Convex Optimization — Coordinate Descent Method

1. Problem 1.

(a). The solution to $Ax=b$ can be equivalently related to the optimizer of the problem. $\min_x \frac{1}{2}x^TAx - b^Tx := f(x)$.

For $A \in S^n_{++}$, we decompose it as $A = D+E+F$

where D is a diagonal matrix, E is strict lower triangle and F is strict upper triangle matrix.

Hence, the modified Gauss-Seidel update can be written in vector form.

$$x^{k+1} = x^k - \gamma D^{-1}\left[Dx^k + Ex^{k+1} + Fx^k - b\right].$$

$$\Rightarrow (I+\gamma D^{-1}E)x^{k+1} = \left[(1-\gamma)I - \gamma D^{-1}F\right]x^k + \gamma D^{-1}b.$$

$$\Rightarrow (D+\gamma E)x^{k+1} = \left[D+\gamma E - \gamma(E+F+D)\right]x^k + \gamma b$$

$$= (D+\gamma E)x^k - \gamma\cdot(Ax^k - b)$$

$$= (D+\gamma E)x^k - \gamma\cdot r^k$$

$$\Rightarrow x^{k+1} = x^k - \left(\frac{D}{\gamma}+E\right)^{-1} r^k.$$

therefore, $f(x^{k+1}) - f(x^k) = \frac{1}{2}\left[x^k - \left(\frac{D}{\gamma}+E\right)^{-1}r^k\right]^T A\left[x^k - \left(\frac{D}{\gamma}+E\right)^{-1}r^k\right]$

$$-\frac{1}{2}(x^k)^TAx^k - b^T(x^{k+1}-x^k)$$

$$= -\frac{1}{2}\left[\left(\frac{D}{\gamma}+E\right)^{-1}r^k\right]^TAx^k\cdot 2 + \frac{1}{2}(r^k)^T\left(\frac{D^T}{\gamma}+E^T\right)^{-1}A\left(\frac{D}{\gamma}+E\right)^{-1}r^k$$

$$+ b^T\left(\frac{D}{\gamma}+E\right)^{-1}r^k$$

$$= \left[\left(\frac{D}{\gamma}+E\right)^{-1}r^k\right]^T\left[b - Ax^k\right] + \frac{1}{2}(r^k)^T\left(\frac{D^T}{\gamma}+E^T\right)^{-1}A\left(\frac{D}{\gamma}+E\right)^{-1}r^k$$

$$= -(r^k)^T\cdot\left(\frac{D^T}{\gamma}+E^T\right)^{-1}\cdot r^k + \frac{1}{2}(r^k)^T\left(\frac{D^T}{\gamma}+E^T\right)^{-1}A\left(\frac{D}{\gamma}+E\right)^{-1}r^k$$

$$= \frac{1}{2}(r^k)^T\left[\left(\frac{D}{\gamma}+E\right)^{-T}A\left(\frac{D}{\gamma}+E\right)^{-1} - 2\left(\frac{D}{\gamma}+E\right)^{-T}\right]r^k.$$

we need to show $(r_k)^T\left[\left(\frac{D}{\gamma}+E\right)^{-T}A\left(\frac{D}{\gamma}+E\right)^{-1} - 2\left(\frac{D}{\gamma}+E\right)^{-T}\right](r_k) > 0$ if $\gamma \notin (0,2)$ $\forall\, r_k \neq 0$.

Let $W = \frac{D}{\gamma} + E$. $\Rightarrow$ $W^T = \frac{D}{\gamma} + F$.

$\Rightarrow$

$$A - W - W^T = (1 - \frac{2}{\gamma}) D.$$

$$(W^T)^{-1} [A - W - W^T] W^{-1} = (1 - \frac{2}{\gamma}) \cdot (W^T)^{-1} \sqrt{D} \cdot \sqrt{D} \cdot W^{-1}$$

$\Rightarrow$

$$(W^{-1})^T A W^{-1} - (W^T)^{-1} - W^{-1} = (1 - \frac{2}{\gamma}) \cdot (D^{1/2} W^{-1})^T \cdot D^{1/2} W^{-1}.$$

Besides, we have

$$(r^k)^T W^{-1} r^k = (r^k)^T (W^T)^{-1} r^k.$$

Therefore.

$$f(x^{k+1}) - f(x^k) = \frac{1}{2} (1 - \frac{2}{\gamma}) \cdot (r_\gamma^k)^T (D^{1/2} W^{-1})^T D^{1/2} W^{-1} r^k$$

$$= \frac{1}{2} (1 - \frac{2}{\gamma}) \cdot \| D^{1/2} W^{-1} r^k \|_2^2.$$

If $\gamma \notin (0, 2)$. then. $\forall\, r^k \neq 0$.

$$f(x^{k+1}) > f(x^k).$$

Therefore. the sequence doesn't converge. if $x_0 \neq x^*$. □.

