MLoN Homework1

Group 5

1 Problem 1

A differentiable function f is μ-strongly convex iff $\forall x_1, x_2 \in \chi, \mu > 0$

$$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2 \tag{1}$$

1.1

Prove Equation 1 is equivalent to a minimum positive curvature

$$\bigtriangledown^2 f(x) \geq \mu \mathrm{I}_d, \forall x \in \chi \tag{2}$$

ⓐ Equation 2 $\Rightarrow$ Equation 1:
Taylor expansion:

$$\begin{aligned} f(x_2) &= f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{1}{2}(x_2 - x_1)^T \bigtriangledown^2 f(x_1)(x_2 - x_1) + R_2(X) \\ &= f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{1}{2}(x_2 - x_1)^T \bigtriangledown^2 f(x_3)(x_2 - x_1) \\ &\geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2 \end{aligned}$$

ⓑ Equation 1 $\Rightarrow$ Equation 2:

$$\begin{aligned} f(x_2) &= f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{1}{2}(x_2 - x_1)^T \bigtriangledown^2 f(x_1)(x_2 - x_1) + R_2(X) \\ &= f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{1}{2}(x_2 - x_1)^T \bigtriangledown^2 f(x_3)(x_2 - x_1) \\ &\geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2 \\ &\Rightarrow \\ &(x_2 - x_1)^T \bigtriangledown^2 f(x_3)(x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2 \\ &\Rightarrow \\ &\bigtriangledown^2 f(x) \geq \mu \mathrm{I}_d, \forall x \in \chi \end{aligned}$$

.
ⓐ & ⓑ $\Rightarrow$ Equation 1 $\equiv$ Equation 2

1.2

Prove Equation 1 is equivalent to

$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2 \tag{3}$$

ⓐ Equation 1 $\Rightarrow$ Equation 3:

$$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2$$
$$f(x_1) \geq f(x_2) + \bigtriangledown f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \|x_1 - x_2\|_2^2$$
$$\Rightarrow$$
$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$

ⓐ Equation 3 $\Rightarrow$ Equation 1:

$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$
$$\Rightarrow$$
$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \geq (x_2 - x_1)^T \mu (x_2 - x_1)$$
$$\Rightarrow$$
$$[(\bigtriangledown f(x_2) - \mu x_2) - (\bigtriangledown f(x_1) - \mu x_1)]^T (x_2 - x_1) \geq 0$$
$$\Rightarrow$$
$$(\bigtriangledown g(x_2) - \bigtriangledown g(x_1))^T (x_2 - x_1) \geq 0,$$
$$g(x) = f(x) - \frac{\mu}{2} \|x\|^2$$
$$\Rightarrow$$
$$g(x) \text{ is convex (the monotone gradient condition for convexity).}$$
$$\text{g(x) is convex iff. } (g(x_2) - g(x_1))^T (x_2 - x_1) \geq 0.$$
$$\Rightarrow$$
$$g(x_2) \geq g(x_1) + \bigtriangledown g(x_1)^T (x_2 - x_1)$$
$$\Rightarrow$$
$$f(x_2) - \frac{\mu}{2} \|x_2\|^2 \geq f(x_1) - \frac{\mu}{2} \|x_1\|^2 + (\bigtriangledown f(x_1) - \mu x_1)^T (x_2 - x_1)$$
$$\Rightarrow$$
$$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2\|^2 + \frac{\mu}{2} \|x_1\|^2 - \mu x_1^T x_2$$
$$\Rightarrow$$
$$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2$$

ⓐ & ⓑ $\Rightarrow$ Equation 1 $\equiv$ Equation 3

1.3 a

Implies Polyak-Łojasiewicz (PL) Inequality

$$f(x) - f^* \leq \frac{1}{2\mu} \left\| \bigtriangledown f(x) \right\|_2^2, \forall x \tag{4}$$

