In the name of God

Convex Optimization Project

Dr. Yassaee

Mobin Khatib

99106114

Amir Ali Loghmani

99102145

1 Theory Question

Given the function

$$f(x) = \frac{۱}{۲}x^T Ax - b^T x + c \tag{۱}$$

, we use the gradient descent update rule at step : t with learning rate η

$$x_{t+۱} = x_t - \eta_t \nabla f(x_t) \tag{۲}$$

We want to show that:

$$x_{t+۱} = (I - \eta_t A)x_t + \eta_t b \tag{۳}$$

Let's assume $f(x)$ has an optimal solution $x*$:

$$x_{t+۱} = (I - \eta_t A)x_t + \eta_t b \tag{۴}$$

$$x_{t+۱} - x^* = (x_t - \eta_t(Ax_t - b)) - (x^* - \eta_t(Ax^* - b)) \tag{۵}$$

$$x_{t+۱} - x^* = (I - \eta_t A)(x_t - x^*) = \prod_{k=۱}^{n}(I - \eta_k A)(x_۱ - x^*) \tag{۶}$$

Considering above equation The convergence of gradient descent for quadratic functions can be analyzed mathematically. In particular, for a strictly convex quadratic function, the sequence of iterates generated by gradient descent will converge to the optimal solution x^* as the number of iterations increases. The rate of convergence depends on the eigenvalues of the matrix A and the choice of the learning rate

$$\lim_{T\to\infty} \|x(T) - x^*\| = ۰ \tag{۷}$$

When we want this because all of the x's from 1 to n are multiplied in equation 6 if one of the answers for $x_t - x^* = 0$ then equation number 6 becomes zero and equation number 7 will be satisfied

2 Theory Question

$$x_{(t+1)} = \arg\min_{x \in R^n} f(x_{(t)}) + < \nabla f(x_{(t)}), x - x_{(t)} > + \frac{1}{2\eta}|x - x_{(t)}|^2 \quad (8)$$

where:

$f(x_{(t)})$ represents the objective function evaluated at $x_{(t)}$. $\nabla f(x_{(t)})$ denotes the gradient of the objective function with respect to x, evaluated at $x_{(t)}$. $\langle \nabla f(x_{(t)}), x - x_{(t)} \rangle$ represents the inner product (dot product) between the gradient $\nabla f(x(t))$ and the difference $x - x_{(t)}$. $\frac{1}{2\eta}|x - x_{(t)}|^2$ is a regularization term that penalizes the distance between x and $x_{(t)}$, scaled by the parameter η. To find $x_{(t+1)}$, we need to compute the value of x that minimizes this expression. This can be done by taking the derivative of the expression with respect to x, setting it equal to zero, and solving for x. Let's differentiate the expression:

$$\frac{d}{dx} f(x_{(t)}) + \langle \nabla f(x_{(t)}), x - x_{(t)} \rangle + \frac{1}{2\eta}||x - x_{(t)}||^2 = 0 \quad (9)$$

To simplify the calculation, let's introduce some intermediate variables:

Let $A = \nabla f(x_{(t)})$, which is the gradient of the objective function. Let $B = x - x_{(t)}$, which is the difference between x and $x_{(t)}$. Now, let's differentiate each term of the expression:

since $f(x_{(t)})$ is a constant with respect to x then:

$$\frac{d}{dx} f(x_{(t)}) = 0 \quad (10)$$

$$\frac{d}{dx}\langle \nabla f(x_{(t)}), x - x_{(t)} \rangle = \langle \nabla f(x_{(t)}), \frac{d}{dx}(x - x(t)) \rangle = \quad (11)$$

$$\langle \nabla f(x_{(t)}), dx \rangle = \langle A, dx \rangle. \quad (12)$$

$$\frac{d}{dx}\frac{1}{2\eta}||x - x_{(t)}||^2 = \frac{1}{2\eta}\frac{d}{dx}||B||^2 = \quad (13)$$

$$\frac{1}{2\eta}\frac{d}{dx}\langle B, B \rangle = \frac{1}{2\eta} \cdot 2\langle B, \frac{dB}{dx} \rangle = \quad (14)$$

$$\frac{1}{\eta}\langle B, \frac{dB}{dx} \rangle. \quad (15)$$

Putting it all together:

$$۰ = \frac{d}{dx} f(x_{(t)}) + \langle \nabla f(x_{(t)}), x - x_{(t)} \rangle + \frac{۱}{۲\eta} ||x - x_{(t)}||^۲ = \tag{۱۶}$$

$$\langle A, dx \rangle + \frac{۱}{\eta} \langle B, \frac{dB}{dx} \rangle = \tag{۱۷}$$

$$\langle A, dx \rangle + \frac{۱}{\eta} \langle B, (dx - dx_{(t)}) \rangle = \tag{۱۸}$$

$$\langle A, dx \rangle + \frac{۱}{\eta} \langle B, dx \rangle = \tag{۱۹}$$

$$\langle A + \frac{۱}{\eta} B, dx \rangle. \tag{۲۰}$$

To satisfy the equation for all values of dx, the coefficient of dx must be zero:

