)} - \eta_t v^{(t)} - X^*\right\|_2^2 \leq \left\|X^{(t)} - X^*\right\|_2^2$$

above all we know:

$$\left\|X^{(t)} - \eta_t v^{(t)} - X^*\right\|_2^2 = \left\|X^{(t)} - X^*\right\|_2^2 + \eta_t^2\left\|v^{(t)}\right\|_2^2 - 2\langle X^{(t)} - X^*, \eta_t v^{(t)} \rangle$$

on the other hand :

$$\left\|X^{(t)} - X^*\right\|_2^2 + \eta_t^2\left\|v^{(t)}\right\|_2^2 - 2\langle X^{(t)} - X^*, \eta_t v^{(t)} \rangle \leq \left\|X^{(t)} - X^*\right\|_2^2 - \eta_t^2\left\|v^{(t)}\right\|_2^2$$

and:

$$\left\|X^{(t)} - X^*\right\|_2^2 - \eta_t^2\left\|v^{(t)}\right\|_2^2 \leq \left\|X^{(t)} - X^*\right\|_2^2$$

so :

$$\left\|X^{(t+1)} - X^*\right\|_2^2 \leq \left\|X^{(t)} - X^*\right\|_2^2$$

Q4.

$$due\ to\ the\ definition\ of\ sub-gradient\ such\ \ v\ \ is\ sub-gradient\ of\ \ f\ \ in\ point\ \ x, y \epsilon \mathbb{R}^n\ \ so:$$

$$f(y) \geq f(x) + \langle v, y - x \rangle$$

$$and\ f\ is\ \rho - lipshitz:$$

$$|f(y) - f(x)| \leq\ \rho \|y - x\|_2$$

$$go\ on:$$

$$|f(y) - f(x)| \geq \|v\|_2 \|y - x\|_2$$

$$so:$$

$$\|v\|_2 \leq\ \rho$$

Q5.

$$we\ know(I):$$

$$\|X^{(t+1)} - X^*\|_2^2 = \|X^{(t)} - \eta_t v^{(t)} - X^*\|_2^2 = \|X^{(t)} - X^*\|_2^2 + \eta_t^2 \|v^{(t)}\|_2^2 - 2\eta_t \langle X^{(t)} - X^*, v^{(t)} \rangle$$

$$on\ the\ other\ hand\ X^* = argmin\ f(x)\ so\ we\ would\ have:$$

$$f(X^{(t)}) \geq f(X^*)$$

$$and:$$

$$|f(X^{(t)}) - f(X^*)| \leq\ \rho \|X^{(t)} - X^*\|_2\ , \|v^{(t)}\|_2 \leq\ \rho$$

$$so:$$

$$f(X^{(t)}) - f(X^*) \leq\ \rho \|X^{(t)} - X^*\|_2 \rightarrow multiply\ 2\eta_t \|v^{(t)}\|_2\ , we\ have:$$

$$2\eta_t \|v^{(t)}\|_2 (f(X^{(t)}) - f(X^*)) \leq 2\eta_t \|v^{(t)}\|_2 \rho \|X^{(t)} - X^*\|_2$$

$$on\ the\ other\ hand\ we\ said\ \|v^{(t)}\|_2 \leq\ \rho\ , so:$$

$$2\eta_t (f(X^{(t)}) - f(X^*)) \leq 2\eta_t \|v^{(t)}\|_2 \|X^{(t)} - X^*\|_2$$

$$now\ we\ try\ to\ build\ (I):$$

$$-2\eta_t \left(f(X^{(t)}) - f(X^*)\right) \geq -2\eta_t \|v^{(t)}\|_2 \|X^{(t)} - X^*\|_2$$

$$\rightarrow$$

$$\|X^{(t)} - X^*\|_2^2 + \eta_t^2 \|v^{(t)}\|_2^2 - 2\eta_t \left(f(X^{(t)}) - f(X^*)\right)$$

$$\geq \|X^{(t)} - X^*\|_2^2 + \eta_t^2 \|v^{(t)}\|_2^2 - 2\eta_t \|v^{(t)}\|_2 \|X^{(t)} - X^*\|_2$$

$$so:$$

$$\|X^{(t+1)} - X^*\|_2^2 \le \|X^{(t)} - X^*\|_2^2 + \eta_t^2\|v^{(t)}\|_2^2 - 2\eta_t\left(f(X^{(t)}) - f(X^*)\right)$$

$$\rightarrow \times \frac{1}{2}:$$

$$\frac{1}{2}\|X^{(t+1)} - X^*\|_2^2 \le \frac{1}{2}\|X^{(t)} - X^*\|_2^2 - \eta_t\left(f(X^{(t)}) - f(X^*)\right) + \frac{\eta_t^2}{2}\|v^{(t)}\|_2^2$$

Q6.

$$first\ consider\ X_0 = 0 = X^{(1)}, also\ We\ consider\ the\ assumptions\ of\ the\ question:$$

$$\sum_{t=1}^{T}\eta_t\left(f(X^{(t)}) - f(X^*)\right) \le \frac{1}{2}B^2 + \frac{1}{2}\rho^2\sum_{t=1}^{T}\eta_t^2$$

$$on\ the\ other\ hand\ \|v^{(t)}\|_2 \le \rho\ and\ \|X^*\|_2^2 \le B^2 and\ from\ part\ 3:$$

$$\frac{1}{2}\|X^{(t+1)} - X^*\|_2^2 \le \frac{1}{2}\|X^{(t)} - X^*\|_2^2 - \eta_t\left(f(X^{(t)}) - f(X^*)\right) + \frac{\eta_t^2}{2}\|v^{(t)}\|_2^2$$

$$\rightarrow \sum_{t=1}^{T}(each\ side\ of\ inequality):$$

$$\frac{1}{2}\sum_{t=1}^{T}\|X^{(t+1)} - X^*\|_2^2 \le \frac{1}{2}\sum_{t=1}^{T}\|X^{(t)} - X^*\|_2^2 - \sum_{t=1}^{T}\eta_t\left(f(X^{(t)}) - f(X^*)\right) + \frac{1}{2}\sum_{t=1}^{T}\eta_t^2\|v^{(t)}\|_2^2$$

$$subtraction:$$

$$\frac{1}{2}\left(\|X^{(T)} - X^*\|_2^2\right) \le -\sum_{t=1}^{T}\eta_t\left(f(X^{(t)}) - f(X^*)\right) + \frac{1}{2}\sum_{t=1}^{T}\eta_t^2\|v^{(t)}\|_2^2$$

$$\rightarrow \|X^{(T)} - X^*\|_2^2 \le \|X^{(1)} - X^*\|_2^2 = \|X^*\|_2^2 \le B^2, \|v^{(t)}\|_2 \le \rho, so:$$

$$\sum_{t=1}^{T}\eta_t\left(f(X^{(t)}) - f(X^*)\right) \le \frac{1}{2}\rho^2\sum_{t=1}^{T}\eta_t^2 + \frac{1}{2}B^2$$

Q7.