Taking minimization respect to x_2 on both sides of Equation 1:

$$\begin{aligned}
& f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \left\| x_2 - x_1 \right\|_2^2 \\
\Rightarrow \\
& f(x^*) \geq f(x_1) + \bigtriangledown f(x_1)^T (x^* - x_1) + \frac{\mu}{2} \left\| x^* - x_1 \right\|_2^2 \\
& (f'(x^*) = 0 \Rightarrow x^* = x_1 - \frac{1}{\mu} \bigtriangledown f(x_1)) \\
\Rightarrow \\
& f(x^*) \geq f(x) - \frac{1}{2\mu} \left\| \bigtriangledown f(x) \right\|_2^2 \\
\Rightarrow \\
& \text{Equation 4}
\end{aligned}$$

1.4 b

Implies

$$\left\| x_2 - x_1 \right\|_2 \leq \frac{1}{\mu} \left\| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \right\|_2, \forall x_1, \forall x_2 \tag{5}$$

Applying Cauchy–Schwarz inequality $|\langle u, v \rangle| \leq \|u\| \, \|v\|$ on the equivalent condition Equation 3.

$$\begin{aligned}
& \left(\bigtriangledown f(x_2) - \bigtriangledown f(x_1) \right)^T (x_2 - x_1) \geq \mu \left\| x_2 - x_1 \right\|_2^2 \\
\Rightarrow \\
& \left\| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \right\|_2 \left\| x_2 - x_1 \right\|_2 \geq \mu \left\| x_2 - x_1 \right\|_2^2 \\
\Rightarrow \\
& \text{Equation 5}
\end{aligned}$$

1.5 c

Implies

$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \leq \frac{1}{\mu} \left\| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \right\|_2^2, \forall x_1, \forall x_2 \tag{6}$$

Multiplying Equation 5 and $\|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2$

$\Rightarrow$

$\|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2 \geq \mu \|x_2 - x_1\|_2 \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2$

Applying Cauchy–Schwarz inequality

$\Rightarrow$

$\|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2 \geq \mu(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T(x_2 - x_1)$

$\Rightarrow$

Equation 6

1.6 d

Implies $f(x) + r(x)$ is strongly convex for any convex f and strongly convex r.

f is convex:

$$f(y) \geq f(x) + \bigtriangledown f(x)^T(x_2 - x_1) \tag{7}$$

r is strongly convex:

$$r(y) \geq r(x) + \bigtriangledown r(x)^T(x_2 - x_1) + \frac{\mu}{2}\|y - x\|^2 \tag{8}$$

$$\begin{aligned} g(x) &= f(x) + r(x) \\ &\geq f(x) + \bigtriangledown f(x)^T(x_2 - x_1) + r(x) + \bigtriangledown r(x)^T(x_2 - x_1) + \frac{\mu}{2}\|y - x\|^2 \\ &= g(x) + \bigtriangledown g(x)^T(x_2 - x_1) + \frac{\mu}{2}\|y - x\|^2 \end{aligned}$$

$\Rightarrow$

$g(x) = f(x) + r(x)$ is strongly convex.

2 Problem 2

A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth iff it is differentiable and its gradient is L-Lipschitz-continuous (usually w.r.t. norm-2):

$$\forall x_1, \forall x_2 \in \mathbb{R}^d, \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2 \leq L \|x_2 - x_1\|_2 \tag{9}$$

2.1 a

Implies

$$f(x_2) \leq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{L}{2} \|x_2 - x_1\|_2^2 \tag{10}$$

f is convex and L-smooth.

$\|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2 \leq L \|x_2 - x_1\|_2$

$\Rightarrow$

$\|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2 \cdot \|x_2 - x_1\|_2 \leq L \|x_2 - x_1\|_2^2$

(Applying Cauchy-Schwartz)

$\Rightarrow$

$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \leq L \|x_2 - x_1\|_2^2$

$\Rightarrow$

$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x_2 - x_1) \leq L \|x_2 - x_1\|_2^2 = L(x_2 - x_1)^T (x_2 - x_1)$

$\Rightarrow$

$(Lx_2 - \bigtriangledown f(x_2) - Lx_1 + \bigtriangledown f(x_1)))^T (x_2 - x_1) \geq 0$

$\Rightarrow$

$(\bigtriangledown g(x_2) - \bigtriangledown g(x_1)))^T (x_2 - x_1) \geq 0$ (define $g(x) = \frac{L}{2} x^T x - f(x)$)

$\Rightarrow$

$g(x)$ is convex (the monotone gradient condition).