$$A + \frac{۱}{\eta} B = ۰. \tag{۲۱}$$

Substituting the values of A and B, we get:

$$\nabla f(x_{(t)}) + \frac{۱}{\eta} (x - x_{(t)}) = ۰. \tag{۲۲}$$

Rearranging the equation, we find:

$$\nabla f(x_{(t)}) = -\frac{۱}{\eta} (x - x_{(t)}). \tag{۲۳}$$

Now, solving for x:

$$x = x_{(t)} - \eta \nabla f(x_{(t)}). \tag{۲۴}$$

Therefore, the value of x that minimizes the given expression is given by:

$$x_{(t+۱)} = x_{(t)} - \eta \nabla f(x_{(t)}). \tag{۲۵}$$

This expression represents an update rule commonly used in optimization algorithms like gradient descent, where $x_{(t+1)}$ is updated based on the gradient of the objective function at $x_{(t)}$, scaled by a learning rate η.

The equation suggests that we can improve our current estimate or solution $x(t)$ by taking a step in the opposite direction of the gradient $\nabla f(x(t))$. The gradient $\nabla f(x(t))$ provides information about the direction of steepest ascent of the objective function $f(x)$, so moving in the opposite direction corresponds to descending or The relationship captures an iterative optimization process minimizing the function. where each step involves updating the current estimate $x(t)$ based on the gradient information and the regularization term to minimize the objective function $f(x)$. The algorithm aims to converge to a solution that optimizes the objective function within the given constraints.

3 Theory Question

With replacing $x_{(t+1)} = x_{(t)} - \eta_t v_t$

$$||x_{t+۱} - x^*||_۲^۲ = ||x_t - \eta_t v_t - x^*||_۲^۲ \tag{۲۶}$$

As we have $\eta_t \geq 0$ and $v_t \geq 0$(because it is in $\sigma f(x_t)$) thus:

$$||x_{t+۱} - x^*||_۲^۲ \leq ||x_t - x^*||_۲^۲ \tag{۲۷}$$

4 Theory Question

We know that we can replace x, y with any two other things in the L-lipchitz formula and so with this, we should prove that:

$$|f(x) - f(y)| \leq L||x - y|| \leftrightarrow ||v|| \leq L, v \in \sigma f(w) \quad ** * \tag{۲۸}$$

Because the left side of inequality *** is true for any x,y , therefore we can choose x,y in the way of that norm of z become equal or bigger than zero:

$$|x - y| = ۱, (x - y)^T z = ||z||_۲, z \in \sigma f(w) \tag{۲۹}$$

And we know that f is convex so:

$$f(x) \geq f(y) + (x - y)^T z \tag{۳۰}$$

$$\rightarrow |f(x) - f(y)| \geq z^T (x - y) \tag{۳۱}$$

$$|f(x) - f(y)| \leq L||x - y|| \tag{۳۲}$$

we have $|x - y| = 1$ and $(x - y)^T z = ||z||_2$ so with this above equations:

$$||z|| \leq L \tag{۳۳}$$

And for right side of the *** inequality (because f is convex):

$$f(x) \geq f(y) + (x - y)^T z, z^T \in f(y) \rightarrow |f(x) - f(y)| \geq z^T(x - y) \tag{۳۴}$$

$$\leq ||z||||x - y|| \leq L||x - y|| \rightarrow \tag{۳۵}$$

$$|f(x) - f(y)| \leq L||x - y|| \tag{۳۶}$$

5 Theory Question

$$||x_{t+۱} - x^*||_۲^۲ = ||x_t - \eta v_t - x^*||_۲^۲ \tag{۳۷}$$

$$= ||x_t - x^*||_۲^۲ - ۲\eta_t v_t^T(x_t - x^*) + \eta_t^۲||v_t||_۲^۲ \tag{۳۸}$$

$$\leq ||x_t - x^*||_۲^۲ - ۲\eta_t(f(x_t) - f^*) + \eta_t^۲||v_t||_۲^۲ \tag{۳۹}$$

where $f^* = f(x^*)$ The last line follows from the definition of subgradient, which gives

$$f(x^*) \geq f(x_{(t)}) + v_{(t)}^T(x^* - x_{(t)}) \tag{۴۰}$$

So we have:

$$\frac{۱}{۲}||x_{(t+۱)} - x^*||_۲^۲ \leq \frac{۱}{۲}||x_{(t)} - x^|_۲^۲ - \eta_t(f(x_{(t)}) - f(x^*)) + \frac{\eta_t^۲}{۲}||v_{(t)}||_۲^۲ \tag{۴۱}$$

6 Theory Question

$$\eta_t(f(x_{(t)}) - f(x^*)) \leq \frac{۱}{۲}||x_{(t)} - x^*||^۲ - \frac{۱}{۲}||x_{(t+۱)} - x^*||^۲ + \frac{۱}{۲}\eta_t^۲||v_{(t)}||^۲ \tag{۴۲}$$

