HW5

2020年11月25日 4:46

HW5Qu…

MATH 3332 Data Analytic Tools
Homework 5

Due date: 30 November, 6pm, Monday

1. Consider the following optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{B}\boldsymbol{x}\|_2^2,$$

where $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\boldsymbol{b}\in\mathbb{R}^m$, and $\boldsymbol{B}\in\mathbb{R}^{p\times n}$ are given. Find the steepest descent (i.e. gradient descent with the exact line search) algorithm for solving this optimization.

In the following two problems, we consider variants of the gradient descent algorithm for solving

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function. Given $\boldsymbol{x}^{(k)}$, the gradient descent algorithm finds $\boldsymbol{x}^{(k+1)}$ by

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}^{(k)}) + \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle + \frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_2^2, \tag{2}$$

where $\alpha_k > 0$ is a step size.

2. In the gradient descent (2), for distance between $\boldsymbol{x}$ and $\boldsymbol{x}^{(k)}$, we used the 2-norm, which can be replaced by other norms to obtain variants of the gradient descent algorithm.

   (a) Prove that (2) is equivalent to

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle, \quad \text{subject to} \quad \|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_2 \le \alpha_k\|\nabla f(\boldsymbol{x}^{(k)})\|_2. \tag{3}$$

   (*Hint: Use Cauchy-Schwartz inequality.*)

   (b) We may replace the 2-norm in (3) by 1-norm, i.e.,

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle, \quad \text{subject to} \quad \|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_1 \le \alpha_k\|\nabla f(\boldsymbol{x}^{(k)})\|_\infty. \tag{4}$$

   Give an explicit expression of $\boldsymbol{x}^{(k+1)}$ in (4). (*Hint: Use the inequality* $|\langle \boldsymbol{x},\boldsymbol{y}\rangle| \le \|\boldsymbol{x}\|_\infty\|\boldsymbol{y}\|_1$ *with equality attained if* $\boldsymbol{x} = c\cdot\operatorname{sign}(\boldsymbol{y})$ *or* $\boldsymbol{y} = c\cdot\boldsymbol{e}_{i_{\max}(\boldsymbol{x})}$ *for some* $c\in\mathbb{R}$. *Here* $\operatorname{sign}(\cdot)$ *takes the sign of each entry of a vector, and* $i_{\max}(\boldsymbol{x}) = \arg\max_i |x_i|$.)

   (c) We may replace the 2-norm in (3) by ∞-norm, i.e.,

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle, \quad \text{subject to} \quad \|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_\infty \le \alpha_k\|\nabla f(\boldsymbol{x}^{(k)})\|_1. \tag{5}$$

   Give an explicit expression of $\boldsymbol{x}^{(k+1)}$ in (5).

3. Gradient depends on inner product. Therefore, one variant of gradient descent is obtained by choosing a weighted inner product of $\mathbb{R}^n$ rather than the standard one. In particular, given a symmetric positive definite matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$, we know that $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\boldsymbol{W}} := \boldsymbol{x}^T\boldsymbol{W}\boldsymbol{y}$ defines an inner product of $\mathbb{R}^n$.

   (a) Express $\nabla_{\boldsymbol{W}} f$ (the gradient of f in $\mathbb{R}^n$ with the weighted inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{W}}$) in terms of $\boldsymbol{W}$ and ∇f (the gradient of f in $\mathbb{R}^n$ with the standard inner product).

   (b) Give the gradient descent algorithm for solving (1) under the weighted inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{W}}$ (i.e., find an explicit formula of $\boldsymbol{x}^{(k+1)}$ when we replace $\langle\cdot,\cdot\rangle$ and $\|\cdot\|_2^2$ in (2) by $\langle\cdot,\cdot\rangle_{\boldsymbol{W}}$ and $\|\cdot\|_{\boldsymbol{W}}^2$ respectively). (*By choosing a suitable $\boldsymbol{W}$, we may obtain faster algorithms than the standard gradient descent. This technique is known as preconditioning.*)

1. Consider the following optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{B}\boldsymbol{x}\|_2^2,$$

where $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\boldsymbol{b}\in\mathbb{R}^m$, and $\boldsymbol{B}\in\mathbb{R}^{p\times n}$ are given. Find the steepest descent (i.e. gradient descent with the exact line search) algorithm for solving this optimization.