$$let's\ divide\ \sum_{t=1}^{T}\eta_t\left(f\left(X^{(t)}\right)-f(X^*)\right)\ by\ \sum_{t=1}^{T}\eta_t^2:$$

$$\frac{\sum_{t=1}^{T}\eta_t\left(f\left(X^{(t)}\right)-f(X^*)\right)}{\sum_{t=1}^{T}\eta_t^2}$$

$$if\ we\ consider\ the\ form\ of\ the\ X_T^{bar}=\frac{\sum_{t=1}^{T}\eta_t X^{(t)}}{\sum_{t=1}^{T}\eta_t}\ we\ conclude:$$

$$\frac{\sum_{t=1}^{T}\eta_t\left(f\left(X^{(t)}\right)-f(X^*)\right)}{\sum_{t=1}^{T}\eta_t^2}=f\left(X_T^{bar}\right)-f(X^*)$$

$$due\ to\ Q6\ and\ divide\ each\ side\ of\ inequality\ by\ \sum_{t=1}^{T}\eta_t^2\ we\ have:$$

$$\frac{\sum_{t=1}^{T}\eta_t\left(f\left(X^{(t)}\right)-f(X^*)\right)}{\sum_{t=1}^{T}\eta_t^2}=f\left(X_T^{bar}\right)-f(X^*)\le\frac{B^2+\rho^2\sum_{t=1}^{T}\eta_t^2}{2\sum_{t=1}^{T}\eta_t}$$

$$so:$$

$$f\left(X_T^{bar}\right)-f(X^*)\le\frac{B^2+\rho^2\sum_{t=1}^{T}\eta_t^2}{2\sum_{t=1}^{T}\eta_t}$$

Q8.

$$***\ we\ should\ consider\ t=0\ to\ T\ ***$$

$$let\ \eta=\eta_t:$$

$$f\left(X_T^{bar}\right)-f(X^*)\le\frac{B^2+(T+1)\rho^2\eta_t^2}{2(T+1)\eta_t}=\frac{1}{2}\left(\frac{B^2}{(T+1)\eta}+\eta\rho^2\right)$$

$$for\ best\ \eta:$$

$$argmin\ \frac{1}{2}\left(\frac{B^2}{(T+1)\eta}+\eta\rho^2\right)\rightarrow\frac{\rho^2}{2}-\frac{B^2}{2(T+1)\eta^2}=0\rightarrow\ \eta_*=\frac{B}{\rho\sqrt{T+1}}$$

$$best\ bound:$$

$$\frac{1}{2}\left(\frac{B^2}{(T+1)\eta}+\eta\rho^2\right),\eta_*=\frac{B}{\rho\sqrt{T+1}}\rightarrow\frac{\rho B}{\sqrt{T+1}}$$

Q9.

با توجه به اینکه در *Sub-Gradient Descent* لزوماً در هر مرحله کاهش انجام نمی‌گیرد و بعضاً نقاطی وجود دارند که افزایش می‌یابند می‌توانیم برای این نقاط از میانگین تمامی نقاط در الگوریتم استفاده کنیم و با توجه به اینکه در سوال بالا کران نهایی بر حسب $f\left(X_T^{bar}\right)$ به دست آمد برای مقدار تابع در میانگین کران بالا وجود دارد ولی برای آن نقطه به روزرسانی شده که افزایشی است پیدا کردن کران بالا مقدور نیست.

Q10.

$$f\left(X_T^{bar}\right) - f(X^*) \leq \frac{B^2 + \rho^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t} \text{ (from question7):}$$

$$\frac{\sum_{t=1}^{T} \eta_t \left(f\left(X^{(t)}\right)\right)}{\sum_{t=1}^{T} \eta_t^2} - f(X^*) \leq \frac{B^2 + \rho^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t}$$

$$on\ the\ other\ hand f\left(X^{(t_T^*)}\right) \leq f\left(X^{(t)}\right)\ so:$$

$$\frac{\sum_{t=1}^{T} \eta_t f\left(X^{(t_T^*)}\right)}{\sum_{t=1}^{T} \eta_t^2} \geq f\left(X^{(t_T^*)}\right)$$

$$so:$$

$$f\left(X^{(t_T^*)}\right) - f(X^*) \leq \frac{B^2 + \rho^2 \sum_{t=0or1}^{T} \eta_t^2}{2 \sum_{t=1or0}^{T} \eta_t}$$

Q11.

$$we\ can\ write\ the\ \Pi_C = \underset{y}{\operatorname{argmin}} \{\|y - x\|_2\}\ \ as\ a\ convex\ optimization\ problem:$$

$$minimize\ \|y - x\|$$

$$subject\ to \quad Ay = b$$

$$solve:$$

$$by\ \ lagrangian\ coefs\ we\ have:$$

$$\mathcal{L}(x, y, v) = x^T x + y^T y - 2x^T y + v^T (Ay - b)$$

$$calculate \frac{\partial \mathcal{L}}{\partial y}:$$

$$\frac{\partial \mathcal{L}}{\partial y} = 2y - 2x + A^T v = 0 \rightarrow y_* = \frac{1}{2}(2x - A^T v) \rightarrow \times A\ , Ay_* = b\ , so:$$

$$Ay_* = Ax - \frac{1}{2}AA^T v \rightarrow \frac{1}{2}AA^T v = Ax - b$$

$$v = 2(AA^T)^{-1}(Ax - b)$$

$$now\ we\ calculate\ y:$$

$$y_* = x - \ 2A^T(AA^T)^{-1}(Ax - b)$$

$$so:$$

$$\Pi_C = \left\| 2A^T\left(AA^T\right)^{-1}(Ax - b) \right\|_2$$

Q12.

$we\ can\ write\ the\ \Pi_C = \underset{y}{\text{argmin}}\ \{\|y - x\|_2\}\ \ as\ a\ convex\ optimization\ problem:$

$$minimize\ \|y - x^2\|$$

$$subject\ to \quad Ay \leq b$$

$$solve:$$

$by\ \ lagrangian\ coefs\ we\ have:$

$$\mathcal{L}(y, \lambda) = \|y - x^2\| + \lambda^T(Ay - b)$$

$$calculate\ \frac{\partial \mathcal{L}}{\partial y}:$$

$$\frac{\partial \mathcal{L}}{\partial y} = 2y - 2x + A^T\lambda = 0$$

$$KKT: \begin{cases} 2y - 2x + A^T\lambda = 0 \\ Ay \leq b \\ \lambda_i\left(A_i^T\lambda - b_i\right) = 0 \end{cases}$$

Q13.

$C\ determine\ the\ set\ of\ inner\ points\ of\ a$ sphere $and\ we\ can\ say:$

The closest point of the set to the desired point is in the line between this point and the center of the sphere (on the sphere). Therefore, y is in the direction of x and at a distance b from the center of the sphere.

خلاصه : y در راستای x در فاصله b از مرکز کره است.

Q14.

due to algorithm 3 we have :

$$X^{(t+1)} = \Pi_C(X^{(t)} - \eta_t v^{(t)})$$

we want to prove that above $X^{(t+1)}$ is same as below eq.:

$$X^{(t+1)} = \underset{X}{\operatorname{argmin}}\{f(X^{(t)}) + \langle v^{(t)}, X - X^{(t)}\rangle + \frac{1}{2\eta_t}\|X - X^{(t)}\|_2\}$$

go on:

$$X^{(t+1)} = \Pi_C(X^{(t)} - \eta_t v^{(t)}) = \underset{y}{\operatorname{argmin}}\|y - X^{(t)} + v^{(t)}\|_2$$

$$= \underset{y}{\operatorname{argmin}}\left(y^T y - 2y^T(X^{(t)} - \eta_t v^{(t)}) + \|X^{(t)} - \eta_t v^{(t)}\|_2^2\right)$$

$$= \underset{y}{\operatorname{argmin}}(\|y\|_2^2 - 2y^T(X^{(t)} - \eta_t v^{(t)}))$$

from eq.:

$$X^{(t+1)} = \underset{X}{\operatorname{argmin}}\{f(X^{(t)}) + \langle v^{(t)}, X - X^{(t)}\rangle + \frac{1}{2\eta_t}\|X - X^{(t)}\|_2\}$$

$$X^{(t+1)} = \underset{X}{\operatorname{argmin}}\{v^{(t)}.x + \frac{1}{2\eta_t}(\|x\|_2^2 - 2x^T x^{(t)})\}$$

$$X^{(t+1)} = \underset{y}{\operatorname{argmin}}(\|x\|_2^2 - 2x^T(X^{(t)} - \eta_t v^{(t)}))$$

as we see Constrained Sub − Gradient Descent method will leads to same answer.