Given:

$$||v||^{۲} \leq \rho \tag{۴۳}$$

and from the result of question 4:

$$||x^*||^{۲} \leq B \tag{۴۴}$$

By recursively substituting, we have:

$$t = ۱ \Rightarrow ||x^* - x_{(۱)}|^{۲} \leq B\eta_{۱}(f(x_{(۱)}) - f(x^*)) \leq \frac{۱}{۲}B^{۲} - \frac{۱}{۲}||x_{(۲)} - x^*||^{۲} + \frac{۱}{۲}\eta_{۱}^{۲}\rho^{۲} \tag{۴۵}$$

$$t = ۲ \Rightarrow \eta_{۲}(f(x_{(۲)}) - f(x^*)) \leq \frac{۱}{۲}B^{۲} - \frac{۱}{۲}||x_{(۲)} - x^*||^{۲} - \frac{۱}{۲}||x_{(۳)} - x^*||^{۲} + \frac{۱}{۲}\eta_{۲}^{۲}\rho^{۲} \tag{۴۶}$$

...

$$t = T \Rightarrow \eta_{T}(f(x_{(T)}) - f(x^*)) \leq \frac{۱}{۲}B^{۲} - \frac{۱}{۲}||x_{(T)} - x^*||^{۲} - \frac{۱}{۲}||x_{(T+۱)} - x^*||^{۲} + \frac{۱}{۲}\eta_{T}^{۲}\rho^{۲} \tag{۴۷}$$

$$\sum_{t=۱}^{T} \eta_t \left(f(x_{(t)}) - f(x^*)\right) \leq \frac{۱}{۲}B^{۲} + \frac{۱}{۲}\rho^{۲}\sum_{t=۱}^{T}\eta_t^{۲} - \frac{۱}{۲}||x_{(T+۱)} - x^*||^{۲} \tag{۴۸}$$

Since we have:

$$\frac{۱}{۲}||x_{(T+۱)} - x^*||^{۲} \geq ۰ \tag{۴۹}$$

From this, we can conclude that:

$$\sum_{t=۱}^{T} \eta_t \left(f(x_{(t)}) - f(x^*)\right) \leq \frac{۱}{۲}B^{۲} + \frac{۱}{۲}\rho^{۲}\sum_{t=۱}^{T}\eta_t^{۲} \tag{۵۰}$$

7 Theory Question

We define this:

$$\sum_{t=۱}^{T} \eta_t = a \tag{۵۱}$$

$$\sum_{t=۱}^{T} \eta_t f(x_{(t)}) - a f(x^*) \leq \frac{۱}{۲} B^۲ + \frac{۱}{۲} \rho^۲ \sum_{t=۱}^{T} \eta_t^۲ \tag{۵۲}$$

Now we divide this by a and we will have:

$$\sum_{t=۱}^{T} \frac{\eta_t}{a} f(x_{(t)}) - f(x^*) \leq \frac{\frac{۱}{۲} B^۲}{a} + \frac{\frac{۱}{۲} \rho^۲ \sum_{t=۱}^{T} \eta_t^۲}{a} \tag{۵۳}$$

On the other hand because f is convex, we have:

$$f(\bar{x}_T) = f\left(\sum_{t=۱}^{T} \frac{\eta_t}{a} x_{(t)} \right) \leq \sum_{t=۱}^{T} \frac{\eta_t}{a} f(x_{(t)}) \tag{۵۴}$$

From the two above equations, it follows that:

$$f(\bar{x}_T) - f(x^*) \leq \frac{B^۲ + \rho^۲ \sum_{t=۱}^{T} \eta_t^۲}{۲ \sum_{t=۱}^{T} \eta_t} \tag{۵۵}$$

8 Theory Question

with substitution $\eta_t = \eta$

$$f(\bar{x}_T) - f(x^*) \leq \frac{B^۲ + T \rho^۲ \eta}{۲ T \eta} \tag{۵۶}$$

To obtain the optimal bound, we take the derivative of the right-hand side and set it equal to zero:

$$\frac{d}{d\eta} \left(\frac{B^۲ + T \rho^۲ \eta}{۲ T \eta} \right) = ۰ \rightarrow \eta = \frac{B}{\rho} \sqrt{\frac{۱}{T}} \tag{۵۷}$$

By substituting this value back, we have:

$$f(\bar{x}_T) - f(x^*) \leq B \rho \sqrt{\frac{۱}{T}} \tag{۵۸}$$

9 Theory Question

Based on the convexity property, we arrived at this particular form of expression by turning the coefficient into a sum of functions and incorporating them inside the function to make it a smaller set.

Indeed, let η be a vector, but the main intuition comes from the fact that we previously proved that for every random subgradients, we need to find η such that the function is decreasing in that direction. However, for this question, it is not necessary that all our steps be optimal. So, a reasonable approach is to not focus solely on the final solution and instead provide an effect from all the steps as an output. One such effect could be the arithmetic mean, which was examined in this question.