(b). $f(x_1, x_2) = \max((x_1 - 1)^2 + (x_2 + 1)^2,\ (x_1 + 1)^2 + (x_2 - 1)^2)$.

First, notice that f is the maximum of convex functions.

hence f is also convex. Any local optimum is global optimum.

Assume. (x_1, x_2) is an optimizer.

$$f(-x_1, -x_2) = \max((-x_1 - 1)^2 + (-x_2 + 1)^2,\ (-x_1 + 1)^2 + (-x_2 - 1)^2)$$

$$= \max((x_1 + 1)^2 + (x_2 - 1)^2,\ (x_1 - 1)^2 + (x_2 + 1)^2)$$

$$= f(x_1, x_2)$$

$\Rightarrow$ $(-x_1, -x_2)$ is also an optimizer.

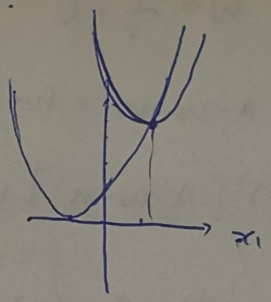
Therefore $0 = \frac{1}{2}(x_1, x_2) + \frac{1}{2}(-x_1, -x_2)$ is always an optimizer.

Besides. $f_1(x) = (x_1 - 1)^2 + (x_2 + 1)^2$, $f_2(x) = (x_1 + 1)^2 + (x_2 - 1)^2$ are strong convex functions.

$\Rightarrow$ ~~$f(x) \geq f(x)$ is also~~

$f(x)$ is also strong convex function. $\Rightarrow$ Unique minimizer.

Therefore, the only minimizer is $x^* = (0,0)^T$.



If we start from $x^0 = (1,1)$.

$$x_1^1 \in \arg\min_{x_1} \max\left((x_1-1)^2+4,\ (x_1+1)^2\right)$$

$$\in \arg\min_{x_1} \begin{cases} (x_1-1)^2+4, & x \le 1 \\ (x_1+1)^2, & x > 1 \end{cases}$$

$$= 1.$$

$$x_2^1 \in \arg\min_{x_2} \max\left((x_2+1)^2,\ 4+(x_2-1)^2\right)$$

$$= 1.$$

Therefore. $(1,1)$ is a fixed point of the Gauss-Seidel algorithm. this doesn't converge to $(0,0)$. □

2. Problem 2

(a) $$\max_z \ b^T z - \frac{1}{2}\|A^T z\|_2^2.$$

$z \in \mathbb{R}^m$. denote $z^k = (z_1^k, z_2^k, \dots, z_m^k)^T$.

$$f(z) = b^T z - \frac{1}{2}\|A^T z\|_2^2 = b^T z - \frac{1}{2} z^T A A^T z.$$

$$\nabla f(z) = b - AA^T z \in \mathbb{R}^m.$$

Therefore

$$z_1^{k+1} = b_1 - a_1^T A^T [z_1^k, z_2^k, \dots, z_m^k]^T$$

$$z_2^{k+1} = b_2 - a_2^T A^T [z_1^{k+1}, z_2^k, \dots, z_m^k]^T$$

$$\vdots$$

$$z_i^{k+1} = b_i - a_i^T A^T [z_1^{k+1}, z_2^{k+1}, \dots z_{i-1}^{k+1}, z_i^k, \dots z_m^k].$$

According to coordinate descent.

$$x_i^{k+1} \in \arg\min_{x_i} f(x_1^{k+1}, \dots x_{i-1}^{k+1}, x_i, x_{i+1}^k, \dots x_m^k).$$