$\Rightarrow$

$g(x_2) \geq g(x_1) + \bigtriangledown g(x_1)^T (x_2 - x_1)$

$\Rightarrow$

$\frac{L}{2} x_2^T x_2 - f(x_2) \geq \frac{L}{2} x_1^T x_1 - f(x_1) + (Lx_1 - \bigtriangledown f(x_1))^T (x_2 - x_1)$

$\Rightarrow$

$f(x_2) \leq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{L}{2} x_2^T x_2 + \frac{L}{2} x_1^T x_1 - Lx_1^T x_2$

$\Rightarrow$

$f(x_2) \leq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{L}{2} \|x_2 - x_1\|_2^2$

2.2 b

Implies

$$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T(x_2 - x_1) + \frac{1}{2L} \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2 \qquad (11)$$

f is convex and L-smooth.

$f(x_2) - f(x_1) = f(x_2) - f(x_3) + f(x_3) - f(x)$

(define $x_3 = x_2 - \frac{1}{L}(\bigtriangledown f(x_1) - \bigtriangledown f(x_2))$)

$\Rightarrow$

$f(x_2) - f(x_3) \geq - \bigtriangledown f(x_2)^T(z - x_2) - \frac{L}{2} \|x_2 - x_3\|_2^2$ (quadratic upper bound)

$f(x_3) - f(x_1) \geq \bigtriangledown f(x_1)^T(x_3 - x_1)$ (convex)

$\Rightarrow$

$f(x_2) - f(x_1) \geq - \bigtriangledown f(x_2)^T(x_3 - x_2) - \frac{L}{2} \|x_2 - x_3\|_2^2 + \bigtriangledown f(x_1)^T(x_3 - x_1)$

$\Rightarrow$

$f(x_2) - f(x_1) \geq -\frac{1}{L} \bigtriangledown f(x_2)^T(\bigtriangledown f(x_1) - \bigtriangledown f(x_2)) - \frac{1}{2L} \|\bigtriangledown f(x_1) - \bigtriangledown f(x_2)\|_2^2$

$+ \bigtriangledown f(x_1)^T(x_2 - x_1) + \frac{1}{L} \bigtriangledown f(x_1)^T(\bigtriangledown f(x_1) - \bigtriangledown f(x_2))$ (substitute z)

$\Rightarrow$

$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T(x_2 - x_1) + \frac{1}{2L} \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2$

2.3 c

Implies

$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T(x_2 - x_1) \geq \frac{1}{L} \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2 \qquad (12)$$

$$f(x_2) \geq f(x_1) + \bigtriangledown f(x_1)^T(x_2 - x_1) + \frac{1}{2L} \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2$$

$$f(x_1) \geq f(x_2) + \bigtriangledown f(x_2)^T(x_1 - x_2) + \frac{1}{2L} \|\bigtriangledown f(x_1) - \bigtriangledown f(x_2)\|_2^2$$

$\Rightarrow$

$$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T(x_2 - x_1) \geq \frac{1}{L} \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2$$

3 Problem 3

Define, discuss the benefits, and give examples for the different convergence rates of a sequence of updates xk: (a) Sublinear, (b) Linear, (c) SuperLinear, (d) Quadratic.

3.1 Sublinear

The sequence of updates x_k sublinearly converges to L if satisfies the following equation:

$$\lim_{k\to\infty} \frac{|x_{k+1} - L|}{|x_k - L|} = 1$$

In this case, x_k converges slowly. For example, the x_k below converges to 0 sublinearly:

$$x_k = \frac{1}{k+1}$$

Because we have:

$$\lim_{k\to\infty} \frac{|x_{k+1} - 0|}{|x_k - 0|} = \lim_{k\to\infty} \frac{\left|\frac{1}{k+2}\right|}{\left|\frac{1}{k+1}\right|} = \lim_{k\to\infty} \frac{\left|\frac{\partial}{\partial k}(\frac{1}{k+2})\right|}{\left|\frac{\partial}{\partial k}(\frac{1}{k+1})\right|} = \lim_{k\to\infty} \frac{\left|(k+2)^2\right|}{|(k+1)^2|} = 1$$