10 Theory Question

$$\sum_{t=۱}^{T} \eta_t f(x_{(t)}) \geq (\sum_{t=۱}^{T} \eta_t) f(x_{(t_T^*)}) \tag{۵۹}$$

Using the above equation and also question 7, we have:

$$\sum_{t=۱}^{T} \eta_t f(x_{(t)}) - a f(x^*) \leq \frac{۱}{۲} B^۲ + \frac{۱}{۲} \rho^۲ \sum_{t=۱}^{T} \eta_t^۲ \tag{۶۰}$$

Deviding by $a = \sum_{t=1}^{T} \eta_t$, we have:

$$\rightarrow f(x_{(t_T^*)}) - f(x^*) \leq \frac{۱}{۲} B^۲ + \frac{\frac{۱}{۲} \rho^۲ \sum_{t=۱}^{T} \eta_t^۲}{\sum_{t=۱}^{T} \eta_t} \tag{۶۱}$$

11 Theory Question

$$C = (x \in \mathbb{R}^n \mid Ax = b) \tag{۶۲}$$

The goal is to find the closed-form expression for the projection of x onto C, denoted as $\Pi_C(x)$, which minimizes the squared Euclidean distance between y and x:

$$\Pi_C(x) = \arg\min_{y \in C} ||y - x||^۲ \tag{۶۳}$$

To solve this optimization problem, we can utilize the concept of the Moore-Penrose pseudoinverse.

Formulating the Lagrangian function:

$$L(y, \lambda) = ||y - x||^۲ + \lambda^T (Ay - b) \tag{۶۴}$$

where λ is a Lagrange multiplier vector associated with the equality constraint $Ax = b$.

Taking the partial derivative of L with respect to y and setting it to zero:

$$\frac{\partial L}{\partial y} = ۲(y - x) + A^T \lambda = ۰ \tag{۶۵}$$

This yields the following equation:

$$y = x - ۰/۵ A^T \lambda \tag{۶۶}$$

Substituting this expression for y into the equality constraint $Ax = b$, we get:

$$Ax - ۰/۵ A A^T \lambda = b \tag{۶۷}$$

To solve for λ, we can left-multiply the equation by $A^\dagger$ (the left inverse of A) to obtain:

$$A^\dagger Ax - ۰/۵ A^\dagger A A^T \lambda = A^\dagger b \tag{۶۸}$$

Simplifying further:

$$۰/۵ A^\dagger A A^T \lambda = x - A^\dagger b \tag{۶۹}$$

Rearranging the equation:

$$\lambda = ۲(x - A^\dagger b)^\dagger A^\dagger A (A^T)^\dagger \tag{۷۰}$$

Multiplying both sides by 2 and solving for λ:

$$\lambda = ۲ \left(x - A^{\dagger}b\right)^{\dagger} A^{\dagger}A(A^{T})^{\dagger} \tag{۷۱}$$

Finally, substituting the expression for λ back into $y = x - 0.5A^{T}\lambda$, we obtain the closed-form expression for $\Pi_C(x)$:

$$\Pi_C(x) = x - ۰٫۵A^{T} \left(A^{\dagger}A(A^{T})^{\dagger}\right) \left(۲(x - A^{\dagger}b)\right) \tag{۷۲}$$

TA add the knowledge that AA^T is positive definite and m<n so the† will be gonna out and the final answer will be:

$$y = x - A^{T}(AA^{T})^{-۱}(Ax - b) \tag{۷۳}$$

This represents the projection of x onto the set C in closed form.

12 Theory Question

When the set C is defined as:

$$C = (x \in \mathbb{R}^n \mid Ax \leq b) \tag{۷۴}$$

where A is an $m \times n$ matrix, x is an n-dimensional vector, and b is an m-dimensional vector, the closed-form expression for the projection of x onto C, denoted as $\Pi_C(x)$, can be found using the concept of linear programming and optimization.

The projection $\Pi_C(x)$ can be defined as the solution to the following optimization problem:

$$\Pi_C(x) = \arg\min_{y \in C} ||y - x||^{۲} \tag{۷۵}$$

To solve this problem, we can formulate it as a quadratic program with linear constraints. The objective function is to minimize the squared Euclidean distance between y and x:

$$\text{minimize:} ||y - x||^{۲} \tag{۷۶}$$

subject to: $Ax \leq b$

To find the closed-form expression, we can utilize the theory of quadratic programming and the KKT (Karush-Kuhn-Tucker) conditions.

The KKT conditions for this problem are as follows:

Stationarity condition:

$$٢(y - x) + A^\top v = ٠ \tag{٧٧}$$

Primal feasibility:

$$Ax - b \leq ٠ \tag{٧٨}$$

Dual feasibility:

$$v \geq ٠ \tag{٧٩}$$

Complementary slackness:

$$v^\top (Ax - b) = ٠ \tag{٨٠}$$

Here, v is a vector of Lagrange multipliers associated with the inequality constraints $Ax \leq b$.

By solving the KKT conditions, we can find the closed-form expression for $\Pi_C(x)$.