In the following two problems, we consider variants of the gradient descent algorithm for solving

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function. Given $\boldsymbol{x}^{(k)}$, the gradient descent algorithm finds $\boldsymbol{x}^{(k+1)}$ by

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}^{(k)}) + \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle + \frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_2^2, \tag{2}$$

where $\alpha_k > 0$ is a step size.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|Bx\|_2^2$.

We need collary 1 & 2 here to solve $\nabla f(x)$.

Collary 1: Let $\varphi(x) = c^Tx$, $\nabla\varphi(x) = c$

Collary 2: Let $\phi(x) = x^TBx$, $\nabla\phi(x) = (B+B^T)x$

Proving collary 1:

$$\varphi(x) = \sum_{j=1}^{n} c_j x_j$$

$$\frac{\partial\varphi}{\partial x_k} = \sum_{j=1}^{n} c_j \frac{\partial x_j}{\partial x_k} = \sum_{j=1}^{n} c_j \delta_{jk} = c_k$$

$$\Rightarrow \nabla\varphi(x) = c.$$

Proving collary 2:

$$\phi(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} x_i x_j$$

$$\frac{\partial\phi}{\partial x_k} = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} \frac{\partial(x_i x_j)}{\partial x_k}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}(\delta_{ik}x_j + x_i\delta_{jk})$$

$$= \sum_{j=1}^{n} b_{kj}x_j + \sum_{i=1}^{n} b_{ik}x_i$$

$$= (Bx)_k + \sum_{i=1}^{n}(B^T)_{ki}x_i$$

$$= (Bx)_k + (B^Tx)_k$$

$$\therefore \nabla\phi(x) = (B+B^T)x.$$

$$\frac{\partial}{\partial x}\|Ax-b\|_2^2 = \frac{\partial}{\partial x}[(Ax-b)^T(Ax-b)]$$

Consider $\|Ax-b\|_2^2 = [(Ax-b)^T(Ax-b)]$

$\|Ax-b\|_2^2 = [x^TA^TAx - (Ax)^Tb - b^TAx + b^Tb]$

$\|Ax-b\|_2^2 = [x^TA^TAx - 2b^TAx + b^Tb]$

$\|Ax-b\|_2^2 = [x^TA^TAx - 2(A^Tb)^Tx + b^Tb]$

Consider $\|Bx\|^2$,

$\|Bx\|_2^2 = (Bx)^T(Bx)$

$= x^TB^TBx = x^T(B^TB)^Tx$

=> Putting A^TA in collary ③, putting A^Tb in collary ①,

we can get $\nabla(\|Ax-b\|_2^2) = (A^TA + (A^TA)^T)x - 2A^Tb$

and $(A^TA)^T = A^T(A^T)^T = A^TA$.

=> $\nabla(\|Ax-b\|_2^2) = 2A^TAx - 2A^Tb$

=> Putting B^TB in collary ②, we can get $\nabla(\|Bx\|_2^2)$

$\nabla(\|Bx\|_2^2) = (B^TB + (B^TB)^T)x$

$= (B^TB + B^TB)x$

$= 2(B^TB)x$

$\therefore \nabla(f(x)) = \frac{1}{2}(2A^TAx - 2A^Tb) + 2\alpha(B^TB)x$

$= A^TAx - A^Tb + 2\alpha(B^TB)x$

$= (A^TA + 2\alpha(B^TB))x - A^Tb$

Also, it is obvious to check that $f(x)$ is strictly convex and smooth function. since 2-norm function is continuous, strictly convex coercive, thus $f(x)$ is also continuous, strictly convex and coercive.