Q15.

$$as\ we\ know\ from\ th.6\ we\ can\ write:$$

$$\frac{1}{2}\left\|X^{(t+1)} - X^*\right\|_2^2 \le \frac{1}{2}\left\|X^{(t)} - X^*\right\|_2^2 - \eta_t\left(f\left(X^{(t)}\right) - f(X^*)\right) + \frac{\eta_t^2}{2}\left\|v^{(t)}\right\|_2^2$$

$$by\ multipling \frac{1}{\eta_t} and \sum_{t=1}^{T} (each\ side\ of\ inequality)$$

$$\rightarrow \frac{1}{2\eta_t}\left(\left\|X^{(T)} - X^*\right\|_2^2\right) \le -\sum_{t=1}^{T}\eta_t\left(f\left(X^{(t)}\right) - f(X^*)\right) + \frac{1}{2}\sum_{t=1}^{T}\eta_t^2\left\|v^{(t)}\right\|_2^2$$

$$\rightarrow$$

$$on\ the\ other\ hand\ \left\|v^{(t)}\right\|_2 \le \rho\ and\ \|X - X^*\|_2^2 \le B^2:$$

$$\sum_{t=1}^{T}\left(f\left(X^{(t)}\right) - f(X^*)\right) \le \frac{1}{2}\rho^2\sum_{t=1}^{T}\eta_t + \frac{1}{2\eta_T}B^2$$

Q16.

$$let\ \eta_t = \frac{\alpha}{\sqrt{t}}\ and\ X_T^{bar} = \frac{1}{T}\sum_{t=1}^{T}X^{(t)}:$$

$$start\ algorithm\ 3\ \ with\ X_0 \epsilon\ C:$$

$$from\ jensen\ inequlity:$$

$$f\left(X_T^{bar}\right) \le \frac{1}{\mathrm{T}}\sum_{t=1}^{T}f\left(X^{(t)}\right)\ \ , f\left(X_T^{bar}\right) - f(X^*) \le \frac{1}{\mathrm{T}}\sum_{t=1}^{T}f\left(X^{(t)}\right) - f(X^*)$$

$$from\ last\ part\ we\ know:$$

$$\frac{1}{\mathrm{T}}\sum_{t=1}^{T}f\left(X^{(t)}\right) - f(X^*)\ \le \frac{1}{2}\rho^2\sum_{t=1}^{T}\eta_t + \frac{1}{2\eta_T}B^2\ \rightarrow$$

$$\sum_{t=1}^{T}f\left(X^{(t)}\right) - f(X^*)\ \le \frac{1}{2T}\rho^2\sum_{t=1}^{T}\eta_t + \frac{1}{2T\eta_T}B^2$$

$$f\left(X_T^{bar}\right) - f(X^*) \le \frac{\rho^2\alpha}{\sqrt{T}} + \frac{1}{2\alpha\sqrt{T}}B^2$$

$$now\ optimal\ \alpha:$$

$$\alpha_* = argmin\ \frac{\rho^2\alpha}{\sqrt{T}} + \frac{1}{2\alpha\sqrt{T}}B^2 \rightarrow\ \ \alpha_* = \frac{B\sqrt{2}}{\rho}$$

Q17.

می‌دانیم که $v^{(t)}$ ناشی از یک توزیع احتمالاتی است و همچنین $X^{(t)}$ در هر لحظه 0 تا T+1 از رابطه

$$X^{(t+1)} = \Pi_C(X^{(t)} - \eta_t v^{(t)})$$

بدست می‌آید بنابراین می‌توان استدلال نمود X برداری تصادفی است و منشا این خاصیت تصادفی همین بیانات بالا است.

Q18.

تا به الان هیچگاه تابع هدف متغیر تصادفی را شامل نمیشد. در این مسئله ما *gradient-discent* را برای متغیرهای تصادفی که توزیع تصادفی دارند تعریف کردیم و باید مجموعه *subgradient* را باید با استفاده از امید ریاضی داده‌ها و بردارها بدست بیاوریم.

Q19.

$$we\ know\ \mathcal{E}_t = V^{(t)} - \mathrm{E}\left[V^{(t)}\middle|X^{(t)}\right]\ and\ we\ want\ to\ prove:$$

$$\frac{1}{2}\left\|X^{(t+1)} - X^*\right\|_2^2 \le \frac{1}{2}\left\|X^{(t)} - X^*\right\|_2^2 - \eta_t\left(f\left(X^{(t)}\right) - f(X^*)\right) + \frac{\eta_t^2}{2}\left\|v^{(t)}\right\|_2^2 - \eta_t\langle\mathcal{E}_t, X^{(t)} - X^*\rangle$$

$$go\ on:$$

$$due\ to\ question\ we\ know\ \mathrm{E}\left[V^{(t)}|X^{(t)}\right]\ \epsilon\ \partial f\left(X^{(t)}\right)\ \ so:$$

$$f(X^*) \ge f\left(X^{(t)}\right) + \langle X^{(t)} - X^*, \mathrm{E}\left[V^{(t)}|X^{(t)}\right]\rangle \rightarrow\ multiply\ \eta_t:$$

$$-\eta_t\left(f\left(X^{(t)}\right) - f(X^*) + \langle X^{(t)} - X^*, \mathrm{E}\left[V^{(t)}|X^{(t)}\right]\rangle\right) > 0$$

$$we\ have\ to\ prove\ equeivalent:$$

$$\frac{1}{2}\left\|X^{(t+1)} - X^*\right\|_2^2 \le \frac{1}{2}\left\|X^{(t)} - \eta_t v^{(t)} - X^*\right\|_2^2\ \ (I)$$

$$as\ we\ know\ X^{(t+1)} = \Pi_C\left(X^{(t)} - \eta_t v^{(t)}\right)\ and\ if\ (I)\ doesn't\ be\ valid\ then\ X^{(t+1)} = X^*$$

$$and\ \frac{1}{2}\left\|X^{(t+1)} - X^*\right\|_2^2 \le \frac{1}{2}\left\|X^{(t)} - \eta_t v^{(t)} - X^*\right\|_2^2 \le 0\ which\ is\ in\ not\ valid\ so:$$

$$\frac{1}{2}\left\|X^{(t+1)} - X^*\right\|_2^2 \le \frac{1}{2}\left\|X^{(t)} - X^*\right\|_2^2 - \eta_t\left(f\left(X^{(t)}\right) - f(X^*)\right) + \frac{\eta_t^2}{2}\left\|v^{(t)}\right\|_2^2 - \eta_t\langle\mathcal{E}_t, X^{(t)} - X^*\rangle$$

Q20.