13 Theory Question

To find the center of the sphere and the intersection point of the line passing through the origin and a given point x with the sphere, we can proceed as follows:

Let the center of the sphere be denoted as c. If we consider the tangent line to the sphere at the intersection point x, we know that this line must pass through the center of the sphere. Using the right triangle formed by the line connecting x, c, and the origin, we can prove that the line along x must coincide with the line connecting c and the origin.

Let's find the value of b relative to the center of the sphere. The intersection of this line segment with the sphere of radius b can be found by considering the line along $(y : y = tx)$.

$$y = \frac{x}{||x||_۲} \cdot b \tag{۸۱}$$

In this equation, c represents the center of the sphere, x is the given point, and b is the radius of the sphere. By multiplying the normalized vector x by the scalar $b/||x||_2$, we obtain the center of the sphere.

14 Theory Question

A : $\mathrm{x}^{t+1} = \Pi_c(x^t - \eta_t v^t) = argmin_{y \in C} \|y - (x^t - \eta_t v^t)\|_2 =$
$argmin_{y \in C} \|y - (x^t - \eta_t v^t)\|_2^2 = argmin_{y \in C} \left\| y^T.y - 2y^T(x^t - \eta_t v^t) + \|(x^t - \eta_t v^t)\|_2^2 \right\|_2 =$
$argmin_{y \in C}(\|y\|_2^2 - 2y^T(x^t - \eta_t v^t)))$
B : $\mathrm{x}^{t+1} = argmin_{x \in C}(f(x^t) + \langle v^t, x - x^t \rangle + \frac{1}{2\eta_t} \|(x - x^t)\|) =$
$\mathrm{argmin}_{x \in C}((v^t)^T.x + \frac{1}{2\eta_t}(\|(x)\|_2^2 - 2x^T x^t)) = argmin_{x \in C}(\|(x)\|_2^2 - 2x^T(x^t - 2\eta_t v^t))$
$\overset{y=x}{\rightarrow} A \equiv B$

15 Theory Question

$$\xrightarrow[algorithm\ 2\ theorems]{} \frac{1}{2}\left\|x^{t+1}-x^*\right\|_2^2 \le \frac{1}{2}\left\|x^t-x^*\right\|_2^2 - \eta_t(f(x^t)-f(x^*)) + \frac{\eta_t^2}{2}\left\|v^t\right\|_2^2$$
$$\xrightarrow{.\frac{1}{\eta_t}} \frac{1}{2\eta_t}\left\|x^{t+1}-x^*\right\|_2^2 \le \frac{1}{2\eta_t}\left\|x^t-x^*\right\|_2^2 - (f(x^t)-f(x^*)) + \frac{\eta_t}{2}\left\|v^t\right\|_2^2$$
$$\xrightarrow[1.4,2.4\ theorems]{\sum_{t=1}^{T}} \sum_{t=1}^{T}(f(x^t)-f(x^*)) \le \frac{B^2}{2\eta_t} + \frac{\rho^2}{2}\sum_{t=1}^{T}\eta_t$$

16 Theory Question

$$f(\bar{x}_t) = f(\tfrac{x_1+\ldots+x_T}{T}) \xrightarrow[jensen\ inequality]{f\ is\ convex} f(\bar{x}_t) \le \frac{1}{T}\sum_{t=1}^{T} f(x^t)$$
$$\to f(\bar{x}_t) - f(x^*) \le (\frac{1}{T}\sum_{t=1}^{T} f(x^t)) - f(x^*) \overset{15}{\to} \sum_{t=1}^{T}(f(x^t)-f(x^*)) \le \frac{B^2}{2\eta_T} + \frac{\rho^2}{2}\sum_{t=1}^{T}\eta_t$$
$$\overset{*\frac{1}{T}}{\to} (\frac{1}{T}\sum_{t=1}^{T} f(x^t)) - f(x^*) \le \frac{B^2}{2\eta_T T} + \frac{\rho^2}{2T}\sum_{t=1}^{T}\eta_t \to f(\bar{x}_t) - f(x^*) \le \frac{B^2}{2\eta_T T} + \frac{\rho^2}{2T}\sum_{t=1}^{T}\eta_t$$
$$\xrightarrow{\eta_T=\frac{\alpha}{\sqrt{T}}} f(\bar{x}_t) - f(x^*) \le \frac{B^2}{2\alpha\sqrt{T}} + \frac{\rho^2\alpha}{\sqrt{T}}$$
$$\xrightarrow{opt.\ alpha} \frac{\partial}{\partial\alpha}(\frac{B^2}{2\alpha\sqrt{T}} + \frac{\rho^2\alpha}{\sqrt{T}}) = 0 \to (-\frac{B^2}{2\alpha^2\sqrt{T}} + \frac{\rho^2}{\sqrt{T}}) = 0 \to \alpha = \frac{\sqrt{2}B}{\rho}$$

17 Theory Question

با توجه به اینکه X از رابطه $x^{t+1} = \Pi_c(x^t - \eta_t v^t)$ بدست می‌آید و از آنجاییکه در هر مرحله v^t از یک توزیع احتمالاتی بدست می‌آید و مقداری معین نیست، X بردار تصادفی است.