$\therefore x^{(k+1)} = \arg\min_{x \in \mathbb{R}^n} f(x) \Leftrightarrow \nabla f(x^{(k+1)}) = 0$

$\nabla f(x^{(k+1)}) = 0 \Leftrightarrow \nabla f(x^{(k)}) + \frac{1}{\alpha_k}(x^{(k+1)} - x^{(k)}) = 0$

$$\nabla f(x^{(k+1)}) = 0 \iff \nabla f(x^{(k)}) + \frac{1}{\alpha_k}(x^{(k+1)} - x^{(k)}) = 0$$

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

also, $\min_{x \in \mathbb{R}^n} f(x)$ has a unique solution

Next, we want to perform exact line search,

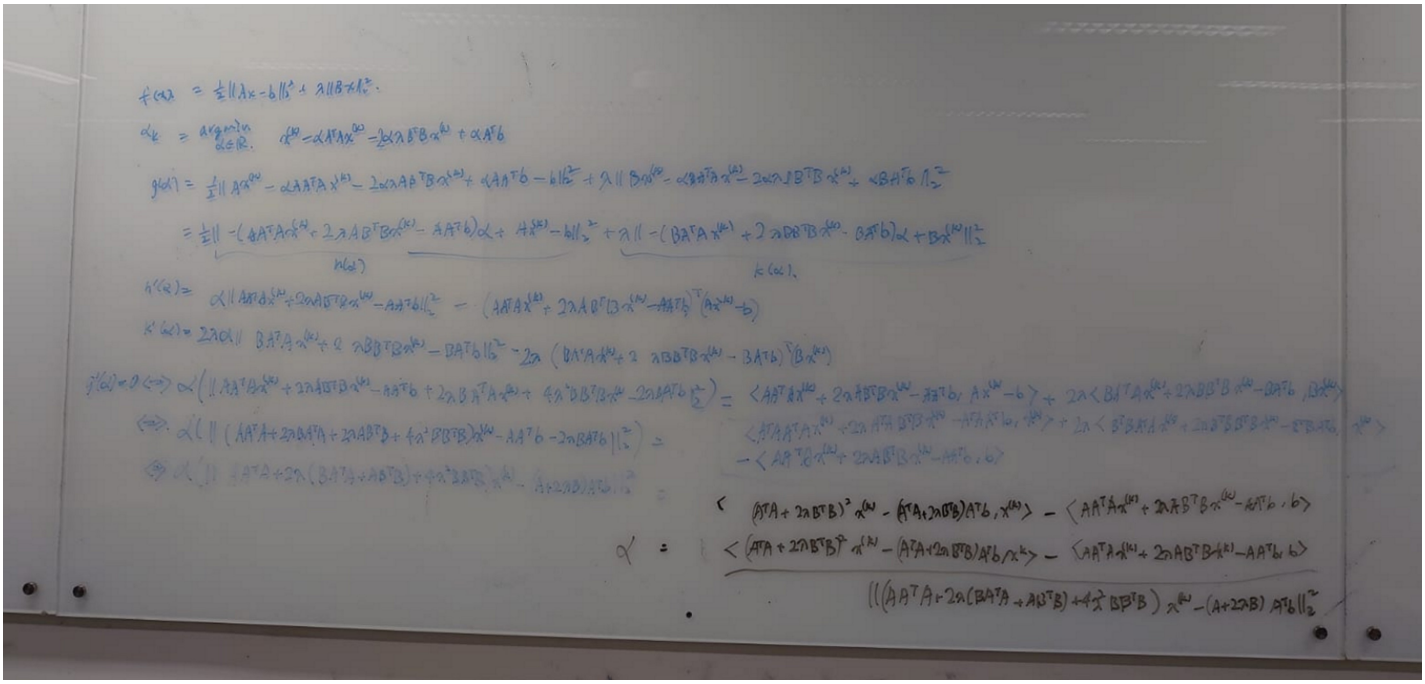
$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f\left(x^{(k)} - \alpha\left(A^TA + 2\lambda B^TB\right)x^{(k)} - A^Tb\right)$$