3.2 Linear

The sequence of updates x_k linearly converges to L if satisfies the following equation:

$$\lim_{k\to\infty} \frac{|x_{k+1} - L|}{|x_k - L|} = \mu, \mu \in (0, 1)$$

In this case, x_k converges faster than the case with sublinear convergence. For example, the x_k below converges to 0 linearly:

$$x_k = (\frac{1}{2})^k$$

Because we have:

$$\lim_{k\to\infty} \frac{|x_{k+1} - 0|}{|x_k - 0|} = \lim_{k\to\infty} \frac{\left|(\frac{1}{2})^{k+1}\right|}{\left|(\frac{1}{2})^k\right|} = \lim_{k\to\infty} \frac{1}{2} = \frac{1}{2}$$

3.3 Superlinear

The sequence of updates x_k sublinearly converges to L if satisfies the following equation:

$$\lim_{k\to\infty} \frac{|x_{k+1} - L|}{|x_k - L|} = 0$$

In this case, x_k converges faster than the case with linear convergence. For example, the x_k below converges to 0 superlinearly:

$$x_k = (\frac{1}{2})^{2^k}$$

Because we have:

$$\lim_{k\to\infty} \frac{|x_{k+1} - 0|}{|x_k - 0|} = \lim_{k\to\infty} \frac{\left|(\frac{1}{2})^{2^{k+1}}\right|}{\left|(\frac{1}{2})^{2^k}\right|} = \lim_{k\to\infty} \left|(\frac{1}{2})^{2k}\right| = 0$$

3.4 Quadratic

Quadratic convergence is one type of superlinear convergence. The sequence of updates x_k converges to L with the order 2. Specifically, it should satisfies the following equation where M is a positive constant value.

$$\lim_{k\to\infty} \frac{|x_{k+1} - L|}{|x_k - L|^2} < M$$

The example we show in the superlinear above also converge to 0 quadratically. Because we have:

$$\lim_{k\to\infty} \frac{|x_{k+1} - 0|}{|x_k - 0|^2} = \lim_{k\to\infty} \frac{\left|(\frac{1}{2})^{2^{k+1}}\right|}{\left|(\frac{1}{2})^{2^k+2^k}\right|} = \lim_{k\to\infty} \frac{\left|(\frac{1}{2})^{2^{k+1}}\right|}{\left|(\frac{1}{2})^{2^{k+1}}\right|} = 1$$

So we can choose any constant value M larger than 1 to satisfy the definition.

3.5 Discussion

In machine learning, in order to find the optimal solution, we are happy when the convergence rate is faster. In other words, we prefer linear or superlinear convergence rate rather than sublinear convergence rate. From the machine learning aspect, we want to minimize the loss function $f(\omega)$ according to the specific problem where ω is the model parameter. We denote the optimal solution as ω^*. A variation of definition of convergence rate is:

$$\lim_{k\to\infty} \frac{f(\omega^{k+1}) - f(\omega^*)}{f(\omega^k) - f(\omega^*)} = \rho$$

$$\rho = 1 \rightarrow sublinear$$
$$\rho \in (0, 1) \rightarrow linear$$
$$\rho = 0 \rightarrow superlinear$$

For some nice designed loss functions f, gradient descent algorithm can have a linear convergence rate. A good example is least square problem. When we have a convex function f, we could introduce an L2 regularization term to make f to be strongly-convex which has Polyak-Lojasiewicz (PL) inequality property. It implies that we could improve gradient methods from sublinear rate to a linear rate. At the same time, when f is fixed, we may try other optimization algorithms like Newton's method which achieves superlinear convergence rate (however requires expensive computation when the dimension is large).

4 Problem 4

Consider

$$\begin{aligned} &\text{minimize } \frac{1}{N}\sum_{i\in[N]} f_i(x_i) \\ &s.t.\ \mathbf{Ax} = \mathbf{b} \end{aligned} \tag{13}$$

for $\mathbf{A} \in \mathbb{R}^{p\times N}$ and $\mathbf{x} = [x_1, \ldots, x_N]^T$.