18 Theory Question

در بخش قبل به الگوریتم کاهش گرادیان مقید در حالتی که تابع هدف شامل متغیر تصادفی نیست رسیدیم. از آنجاییکه به دنبال روشی برای حل مسائل دارای متغیر تصادفی هستیم، منطقی است که به جای استفاده از زیرگرادیان‌های معلوم، به کمک اطلاعاتی که از متغیر تصادفی داریم(امید ریاضی) مقدار زیرگرادیان را حدس بزنیم.

19 Theory Question

$\mathrm{E}[\mathrm{v}^t|x^t] \in \partial f(x^t) \to f(x^*) \geq f(x^t) + \langle x^* - x^t, E[v^t|x^t]\rangle$

$\overset{\eta_t>0}{\to} -\eta_t((f(x^t) - f(x^*)) + \langle x^t - x^*, E[v^t|x^t]\rangle) > 0$

حال کافی است ثابت کنیم:

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

برهان خلف : اگر رابطه بالا برقرار نباشد، با توجه به رابطه $x^{t+1} = \Pi_c(x^t - \eta_t v^t)$ می‌توان گفت x^* نزدیک‌ترین نقطه به $x^t - \eta_t v^t$ درون C است پس $x^{t+1} = x^*$ و در نتیجه باید داشته باشیم :

$$\|(x^t - \eta_t v^t) - x^*\|_2^2 < 0$$

از آنجاییکه نامساوی فوق ناممکن است فرض خلف باطل و حکم برقرار است.

20 Theory Question

$\mathrm{E}[\langle \varepsilon_t, X^t - x^*\rangle] = E[(V^t)^T X^t - (V^t)^T x^* - E[V^t|X^t]^T X^t + E[V^t|X^t]^T x^*] =$
$\mathrm{E}[\mathrm{V}^t X^t] - E[(V^t)^T]x^* - E[E[V^t|X^t]^T X^t] + E[V^t]^T x^* =$
$\mathrm{E}[(\mathrm{V\text{-}E[V]})^T(X - E[X])] - E[(E[V|X] - E[E[V|X]])^T(X - E[X])] =$
$\mathrm{E}[(\mathrm{V\text{-}E[V]})^T(X - E[X])] - E[(E[V|X] - E[V])^T(X - E[X])] = 0$

21 Theory Question

$\frac{1}{2}\|X^{t+1} - x^*\|_2^2 \leq \frac{1}{2}\|X^t - x^*\|_2^2 - \eta_t(f(X^t) - f(x^*)) + \frac{\eta_t^2}{2}\|V^t\|_2^2 - \eta_t\langle \varepsilon_t, X - x^*\rangle$

$\xrightarrow{\cdot\frac{1}{\eta_t}} \frac{1}{2\eta_t}\|X^{t+1} - x^*\|_2^2 \leq \frac{1}{2\eta_t}\|X^t - x^*\|_2^2 - (f(X^t) - f(x^*)) + \frac{\eta_t}{2}\|V^t\|_2^2 - \langle \varepsilon_t, X - x^*\rangle$

$\xrightarrow[\text{1.4,2.4 } \textit{theorems}]{\sum_{t=1}^T,\ Q_{16}\ \textit{proof}} f(\bar{X}^t) - f(x^*) \leq \frac{B^2}{2T\eta_T} + \frac{\rho^2}{2T}\sum_{t=1}^T \eta_t - \frac{1}{T}\sum_{t=1}^T \eta_t\langle \varepsilon_t, X_t - x^*\rangle$

$\xrightarrow[Q_{20}]{E[]} E[f(\bar{X}^t)] - f(x^*) \leq \frac{B^2}{2T\eta_T} + \frac{\rho^2}{2T}\sum_{t=1}^T \eta_t$