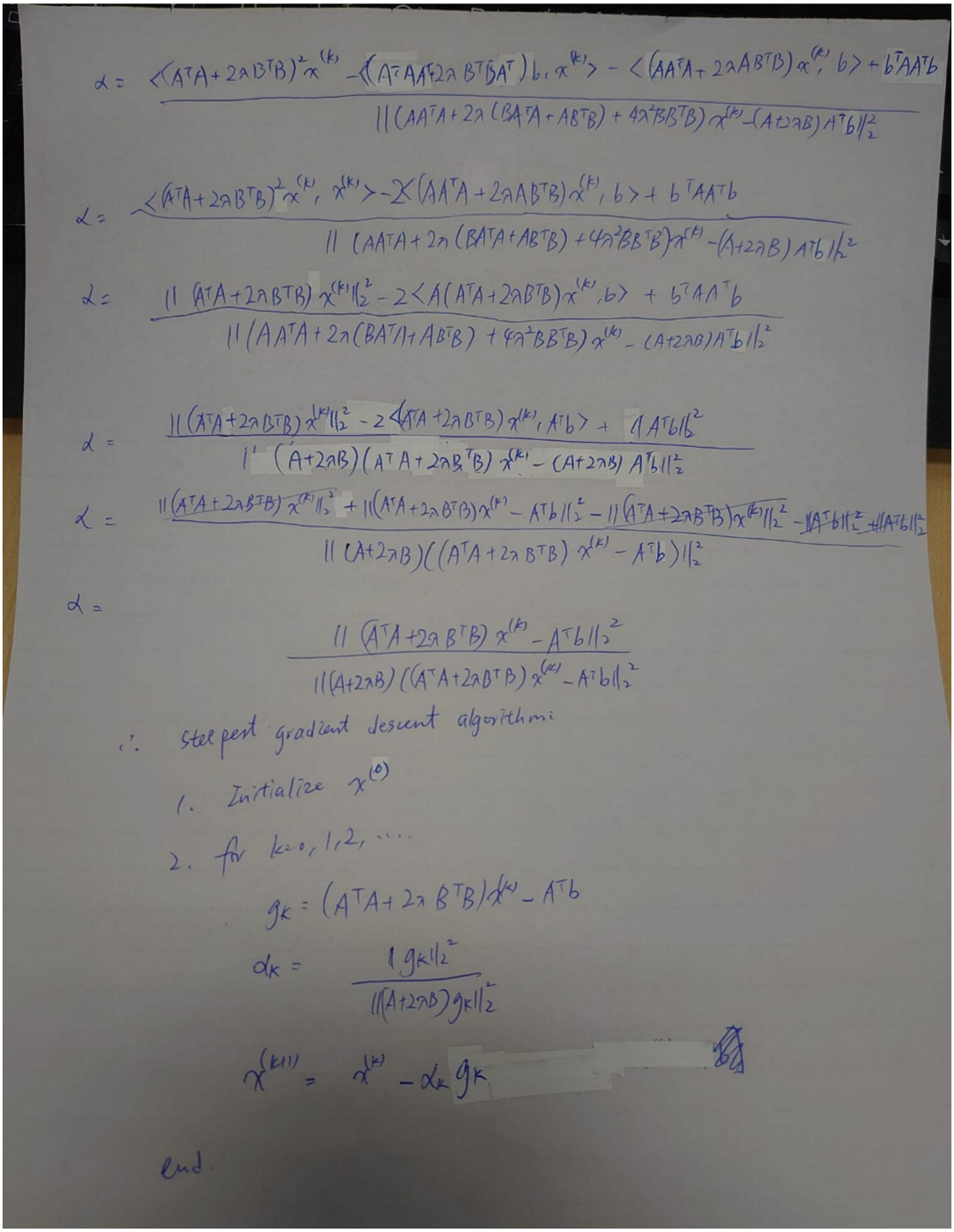
Let $g(\alpha) = f\left(x^{(k)} - \alpha\left(\left(A^TA + 2\lambda B^TB\right)x^{(k)} - A^Tb\right)\right)$