4.1 a,b,c

We have strong-convexity and smoothness on f, assuming f is twice differentiable then we could calculate the Hessian matrix of $f(\mathbf{x}) = \frac{1}{N}\sum_{i\in[N]} f_i(x_i)$:

$$\bigtriangledown^2 f(\mathbf{x}) = \begin{bmatrix} \bigtriangledown^2 f_1(x_1) & 0 & \cdots & 0 \\ 0 & \bigtriangledown^2 f_2(x_2) & & \\ \vdots & & \ddots & \\ 0 & & & \bigtriangledown^2 f_N(x_N) \end{bmatrix} \tag{14}$$

since it is diagonal matrix, the inverse of Hessian matrix is fast no matter when $N = 1000$ or $N = 10^9$. We could use Newton's method with equality constraints to solve it:
1. select a starting point $\mathbf{x}$ that satisfies $\mathbf{Ax} = \mathbf{b}$
2. Compute the Newton step $\Delta\mathbf{x}_{nt}$ by solving:

$$\begin{bmatrix} \bigtriangledown^2 f(\mathbf{x}) & \mathbf{A}^T \\ \mathbf{A} & 0 \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_{nt} \\ w \end{bmatrix} = \begin{bmatrix} -\bigtriangledown f(\mathbf{x}) \\ 0 \end{bmatrix} \tag{15}$$

where w is the associated optimal dual variable for the quadratic problem. And compute $\lambda^2(\mathbf{x})$ as:

$$\lambda^2(\mathbf{x}) = \Delta\mathbf{x}_{nt}^T \bigtriangledown^2 f(\mathbf{x})\Delta\mathbf{x}_{nt} \tag{16}$$

3. Quit if $\lambda^2(\mathbf{x})/2 \leq \epsilon$
4. Update $\mathbf{x} = \mathbf{x} + \Delta\mathbf{x}_{nt}$

The computing Hessian matrix is efficient since it is diagonal.
If f is not twice differentiable, we could only calculate the gradient of f, thus we could use projected gradient descent to solve it.

4.2 d

At this point, the objective function $f(\mathbf{x}) = \frac{1}{N}\sum_{i\in[N]} f_i(x_i) + r(\mathbf{x})$. Since we do not know if $\frac{\partial^2 r(\mathbf{x})}{\partial x_i \partial x_j}$ is zero, then the computing of Hessian matrix and its inverse for each iteration would become too expensive. So we could use quasi-Newton's method to update the Hessian by analyzing successive gradients.

5 Problem 5

Problem
In the convergence proof of GD with constant step size and strongly convex objective function (see slides), prove the coercivity of the gradient:

$$\langle \bigtriangledown f(x) - \bigtriangledown f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\bigtriangledown f(x_2) - \bigtriangledown f(x_1)\|_2^2 \quad (17)$$

Proof
Extend function f to:

$$g(x) = f(x) + \frac{\mu}{2} \|x\|^2$$

According to μ-convexity of f, g is convex which is proven in Problem 1. Now the L-smoothness of f can be extended to $L - \mu$ smoothness of g according to the definition of Lipschitz-smoothness. Then:

$$\begin{aligned}
\langle \bigtriangledown g(x) - \bigtriangledown g(y), x - y \rangle &\geq \frac{1}{(L - \mu)} \|\bigtriangledown g(x) - \bigtriangledown g(y)\|_2^2 \\
&\Rightarrow \\
\langle \bigtriangledown f(x) - \bigtriangledown f(y) - \mu(x - y), x - y \rangle &\geq \frac{1}{(L - \mu)} \|\bigtriangledown f(x) - \bigtriangledown f(y) - \mu(x - y)\|_2^2 \\
&\Rightarrow (Denote(*) = \langle \bigtriangledown f(x) - \bigtriangledown f(y), x - y \rangle) \\
(*) - \mu \|x - y\|_2^2 &\geq \frac{1}{(L - \mu)} \|\bigtriangledown f(x) - \bigtriangledown f(y)\|_2^2 + \frac{\mu^2}{(L - \mu)} \|x - y\|_2^2 - 2\frac{\mu}{L - \mu}(*) \\
&\Rightarrow \\
\frac{L + \mu}{L - \mu}(*) &\geq \frac{1}{(L - \mu)} \|\bigtriangledown f(x) - \bigtriangledown f(y)\|_2^2 + \frac{\mu L}{(L - \mu)} \|x - y\|_2^2 \\
&\Rightarrow Strong\ coercivity
\end{aligned}$$