22 Theory Question

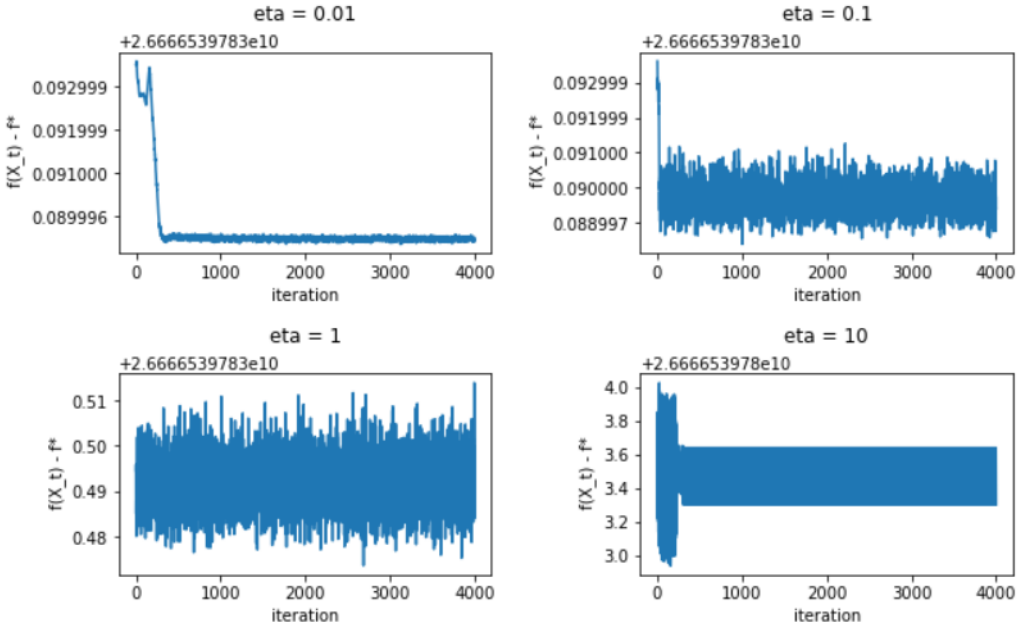
23 Theory Question

24 Theory Question

v$\in \partial F(x,z) \xrightarrow[course\ \ slides]{^{.p(x)>0}} p(x)v \in \partial(p(x)F(x))$
$\xrightarrow{\int} E[V] \in \partial E[F] = \partial f(X) \rightarrow v\ is\ random\ subgradient.$

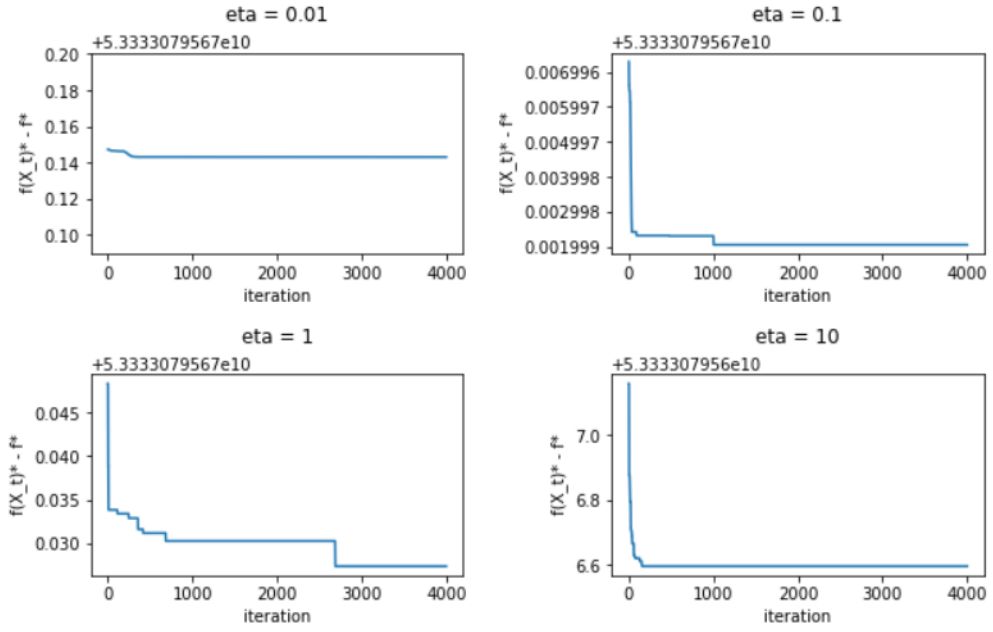
Simulation

1.



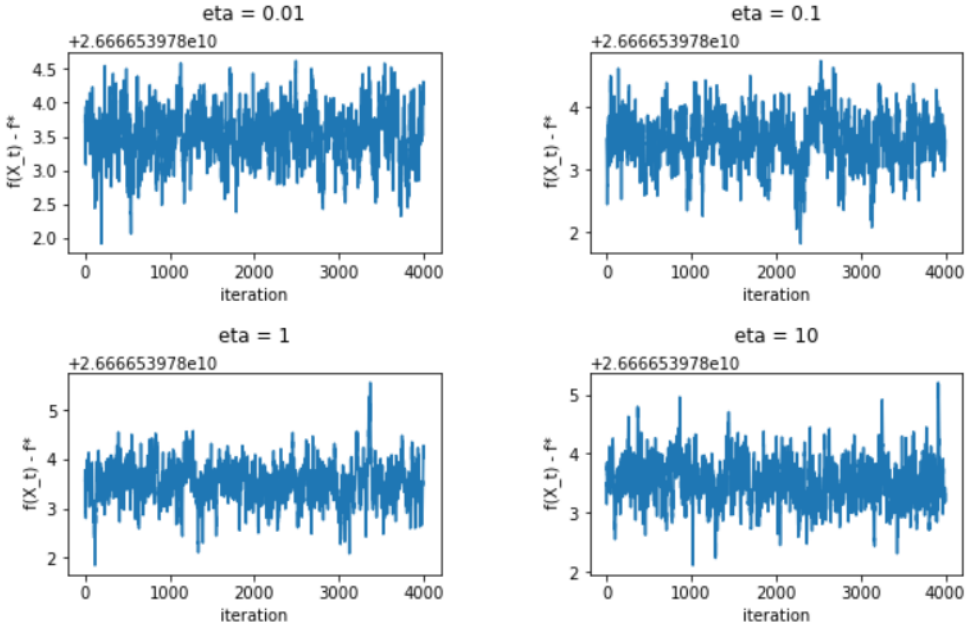
for small η error converge to zero.