It is obvious that $g(\alpha)$ is convex since 2-norm function is convex.

therefore we want to solve $g'(\alpha_k) = 0$,

recall $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|Bx\|_2^2$,





$\alpha = \dfrac{\langle (A^TA+2\lambda B^TB)^2 x^{(k)} - ((A^TAA^T+2\lambda B^TBA^T)b, x^{(k)}\rangle - \langle (AA^TA+2\lambda ABB^TB)x^{(k)}, b\rangle + b^TAA^Tb}{\|(AA^TA+2\lambda(BA^TA+AB^TB)+4\lambda^2BB^TB)x^{(k)}-(A+2\lambda B)A^Tb\|_2^2}$

$\alpha = \dfrac{\langle (A^TA+2\lambda B^TB)^2 x^{(k)}, x^{(k)}\rangle - 2\langle (AA^TA+2\lambda ABB^TB)x^{(k)}, b\rangle + b^TAA^Tb}{\|(AA^TA+2\lambda(BA^TA+AB^TB)+4\lambda^2BB^TB)x^{(k)}-(A+2\lambda B)A^Tb\|_2^2}$

$\alpha = \dfrac{\|(A^TA+2\lambda B^TB)x^{(k)}\|_2^2 - 2\langle A(A^TA+2\lambda B^TB)x^{(k)}, b\rangle + b^TAA^Tb}{\|(AA^TA+2\lambda(BA^TA+AB^TB)+4\lambda^2BB^TB)x^{(k)}-(A+2\lambda B)A^Tb\|_2^2}$

$\alpha = \dfrac{\|(A^TA+2\lambda B^TB)x^{(k)}\|_2^2 - 2\langle (A^TA+2\lambda B^TB)x^{(k)}, A^Tb\rangle + \|A^Tb\|_2^2}{\|(A+2\lambda B)(A^TA+2\lambda B^TB)x^{(k)}-(A+2\lambda B)A^Tb\|_2^2}$

$\alpha = \dfrac{\|(A^TA+2\lambda B^TB)x^{(k)}\|_2^2 + \|(A^TA+2\lambda B^TB)x^{(k)}-A^Tb\|_2^2 - \|(A^TA+2\lambda B^TB)x^{(k)}\|_2^2 - \|A^Tb\|_2^2 + \|A^Tb\|_2^2}{\|(A+2\lambda B)((A^TA+2\lambda B^TB)x^{(k)}-A^Tb)\|_2^2}$

$\alpha = \dfrac{\|(A^TA+2\lambda B^TB)x^{(k)}-A^Tb\|_2^2}{\|(A+2\lambda B)((A^TA+2\lambda B^TB)x^{(k)}-A^Tb)\|_2^2}$

∴ steepest gradient descent algorithm:

1. Initialize $x^{(0)}$
2. for $k=0,1,2,\cdots$

$g_k = (A^TA+2\lambda B^TB)x^{(k)} - A^Tb$

$\alpha_k = \dfrac{\|g_k\|_2^2}{\|(A+2\lambda B)g_k\|_2^2}$

$x^{(k+1)} = x^{(k)} - \alpha_k g_k$

end.

2. In the gradient descent (2), for distance between $\boldsymbol{x}$ and $\boldsymbol{x}^{(k)}$, we used the 2-norm, which can be replaced by other norms to obtain variants of the gradient descent algorithm.