For each iteration i, we have z_i satisfying that

$$\nabla_i f(z)_i = b_i - [AA^T z]_i = 0. \quad b_i - a_i^T A^T z$$

where $\tilde{z} = [\, z_1^{k+1}, \cdots z_{i-1}^{k+1}, z_i^k, z_{i+1}^k, \cdots z_m^k \,]^T$

$$\Rightarrow \quad \nabla_i f(\tilde{z}) = b_i - a_i^T A^T \tilde{z}$$

$$= b_i - \sum_{j=1}^{i-1} a_i^T a_j z_j^{k+1} - \sum_{j=i}^{m} a_i^T a_j z_j^k$$

$$z_i^{k+1} = z_i^k + \nabla_i f(\tilde{z}) \qquad \text{(Because it's a maximum problem, and we assume stepsize is 1)}$$

$$= b_i - \sum_{j=1}^{i-1} a_i^T a_j z_j^{k+1} - \sum_{j=i+1}^{m} a_i^T a_j z_j^k$$

therefore, the update rule is

$$z_i^{k+1} = b_i - \sum_{j=1}^{i-1} a_i^T a_j z_j^{k+1} - \sum_{j=i+1}^{m} a_i^T a_j z_j^k \qquad (2)$$

For this update, we have $\hat{z} = [\, z_1^{k+1}, \cdots z_{i-1}^{k+1}, z_i^{k+1}, z_{i+1}^k, \cdots z_m^k \,]^T$

~~$a_i^T z_i^{k+1} = a_i^T z_i^k$~~

With this new $\hat{z}$.

$$\nabla_i f(\hat{z}) = b_i - a_i^T A^T \hat{z} \qquad (1).$$

$$= b_i - \sum_{j=1}^{i} a_i^T a_j z_j^{k+1} - \sum_{j=i+1}^{m} a_i^T a_j z_j^k - z_i^{k+1}$$

$$= 0.$$

therefore $\quad a_i^T (A^T \hat{z}) = b_i .$

(b) for the primal problem. $\min_x \frac{1}{2}\|x\|_2^2$ s.t $Ax=b$

$$L(x,z) = \frac{1}{2}\|x\|_2^2 + z^T(Ax-b)$$

$$d(z) = \inf_x L(x,z) = \inf_x \frac{1}{2}x^Tx + z^TAx - z^Tb.$$

$$\Rightarrow \quad x + (z^TA)^T = 0 \quad \Rightarrow \quad A^Tz + x = 0$$

therefore. $d(z) = \frac{1}{2}\|A^Tz\|_2^2 - z^TAA^Tz - z^Tb$

$$= -\frac{1}{2}\|A^Tz\|_2^2 - z^Tb.$$

$$d(z') \overset{z'=-z}{=} -\frac{1}{2}\|A^Tz'\|_2^2 + b^Tz'$$

$\Rightarrow$ dual problem is $\max d(z') = b^Tz' - \frac{1}{2}\|A^Tz'\|_2^2$.
s.t. $z' \in \mathbb{R}$.

therefore, the relation between primal variable x and the dual variable is that. $A^Tz' - x = 0.$ (3)

After each update from $z_i^k \to z_i^{k+1}$. with the new $\tilde{z}$.. we update x.

From (1) and (3) we know.
At iteration i. the constraint $b_i = a_i^Tx$ is satisfied. □

(c)

$$g^k = \nabla_{q_t}(y^k) = (a_t^T y^k - b_t)\, a_t.$$

Hence, the update rule for y^k is:

$$y^{k+1} = y^k - (a_t^T y^k - b_t) a_t \quad \text{if we assume step size is 1.}$$

Notice that, if we choose $t = \text{mod}(k, m) + 1$, starting with $k = 0$.

After each update of y^k to y^{k+1}.

we have

$$\begin{aligned} a_t^T y^{k+1} &= a_t^T (y^k - (a_t^T y^k - b_t) a_t) \\ &= a_t^T y^k - a_t^T a_t (a_t^T y^k - b_t) \\ &= b_t. \end{aligned}$$

which means, at iteration $t = \text{mod}(k, m) + 1$,

the constraint $a_t^T y = b_t$ is satisfied.

Which is exactly what we have for primal variable's update property in (b).

In (b), the order is $i = 1, 2, \cdots m, 1, 2, \ldots$ with $a_i^T x = b_i$ satisfied.

which is exactly matching $t = \text{mod}(k, m) + 1$. while k is update ~~index for~~ time counter for dual variable z (Each update in (2) counts once).

Therefore, SGD with $t = \text{mod}(k, m) + 1$ and unit stepsize has the same update rule as part (b) for primal variable. □