2.



error converge to zero faster than part 1.

3.

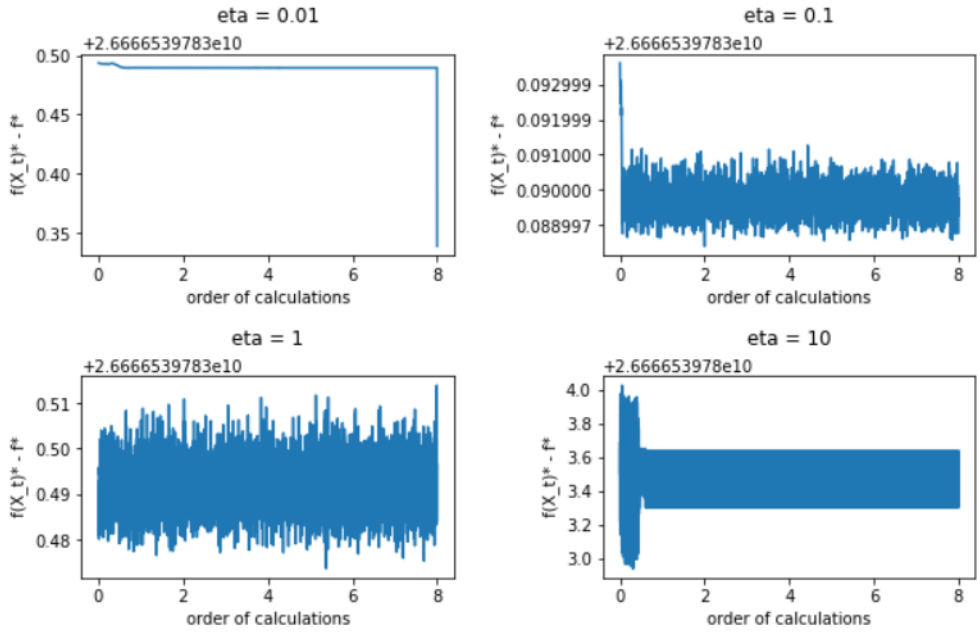


it doesn't converge while in part 1 the error converges to zero.

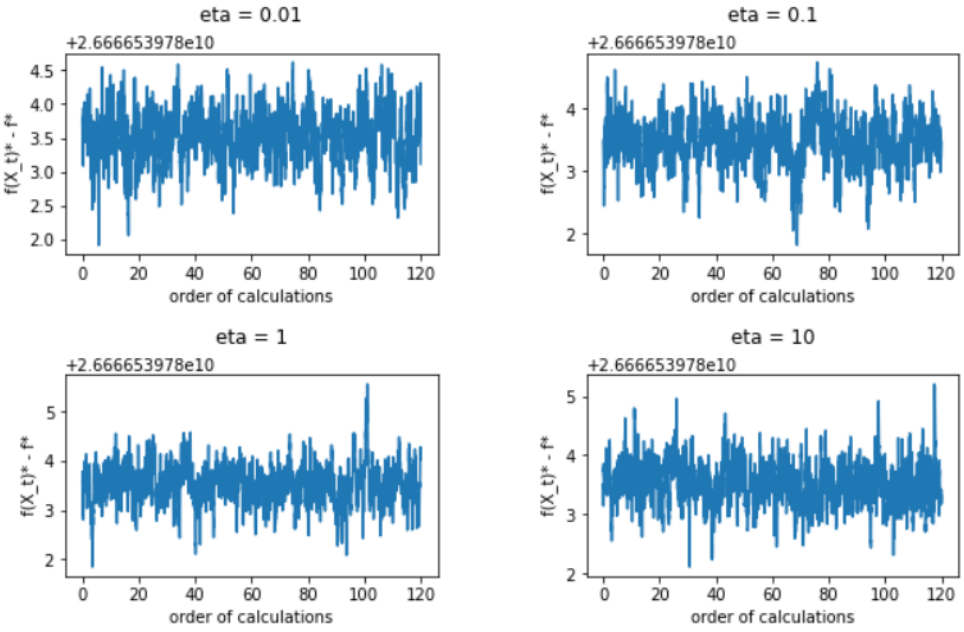
4.

```
Elapsed time: 0.0020360946655273438 seconds
Elapsed time: 0.029249191284179688 seconds
```

numbers above are the times spent for an iteration of each algorithm.



error over calculation order plot for algorithm 2.



error over calculation order plot for algorithm 4.
based on plots above, it is obvious that algorithm 4 takes much more calculation power compared to algorithm 2.
rest of the questions are explained in notebook.