   (a) Prove that (2) is equivalent to

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle, \quad \text{subject to} \quad \|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_2 \le \alpha_k \|\nabla f(\boldsymbol{x}^{(k)})\|_2. \tag{3}$$

   (*Hint: Use Cauchy-Schwartz inequality.*)

$$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}^{(k)}) + \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{x}-\boldsymbol{x}^{(k)}\rangle + \frac{1}{2\alpha_k}\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_2^2, \tag{2}$$

Q2a). We need to show that $x = x^{(k)} - \alpha \nabla f(x^{(k)})$ is the optimal solution to

$$\min_{x \in \mathbb{R}^n} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \text{ subject to } \|x - x_k\|_2 \leq \alpha_k \|\nabla f(x^{(k)})\|_2 \quad -(*)$$

Using cauchy-schwartz inequality we have

$$|\langle \alpha_k \nabla f(x^{(k)}), x - x^{(k)} \rangle| \leq \alpha_k \|\nabla f(x^{(k)})\|_2 \|x - x_k\|_2 \leq \alpha_k^2 \|\nabla f(x^{(k)})\|_2^2$$

$$\Rightarrow \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \geq -\alpha_k \|\nabla f(x^{(k)})\|_2^2$$

$$\Rightarrow \langle \nabla f(x^{(k)}), x \rangle \geq \langle \nabla f(x^{(k)}), x^{(k)} \rangle - \alpha_k \|\nabla f(x^{(k)})\|_2^2$$

Since equality occurs only if $\alpha_k \nabla f(x^{(k)}) = x^{(k)} - x$

$\therefore$ The optimal solution to $(*)$ is $x = x^{(k)} - \alpha_k \nabla f(x^{(k)})$, the given gradient descent algorithm.

b).

$$x^{(k+1)} = \operatorname{argmin}_{x \in \mathbb{R}^n} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \text{ subject to } \|x - x^{(k)}\|_1 \leq \alpha_k \|\nabla f(x^{(k)})\|_\infty$$

$$|\langle \alpha_k \nabla f(x^{(k)}), x - x^{(k)} \rangle| \leq \alpha_k \|\nabla f(x^{(k)})\|_\infty \|x - x^{(k)}\|_1 \leq \alpha_k^2 \|\nabla f(x^{(k)})\|_\infty^2$$

with equality attained if $x - x^{(k)} = c \cdot e_{i_{\max}(\nabla f(x^{(k)}))}$ for some $c \in \mathbb{R}$.

$$\Rightarrow \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \geq -\alpha_k \|\nabla f(x^{(k)})\|_\infty^2$$

$$\Rightarrow \langle \nabla f(x^{(k)}), x \rangle \geq \langle \nabla f(x^{(k)}), x^{(k)} \rangle - \alpha_k \|\nabla f(x^{(k)})\|_\infty^2$$

Since equality occurs only if $x - x^{(k)} = c \cdot e_{i_{\max}(\nabla f(x^{(k)}))}$ for some $c \in \mathbb{R}$,

$$x^{(k+1)} = x^{(k)} + c \cdot e_{i_{\max}(\nabla f(x^{(k)}))} \text{ for some } c \in \mathbb{R}.$$

c). Similarly, $|\langle \alpha_k \nabla f(x^{(k)}), x - x^{(k)} \rangle| \leq \alpha_k \|\nabla f(x^{(k)})\|_1 \|x - x^{(k)}\|_\infty \leq \alpha_k^2 \|\nabla f(x^{(k)})\|_1$

$$\Rightarrow \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \geq -\alpha_k \|\nabla f(x^{(k)})\|_1^2$$

$$\Rightarrow \langle \nabla f(x^{(k)}), x \rangle \geq \langle \nabla f(x^{(k)}), x^{(k)} \rangle - \alpha_k \|\nabla f(x^{(k)})\|_1^2$$