$$show\ \mathrm{E}\left[\langle \mathcal{E}_t, X^{(t)} - X^* \rangle\right] = 0$$

$$\mathrm{E}\left[\langle \mathcal{E}_t, X^{(t)} - X^* \rangle\right] = \mathrm{E}\left[\langle V^{(t)} - \mathrm{E}\left[V^{(t)} \middle| X^{(t)}\right], X^{(t)} - X^* \rangle\right] = \mathrm{E}\left[\langle V^{(t)}, X^{(t)} - X^* \rangle\right] - \mathrm{E}\left[\langle \mathrm{E}\left[V^{(t)} \middle| X^{(t)}\right], X^{(t)} - X^* \rangle\right]$$

$$\mathrm{E}\left[\langle V^{(t)}, X^{(t)} - X^* \rangle\right] - \mathrm{E}\left[\langle \mathrm{E}\left[V^{(t)} \middle| X^{(t)}\right], X^{(t)} - X^* \rangle\right] = \mathrm{E}\left[\left(X^{(t)} - X^*\right)^T V^{(t)}\right] - \mathrm{E}\left[\left(X^{(t)} - X^*\right)^T \mathrm{E}\left[V^{(t)} \middle| X^{(t)}\right]\right]$$

$$\mathrm{E}\left[\left(X^{(t)} - X^*\right)^T V^{(t)}\right] - \mathrm{E}\left[\left(X^{(t)} - X^*\right)^T \mathrm{E}\left[V^{(t)} \middle| X^{(t)}\right]\right] = \mathrm{E}\left[\left(X^{(t)} - X^*\right)^T V^{(t)}\right] - \mathrm{E}\left[\mathrm{E}\left[\left(X^{(t)} - X^*\right)^T V^{(t)} \middle| X^{(t)}\right]\right]$$

$$so:$$

$$\mathrm{E}\left[\left(X^{(t)} - X^*\right)^T V^{(t)}\right] - \mathrm{E}\left[\left(X^{(t)} - X^*\right)^T V^{(t)}\right] = 0 \rightarrow$$

$$\mathrm{E}\left[\langle \mathcal{E}_t, X^{(t)} - X^* \rangle\right] = 0$$

Q21.

$$\mathrm{E}\left[f(X_T^{bar})\right] - f(X^*) = \mathrm{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} X^{(t)}\right)\right] - f(X^*) \leq \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[f\left(X^{(t)}\right)\right] - f(X^*) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[f\left(X^{(t)}\right) - f(X^*)\right]$$

$$so: \quad \mathrm{E}\left[f(X_T^{bar})\right] - f(X^*) \leq \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[f\left(X^{(t)}\right) - f(X^*)\right]$$

$$from\ question\ 19\ ,20:$$

$$f\left(X^{(t)}\right) - f(X^*) \leq \frac{1}{2\eta_t}\left\|X^{(t)} - X^*\right\|_2^2 - \frac{1}{2\eta_t}\left\|X^{(t+1)} - X^*\right\|_2^2 + \frac{\eta_t}{2}\left\|v^{(t)}\right\|_2^2 - \langle \mathcal{E}_t, X^{(t)} - X^* \rangle$$

$$\mathrm{E}\left[f\left(X^{(t)}\right) - f(X^*)\right] \leq \frac{1}{2\eta_t}\,\mathrm{E}\left[\left\|X^{(t)} - X^*\right\|_2^2\right] - \frac{1}{2\eta_t}\mathrm{E}\left[\left\|X^{(t+1)} - X^*\right\|_2^2\right] + \frac{\eta_t}{2}\mathrm{E}\left[\left\|v^{(t)}\right\|_2^2\right]$$

$$\rightarrow \sum_{t=1}^{T} (each\ side\ of\ inequality):$$

$$\rightarrow \sum_{t=1}^{T} \mathrm{E}\left[f\left(X^{(t)}\right) - f(X^*)\right] \leq \frac{1}{2\eta_t}\,\mathrm{E}\left[\left\|X^{(1)} - X^*\right\|_2^2\right] - \frac{1}{2\eta_t T + 1}\mathrm{E}\left[\left\|X^{(T+1)} - X^*\right\|_2^2\right] + \sum_{t=1}^{T}\frac{\eta_t}{2}\mathrm{E}\left[\left\|v^{(t)}\right\|_2^2\right]$$

$$from\ \mathrm{E}[\|v\|_2^2] \leq \rho^2\ ,\left\|X - X^*\right\|_2^2 \leq B^2:$$

$$\sum_{t=1}^{T} \mathrm{E}\left[f\left(X^{(t)}\right) - f(X^*)\right] \leq \frac{1}{2\eta_t}\,B^2 - \frac{1}{2\eta_t T + 1}\mathrm{E}\left[\left\|X^{(T+1)} - X^*\right\|_2^2\right] + \sum_{t=1}^{T}\frac{\eta_t}{2}\rho^2$$

$$go\ on:$$

$$\sum_{t=1}^{T} \mathrm{E}\left[f\left(X^{(t)}\right) - f(X^*)\right] \leq \frac{1}{2\eta_t}\,B^2 + \sum_{t=1}^{T}\frac{\eta_t}{2}\rho^2$$

$$\mathrm{E}[f(X_T^{bar})] - f(X^*) \le \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[f(X^{(t)}) - f(X^*)] \le \frac{1}{2\eta_t}B^2 + \sum_{t=1}^{T}\frac{\eta_t}{2}\rho^2$$

$$so:$$

$$\mathrm{E}[f(X_T^{bar})] - f(X^*) \le \frac{1}{2\eta_t}B^2 + \sum_{t=1}^{T}\frac{\eta_t}{2}\rho^2$$

Q22.

$$let\ \eta_t = \frac{B}{\rho\sqrt{t}}\ and\ from\ Q21:$$

$$\mathrm{E}[f(X_T^{bar})] - f(X^*) \le \frac{1}{2T\eta_t}B^2 + \frac{1}{2T}\rho^2\sum_{t=1}^{T}\eta_t = \frac{B\rho}{2\sqrt{T}} + \frac{B\rho}{2T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$$

$$\frac{B\rho}{2\sqrt{T}} + \frac{B\rho}{2T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le \frac{3B\rho}{2\sqrt{T}} \rightarrow \sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$$

$$by\ induction\ and\ let\ T = 1: 1 < 2\ which\ is\ correct$$

$$for\ T+1\ we\ should\ show\ 2\sqrt{T}\ \le 2\sqrt{T+1} - \frac{1}{\sqrt{T+1}}\ if\ we\ square\ each\ side:$$

$$\frac{1}{T+1} \ge 0 \rightarrow T \ge 1\ \ so\ by\ induction\ proved$$

Q23.

Q24.

$$\mathrm{E}[.]\ Mathematical\ hope\ is\ a\ linear\ operator\ so:$$

$$V \epsilon \partial_X F(X,Z) \rightarrow$$

$$\mathrm{E}[\partial_X F(X,Z)] = \partial \mathrm{E}_Z[F(X,Z)] = \partial f(X)$$

$$so\ V\ is\ a\ Random\ subgradient$$

Simulation questions:

Simulation 1:

First of all, It is better to rewrite function F as below:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} |Z_i^T x - B_i| = \frac{1}{m}||Ax - B||_1$$

Where rows of matrix A include $Z_i's$. In this way, we can better define function f in python. Now we need to calculate sub gradient of function f. In order to calculate sub gradient of function f, we first try to gain a relation for sub gradient of norm 1 function. As we learned in the class:

$$v_i \in \partial g(x) \rightarrow v_i = \begin{cases} 1 & x_i > 0 \\ -1 & x_i < 0 \\ [-1,1] & x_i = 0 \end{cases}$$

Now we use above formula to calculate sub gradient of our function:

$$\partial f(x) = \partial\left(\frac{1}{m} g(Ax - B)\right) = \frac{1}{m} A^T \partial g(Ax - B)$$

In fact we achieved above formula using chain rule for gradient, which is extendable to sub gradients as well. Using these relations, we implement a python code to calculate sub gradient. Finally using sub gradient we can implement Sub-Gradient Descent algorithm:

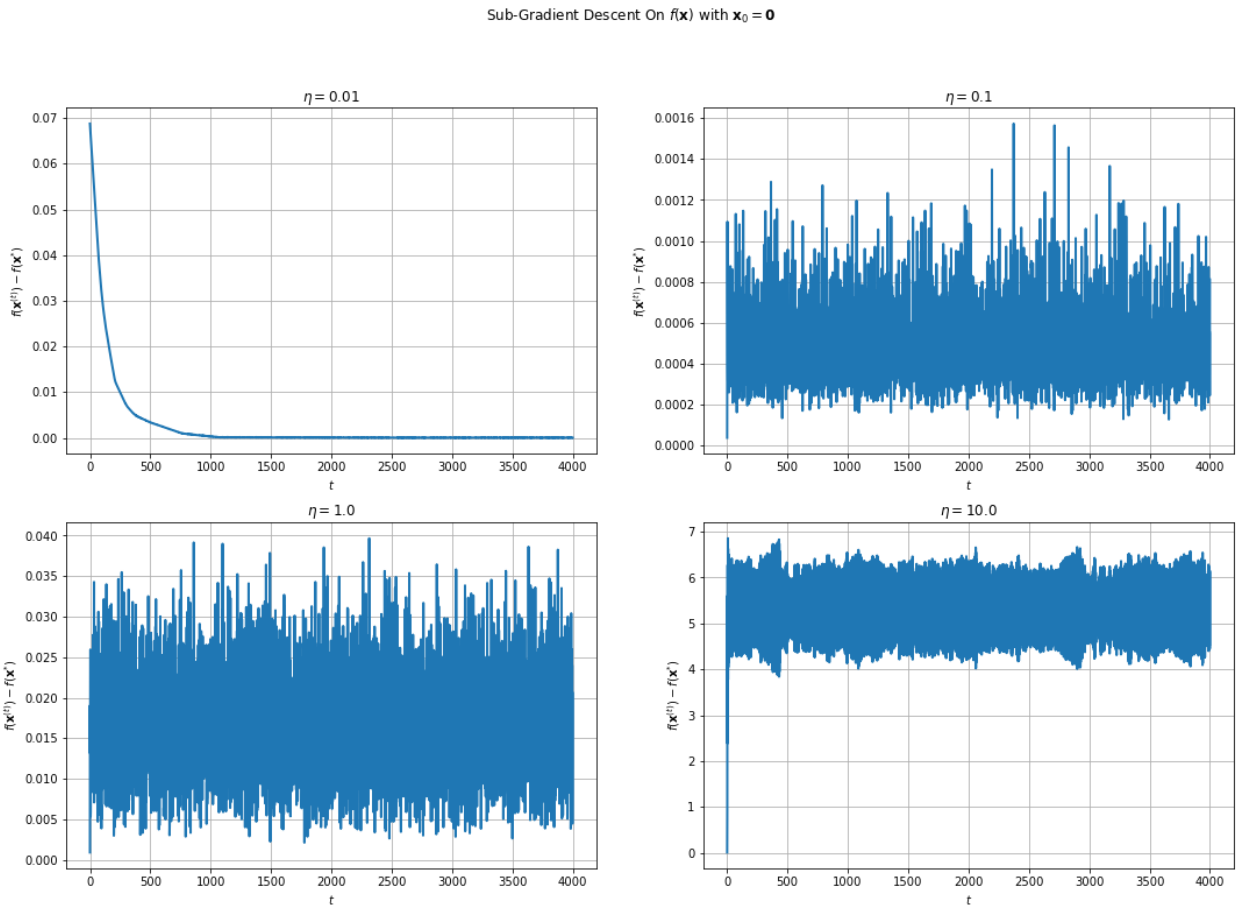
```
# functio below calculates subgradient of the function:

def norm1_subgrad(x):

        g = np.zeros_like(x)
        for i in range (x.size):
            if x[i]>0:
                g[i] = 1
            else:
                g[i] = -1

        return g

def subgrad(x, A, B):                          # This function gives subgradient of f(x)
    m = B.size
    return 1/m * A.T @ norm1_subgrad(A @ x - B)

```

Now we use Sub-Gradient Descent algorithm and update values of x_t:

```
T   = 4000
x0  = np.zeros(n)
eta = np.array([0.01, 0.1, 1, 10])

# Implementing Sub-Gradient Decent algorithm:

plt.figure(figsize = (18, 12))

for i in range (eta.size):
    x_new = x0
    lst = []
    for j in range(T+1):

        diff = obj_func(x_new, A, B)-obj_func(x_opt, A, B)
        lst.append(diff)
        x_new -= eta[i]*subgrad(x_new, A, B)

    plt.subplot(int(np.ceil(np.sqrt(eta.size))), int(np.ceil(np.sqrt(eta.size))), i + 1)
    plt.plot(np.arange(T + 1), lst, linewidth=2)
    plt.xlabel(r'$t$')
    plt.ylabel(r'$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{*})$')
    plt.title(r'$\eta = ${eta}'.format(eta=eta[i]))
    plt.grid(True)

plt.suptitle(r'Sub-Gradient Descent On $f(\mathbf{x})$ with $\mathbf{x}_0 = \mathbf{0}$')
plt.show()
```

Now, we can see the results of updating $x_t's$ with different coefficients of η. Note that we used CVXPY library to calculate the minimum of the function which is denoted as $f(x^*)$ in the code. Running the code, we get the following results:

```
optimal X is :

 [[-0.1667894 ]
 [-0.22008943]
 [ 0.13805849]
 [ 0.09134012]
 [-0.11563373]
 [ 0.04394651]
 [-0.15485976]
 [-0.35966015]
 [-0.04665486]
 [-0.06390327]]

f(x*) = [0.79011531]

f(x0) = [0.85887246]
```