Since equality occurs only if $x - x^{(k)} = c \cdot \operatorname{sign}(\nabla f(x^{(k)}))$ for some $c \in \mathbb{R}$,

$$x^{(k+1)} = x^{(k)} + c \cdot \operatorname{sign}(\nabla f(x^{(k)})) \text{ for some } c \in \mathbb{R}.$$

3. Gradient depends on inner product. Therefore, one variant of gradient descent is obtained by choosing a weighted inner product of $\mathbb{R}^n$ rather than the standard one. In particular, given a symmetric positive definite matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$, we know that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{W}} := \boldsymbol{x}^T \boldsymbol{W} \boldsymbol{y}$ defines an inner product of $\mathbb{R}^n$.

   (a) Express $\nabla_{\boldsymbol{W}} f$ (the gradient of f in $\mathbb{R}^n$ with the weighted inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{W}}$) in terms of $\boldsymbol{W}$ and ∇f (the gradient of f in $\mathbb{R}^n$ with the standard inner product).

   (b) Give the gradient descent algorithm for solving (1) under the weighted inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{W}}$ (i.e., find an explicit formula of $\boldsymbol{x}^{(k+1)}$ when we replace $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2^2$ in (2) by $\langle \cdot, \cdot \rangle_{\boldsymbol{W}}$ and $\| \cdot \|_{\boldsymbol{W}}^2$ respectively). (*By choosing a suitable $\boldsymbol{W}$, we may obtain faster algorithms than the standard gradient descent. This technique is known as preconditioning.*)

Q3. From the definition of gradient, we have:

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0) - \langle v, x - x_0 \rangle_W|}{\|x - x_0\|_W} = 0$$, where v is gradient $\nabla f(x)$ at $x = x_0$.

Consider $\langle v, x - x_0 \rangle$,

$$\langle v, x - x_0 \rangle_W = v^T W (x - x_0) = \langle v, W(x - x_0) \rangle = \langle W^T v, x - x_0 \rangle$$

If $v = \nabla_W f$, $\nabla f = W^T v$

$$\Leftrightarrow \nabla f = W^T \nabla_W f \Leftrightarrow \nabla_W f = (W^{-1})^T \nabla f$$

b). As W is S.P.D,

$\therefore$ $W = U^T U$ for some matrix U.

$$\|x\|_W^2 = x^T W x = (Ux)^T Ux = \langle Ux, Ux \rangle = \|Ux\|_2^2$$

$$\langle x, y \rangle_W = \langle Wx, y \rangle$$

For gradient descent algorithm, we have:

$$x^{(k+1)} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x^{(k)}) + \langle \nabla_W f(x^{(k)}), x - x^{(k)} \rangle_W + \frac{1}{2\alpha_k} \|x - x^{(k)}\|_W^2$$

$$x^{(k+1)} = \operatorname{argmin}_{x \in \mathbb{R}^n} \langle W^T \nabla_W f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\alpha_k} \|U(x - x^{(k)})\|_2^2$$

let $F(x) = \langle W^T \nabla_W f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\alpha_k} \|U(x - x^{(k)})\|_2^2$

Solving $\nabla F(x) = 0$, we have:

$$W^T \nabla_W f(x^{(k)}) + \frac{1}{2\alpha_k} (2U \cdot (Ux - Ux^{(k)})) = 0$$

$W^T \nabla_W f(x^{(k)}) = W^T (W^T)^{-1} \nabla f = \nabla f$; $U^T U = W$

$$\Leftrightarrow \nabla f(x^{(k)}) + \frac{1}{\alpha_k} W(x - x^{(k)}) = 0$$

$$Wx = Wx^{(k)} - \alpha_k \nabla f(x^{(k)})$$

$$\Leftrightarrow x = x^{(k)} - \alpha_k W^{-1} \nabla f(x^{(k)})$$

$\therefore$ we have $x^{(k+1)} = x^{(k)} - \alpha_k W^{-1} \nabla f(x^{(k)})$ for gradient descent.