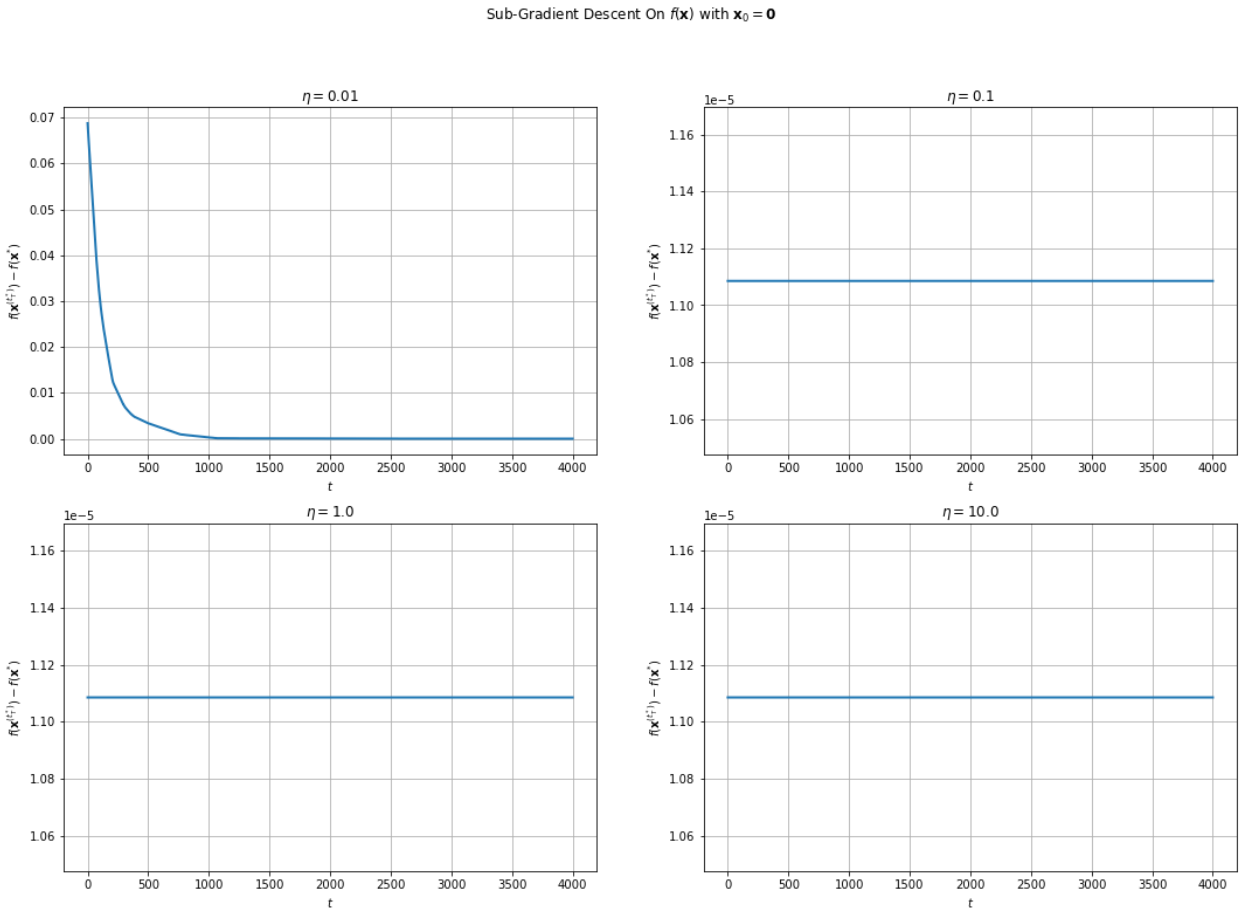
Which shows, selecting small enough value for η is essential for appropriate convergence of the Sub-Gradient Descent algorithm. As you can see $\eta = 0.01$ is the best choice for pretty good convergence in SGD algorithm.

Simulation 2:

We calculate minimum value of the function until current iteration:

```
x0 = np.zeros(n)          # Initial value
f0 = obj_func(x0, A, B)

difference = []      # saves minimum between f(x) and f0 for every x from the begining of the algorithm until current step
eta = np.array([0.01, 0.1, 1, 10])

plt.figure(figsize = (18,12))
for i in range(eta.size):
    x_new = x0
    difference = []
    for t in range(T+1):

        f_xnew = obj_func(x_new, A, B)
        f0 = np.min((f_xnew, f0))
        x_new -= eta[i]*subgrad(x_new, A, B)
        difference.append(f0 - obj_func(x_opt, A, B))     #f(x(t*)-f(x*))

    plt.subplot(int(np.ceil(np.sqrt(eta.size))), int(np.ceil(np.sqrt(eta.size))), i + 1)
    plt.plot(np.arange(T + 1), difference, linewidth=2)
    plt.xlabel(r'$t$')
    plt.ylabel(r'$f(\mathbf{x}^{(t^*_T)}) - f(\mathbf{x}^{*})$')
    plt.title(r'$\eta = ${eta}'.format(eta=eta[i]))
    plt.grid(True)

plt.suptitle(r'Sub-Gradient Descent On $f(\mathbf{x})$ with $\mathbf{x}_0 = \mathbf{0}$')
plt.show()
```

Which leads us to the following results:



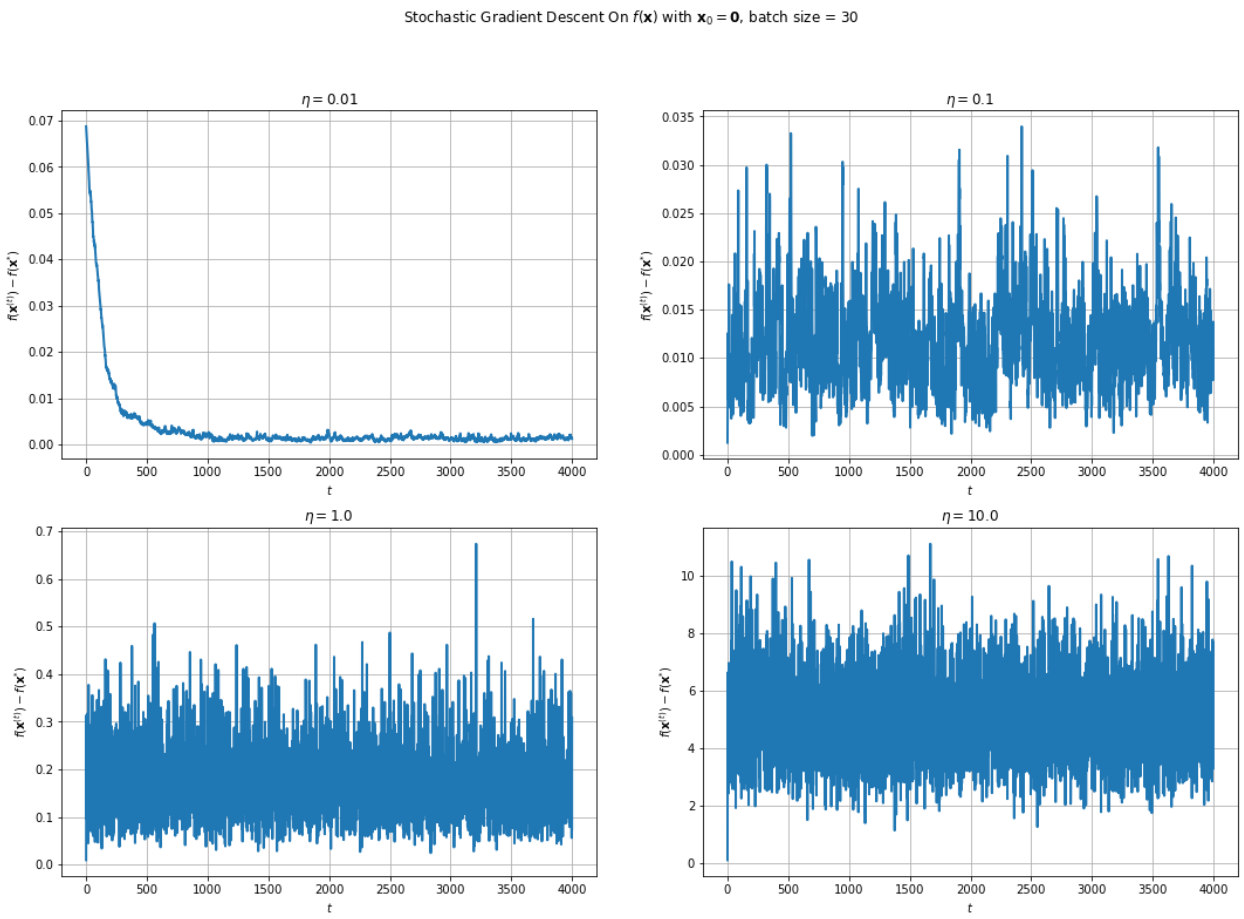
Note that for $\eta = \{0.1, 1, 10\}$ the real plot would be something like L word in English. Because it has been plotted bigger, it is not quite observable.

Simulation 3:

The code below in python implements Stochastic-Gradient Descent algorithm:

```
batch_size = 30
T  = 4000
x0  = np.zeros(n)
eta = np.array([0.01, 0.1, 1, 10])
# Implementing Stochastic_Gradient_Decent algorithm:

plt.figure(figsize = (18, 12))

for i in range (eta.size):
    x_new = x0
    lst = []
    for j in range(T+1):

        diff = obj_func(x_new, A, B)-obj_func(x_opt, A, B)
        lst.append(diff)
        random_samples = np.random.choice(m, batch_size)

        # choose a random sample set from dataset

        A_t = A[random_samples, :]
        B_t = B[random_samples]
        V_t = subgrad(x_new, A_t, B_t)  # random subgradient where expected value of V_T belongs to the subgradient of f(X_t
        x_new -= eta[i]*V_t

    plt.subplot(int(np.ceil(np.sqrt(eta.size))), int(np.ceil(np.sqrt(eta.size))), i + 1)
    plt.plot(np.arange(T + 1), lst, linewidth=2)
    plt.xlabel(r'$t$')
    plt.ylabel(r'$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{*})$')
    plt.title(r'$\eta = ${eta}'.format(eta=eta[i]))
    plt.grid(True)

plt.suptitle(r'Stochastic Gradient Descent On $f(\mathbf{x})$ with $\mathbf{x}_0 = \mathbf{0}$, batch size = $30$')
plt.show()

```

Now we can discuss the results:

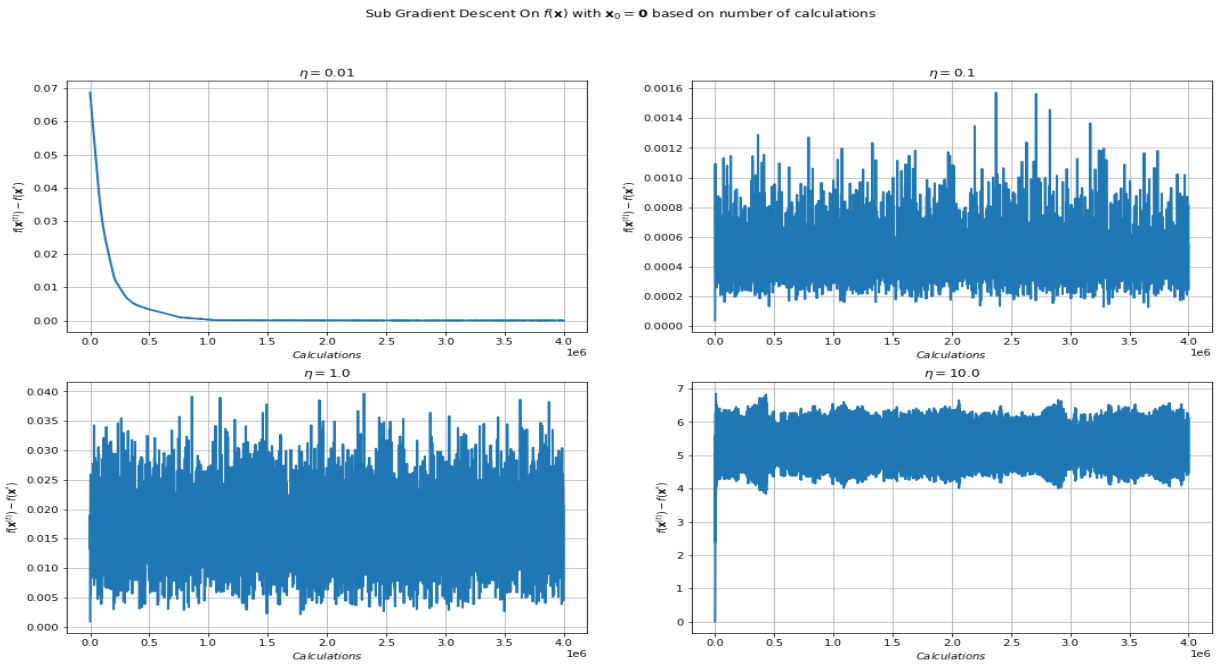


As a matter of difference, we can mention convergence speed and code's speed in running. Using Stochastic-Gradient Descent algorithm will result in faster execution of algorithm, but slightly decreases the convergence speed.

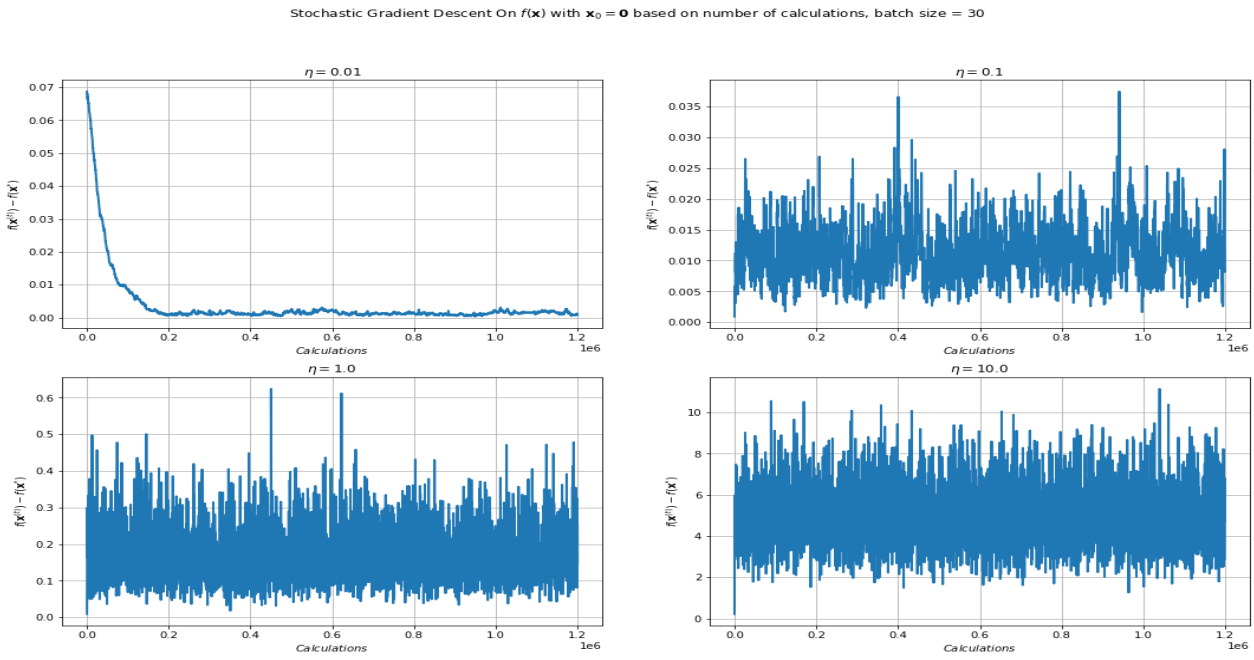
Simulation 4:

Let’s see both results simultaneously.

Below we can see Sub-Gradient Descent versus number of calculations:



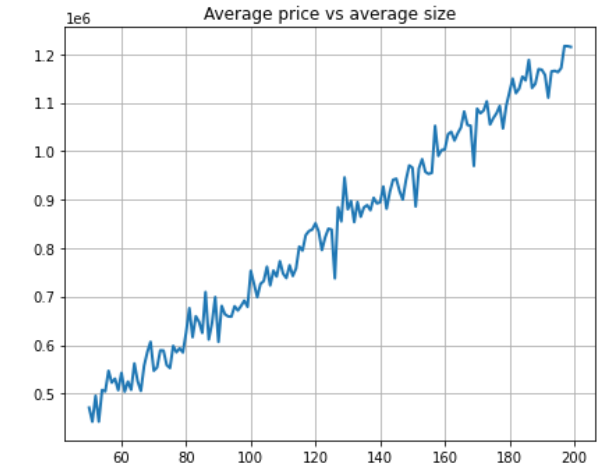
Below we can see Stochastic-Gradient Descent versus number of calculations:

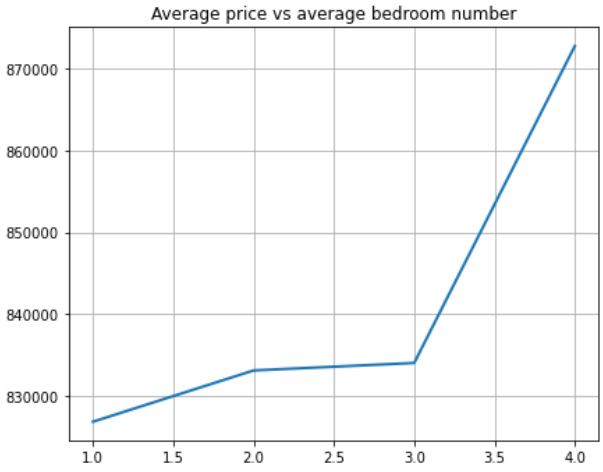


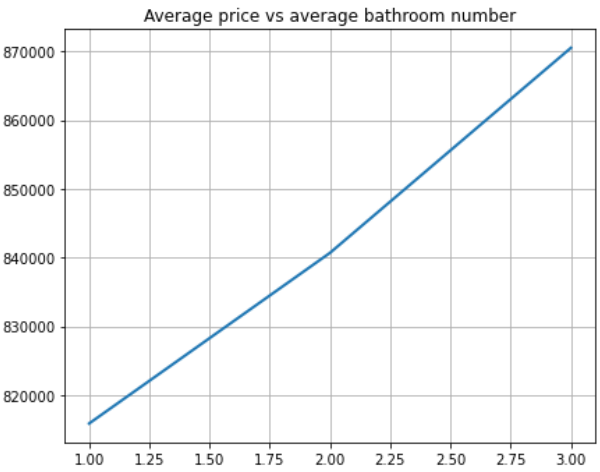
It is obvious that Stochastic-Gradient Descent converges in lower number of calculations. This result totally makes sense with the previous results in simulation 3 which we mentioned that Stochastic-Gradient Descent converges to the final value in lower iterations. These results are for $\eta = 0.01$

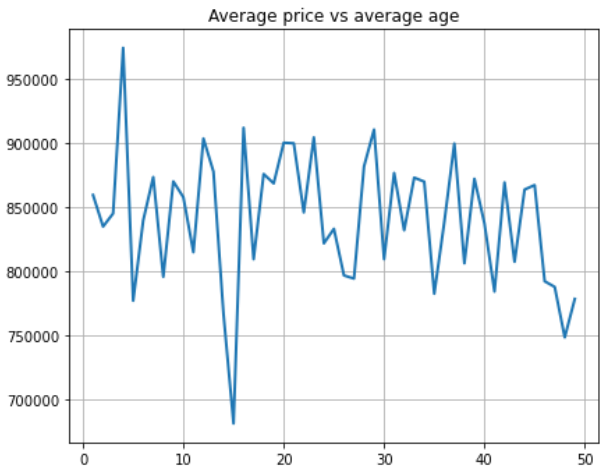
Simulation 5:

Let's take a look at plots that demonstrate relation between a feature of a house and price of the house:









Take size of the house as a matter of illustration. As the size of the house increases, the price of the house increases proportionally. This is true for number of bathrooms as well. On the other hand, age of the house is not that important and lower age of the construction, doesn't necessarily

guarantee higher price of the house, meaning that the relation between average age of the house and the average price is not proportional. Note that repetitions of some features can be a challenge. For example there are lots of houses with 1 bedrooms which vary on price. In order to handle this issue we considered the average value of the bedrooms. That why you can see values such as 2.25 in the bathroom section.

Simulation 6:

Below code, divides data into two parts. Train data and test data. The ratio of this separation is optional and we took 0.75 of the whole data as train data and rest of the data would be test data. We did this using **sample** function in python to shuffle data and then split data into two groups.

```
# Splitting the data to train and test

sampled_house_data = house_data.sample(frac=1).to_numpy()       # we shuffle data using sample function

ratio = 0.75
total_rows = sampled_house_data.shape[0]
train_size = int(total_rows*ratio)
train_data = sampled_house_data[:train_size, :]       # This would be the training data
test_data  = sampled_house_data[train_size:, :]       # This would be the test data


train_x = train_data[:, :-1]
train_y = train_data[:, -1].reshape(-1, 1)
test_x = test_data[:, :-1]
test_y = test_data[:, -1].reshape(-1, 1)

# Normalizing train data
train_x_means = train_x.mean(axis=0)
train_x_stds = train_x.std(axis=0)
train_y_means = train_y.mean(axis=0)
train_y_stds = train_y.std(axis=0)

train_x_normalized = (train_x - train_x_means) / train_x_stds
train_y_normalized = (train_y - train_y_means) / train_y_stds

test_x_normalized = (test_x - train_x_means) / train_x_stds
test_y_normalized = (test_y - train_y_means) / train_y_stds

```

Simulation 7:

Below function, calculates MSE:

```
# Function below calculates MSE:

def MSE_Calculator(X, y, w, b):
    m, n = X.shape
    one = np.ones((m, 1))
    y_hat = X @ w + b * one
    mse = ((y-y_hat)**2).sum()
    return mse / m

```

Simulation 8:

Function below calculates SGD:

```

def SGD(X, y, w, b, lr=1e-3):
    m, n = X.shape
    one = np.ones((m, 1))
    # Calculate gradients

    w_grad = 2*X.T @ X @ w + 2*b * X.T @ one - 2*X.T @ y
    b_grad = 2*one.T @ X @ w + 2*b* one.T @ one - 2*one.T @ y

    w = w - lr * w_grad
    b = b - lr * b_grad

    return w, b

```

Simulation 9:

Take following specification:

$$number\ of\ epochs = 30$$

$$batch\_size = 64$$

$$learning\ rate = 10^{-4}$$

```

NUM_EPOCHS = 30
LEARNING_RATE = 1e-4
BATCH_SIZE = 64

train_errors = []
test_errors = []

m, n = train_x.shape

# Initialize model parameters

w = np.random.rand(n, 1)
b = np.random.rand(1)

for epoch in range(1, NUM_EPOCHS+1):
    w, b = iteration(
        X=train_x_normalized,
        y=train_y_normalized,
        w=w,
        b=b,
        batch_size=BATCH_SIZE,
        lr=LEARNING_RATE
    )
    mse_train = MSE_Calculator(train_x_normalized, train_y_normalized, w, b)
    mse_test = MSE_Calculator(test_x_normalized, test_y_normalized, w, b)
    train_errors.append(mse_train)
    test_errors.append(mse_test)
    print(f'epoch {epoch}, train mse: {np.round(mse_train, 3)}, test mse: {np.round(mse_test, 3)}')
    print()

```

```

fig = plt.figure(figsize=(8, 6))
ax = fig.add_subplot(1, 1, 1)
ax.plot(range(1, NUM_EPOCHS+1), train_errors, label='train error')
ax.plot(range(1, NUM_EPOCHS+1), test_errors, label='test error')
ax.legend();
ax.grid(True)
ax.set_title('Test and train MSEs after each epoch')
ax.set_xlabel('Epochs')
ax.set_ylabel('MSE')

```

And the result of the code would be: