IEOR 262B LECTURE NOTE

SPRING 2024

Hyunin Lee
Ph.D. student
UC Berkeley
hyunin@berkeley.edu

Contents

1 Lecture 1

2 Lecture 2

2.1 Properties of local minimum x_*

2.1.1 First-order necessary condition

We use the abbreviation FOC(necessary) which denotes first-order necessary condition

Theorem 2.1 (FOC(necessary))**.** If x_* is a local min, then $\nabla f(x_*) = 0$.

??

Proof. Let $g(t) = f(x_* + t\Delta x)$. Then by the definition of derivation, we have

$$g'(0) = \lim_{t \to 0} \frac{f(x_* + t\Delta x) - f(x_*)}{t}.$$

Note that $f(x_* + t\Delta x) - f(x_*) > 0$ holds by the definition of optimal point x_*. So we conclude that $g'(0) \geq 0$. Then we have

$$0 \leq g'(0) = \nabla f(x_*)^\top \Delta x$$

for all Δx. Now plug in $\Delta x = -\Delta f(x_*)$, then we have $0 \leq -||\nabla f(x_*)||^2$. Then this concludes that we have $\nabla f(x_*) = 0$. □

2.1.2 Second-order necessary condition

Theorem 2.2 (SOC (necessary))**.** If x_* is a local min, then $\nabla^2 f(x_*) \succcurlyeq 0$.

Proof. Proof by a contradiction. If $\nabla^2 f(x_*) \not\succcurlyeq 0$ holds, then there exist Δx that satisfies $\Delta x^\top \nabla^2 f(x_*)\Delta x < 0$. Then, by Taylor expansion, we have

$$f(\underbrace{x_*}_{\text{nominal}} + \underbrace{t\Delta x}_{\text{perturbation}}) = f(x_*) + t\Delta f(x_*)^\top \Delta x + \frac{1}{2}\Delta x^\top \nabla^2 f(\underbrace{y(t)}_{\text{new point}})\Delta x$$

Then as $t \to 0$, then $y(t) \to x_*$, so we have

1. $\nabla^2 f(y(t)) \to \nabla^2 f(x_*)$.

2. $\Delta x^\top \nabla^2 f(y(t))\Delta x \to \Delta x^\top \nabla^2 f(x_*)\Delta x$.

Therefore, the above two bullets lead us to conclude that $f(x_* + t\Delta x) < f(x_*)$ holds which contradicts the definition of local min. □

2.1.3 Second-order sufficient condition

Theorem 2.3 (SOC (sufficient))**.** If x_* satisfies $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*) > 0$, then the following holds.

1. x_* is a strict local min.
2. $\exists \epsilon > 0, \mu > 0$ such that $f(x) \geq f(x_*) + \mu ||x - x_*||^2$.

Proof. If $\nabla f(x_*) > 0$ holds then there exist $\epsilon, \lambda > 0$ such that $\text{mineig}(\nabla^2 f(x)) \geq \lambda$ for $\forall x \ : \ ||x - x_*|| \leq \epsilon$. Now, let the point y as an intermediate point between x_* and $x_* + \Delta x$. The corresponding Taylor expansion would be

$$f(x_* + \Delta x) = f(x_*) + \nabla f(x_*) \Delta x + \frac{1}{2} \Delta x^\top \nabla^2 f(y) \Delta x.$$

Let $x = x_* + \Delta x$ and note that $\nabla f(x_*) = 0$ holds and $\nabla^2 f(y) \succcurlyeq \lambda \mathbb{I}$. Then we have

$$f(x) \geq f(x_*) + \frac{\lambda}{2} ||\Delta x||^2$$

where setting $\mu = \lambda/2$ finishes the proof. □

2.2 Update rule

So far, we have investigated some properties of local minimum. FOC, SOC. However, the main question that we are interested in is more practical: **how to find local minimum?**. Within this sense, let's come up with **decent method**. We cannot find x_* in one shot. We have to find it gradually. If we generate the point sequence as $\{x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots \to x^{(k)} \to x^{(k+1)} \ \to \cdots\}$ and what we want is the improvement at each iteration, i.e. $f(x^{(0)}) > f(x^{(1)}) > f(x^{(2)}) > \cdots > f(x^{(k)}) > f(x^{(k+1)}) > \cdots$. Within this sense, we introduce the following update rule

Definition 2.4 (Update rule)**.** $\underbrace{x^{(k+1)}}_{\text{new point}} = \underbrace{x^{(k)}}_{\text{old point}} + \underbrace{\alpha^{(k)}}_{\text{stepsize}} \underbrace{\Delta x^{(k)}}_{\text{direction}}$

Theorem 2.5. Three followings are held.

1. If $\nabla f(x^{(k)})^\top \Delta x^{(k)} > 0$, then $\exists \tau > 0$ such that $f(x^{(k+1)}) > f(x^{(k)})$ for $\forall \alpha^{(k)} \in [0, \tau]$.
2. If $\nabla f(x^{(k)})^\top \Delta x^{(k)} < 0$, then $\exists \tau > 0$ such that $f(x^{(k+1)}) < f(x^{(k)})$ for $\forall \alpha^{(k)} \in [0, \tau]$.
3. If $\nabla f(x^{(k)})^\top \Delta x^{(k)} = 0$, then look at $\text{sign}((\Delta x^{(k)})^\top \nabla^2 f(x^{(k)}) \nabla x^{(k)})$.

2.2.1 Descent direction

Definition 2.6 (Descent direction)**.** In the update rule, if $\nabla f(x^{(k)})^\top \Delta x < 0$ is satisfied, then Δx is called descent.

2.2.2 Steepest descent

Note that there are *infinitely many* descent directions. **Then what is the best direction?**. Let's take a look at the following equation to come up with what is the best direction intuitively.

$$\underbrace{f(x^{(k+1)})}_{\text{(I)}} = \underbrace{f(x^{(k)})}_{\text{fixed}} + \underbrace{\alpha^{(k)}}_{\text{fixed}} \underbrace{\nabla f(x^{(k)})^\top \nabla x^{(k)}}_{\text{(II)}} + \cdots$$

One intuition that we could get is reducing term (II) also leads to reducing term (I). Then we can determine the best direction as follows.

$$\begin{aligned} \min_{\Delta x} \quad & \nabla f(x^{(k)})^\top \Delta x \\ \text{s.t.} \quad & ||\Delta x|| \leq 1 \end{aligned}$$

Note that regarding with norm $||\Delta x||$, it is open to either choose $1-, 2-\infty-$ norm, etc... In most case, we use standard norm of $||\Delta x||_2$ where

$$\Delta x = -\frac{\nabla f(x^{(k)})}{||\nabla f(x^{(k)})||}.$$

Since we have a step size, let's make it proportional then we can rewrite it as $\Delta x = -\nabla f(x^{(k)})$.

Definition 2.7 (Steepest descent)**.** If we choose $\Delta x^{(k)} = -\nabla f(x^{(k)})$, then we call this method as steepest descent method (w.r.t $|| \cdot ||_2$). We can generalize this to $\Delta x^{(k)} = -D^{(k)} \nabla f(x^{(k)})$ where $D^{(k)} \succ 0$.

Theorem 2.8. If $D^{(k)} \succ 0$, then the steepest descent direction is a descent direction. Namely for $\Delta x^{(k)} = -D^{(k)} \nabla f(x^{(k)})$ where $D^{(k)} \succ 0$, $\nabla f(x^{(k)})^\top \Delta x^{(k)} \leq 0$ holds.

2.3 Stepsize

How to find a stepsize $\alpha^{(k)}$?

1. Exact line search
2. Limited line search
3. Backtracking

2.3.1 Exact line search

3 Lecture 3

We are interested in solving the problem $\min f(x)$ and we have shown that optimal point x_* satisfies $\nabla f(x_*) = 0$ (FOC). We are interested in computing $\{x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots\}$ by a algorithm $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$. We determine $\alpha^{(k)}$ by 1) exact line search 2) limited line search 3) backtracking and we call $\Delta x^{(k)}$ a descent direction if $\nabla f(x^{(k)})^\top \Delta x^{(k)} < 0$ is satisfied. The important part of previous methods are whether $f(x^{(k+1)}) < f(x^{(k)})$ holds. Now, let's focus on how much it could improve, namely $f(x^{(k+1)}) - f(x^{(k)})$.

3.1 Armijo rule

Definition 3.1 (Armijo rule)**.** We have a pre-set parameter $0 < \sigma < 1$. Pick the smallest t such that $\alpha^{(k)} = \alpha\beta^t$ satisfies the following inequality

$$f(x^{(k+1)}) - f(x^{(k)}) < \sigma \nabla f(x^{(k)})^\top \Delta x^{(k)} \alpha^{(k)} \tag{1}$$

Theorem 3.2. If t is bigger than the threshold, then Inequality (1) is always satisfied.

3.2 Constant stepsize

So far, we have dealt with how to determine step size $\alpha^{(k)}$ that varies as k goes by. Then what if we fix the $\alpha^{(k)} = \alpha$ as a constant? we call this gradient algorithm

$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)}) \tag{2}$$

The first question that naturally we can come up with is whether the gradient algorithm guarantees convergence. Usually, it does not converge.

3.3 Diminishing stepsize

If $\alpha^{(k)} \to 0$ but $\sum_{k=1}^{\infty} \alpha^{(k)} = +\infty$ then it's going to work. For example if $\alpha^{(k)} = \frac{1}{k}$, it does not work but if $\alpha^{(k)} = \frac{1}{k^2}$, then it works. However, we have a convergence issue. Let's think about the algorithm that satisfies a descent direction, i.e. $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$ where $\nabla f(x^{(k)})^\top \Delta x^{(k)} < 0$. The problem happens if the direction is becoming nearly orthogonal to the gradient. So we need the following assumption

$$\lim_{k \to \infty} \frac{\nabla f(x^{(k)})^\top \Delta x^{(k)}}{||\nabla f(x^{(k)})|| \, ||\Delta x^{(k)}||} < 0 \tag{3}$$

Note that (3) tells us that the angle between direction $(\Delta^{(k)})$ and the gradient $(\nabla f(x^{(k)}))$ should not be 90 degree.

3.4 Gradient related

Definition 3.3 (Gradient related)**.** $\{\Delta x^{(k)}\}_{k=1}^{\infty}$ is gradient related if for any subsequence $\{x^{(k)}\}_{k\in K}$ such that $\{x^{(k)}\}_{k\in K}$ converges to a non-stationary point, then

- $\{\Delta x^{(k)}\}_{k=1}^{\infty}$ is bounded
- $\lim_{k\to\infty}\sup_{k\in K}\Delta f(x^{(k)})^{\top}\Delta x^{(k)} < 0$.

Based on the Defintion 3.3, we have the following lemma

Lemma 3.1. Let gradient direction $\Delta x^{(k)} = -D^{(k)}\nabla f(x^{(k)})$ where the matrix $D^{(k)} \succ 0$. The matrix $D^{(k)}$ can be regarded as a scaling factor related to the gradient. If there exists a constant $c_1 > 0, p_1 \geq 0, c_2 > 0, p_2 \geq 0$ that satisfies then this is gradient gradient-related algorithm.

For example, $\Delta x^{(k)} = -\nabla f(x^{(k)})$ is gradient related algorithm (also, later on, we define this as gradient algorithm) since $D^{(k)} = \mathbb{I}$

3.5 When to stop?

We have some candidates to determine when to stop

1. Stop if $||\nabla f(x^{(k)})|| \leq \epsilon$
2. Stop if $\frac{||\nabla f(x^{(k)})||}{||\nabla f(x^{(0)})||} \leq \epsilon$
3. Stop if $||x^{(k+1} - x^{(k)}|| \leq \epsilon$

Theorem 3.4. Suppose we have an optimal point x_* that satisfies SOC sufficient, i.e. $\nabla^2 f(x_*) \succ 0$. Think about a ball B around the x_* that satisfies $\nabla^2 f(x) \succ 0$. Then if $||\nabla f(x|| \leq \epsilon$ for any $x \in B$, then the following holds,

1. $||x - x_*|| \leq \epsilon/m$
2. $f(x) - f(x_*) \leq \epsilon^2/2m$

it also means that **if the gradient $\nabla^2 f(x)$ is small, then it is close enough to solution either x_* or $f(x_*)$.**

Proof. Assume that intermediate point y between x_* and x that is also stated in a ball B. i.e. $x, x_*, y \in B$. The Taylor expansion where x_* is a new point and x is a nominal point is given as follows.

$$\begin{aligned} f(x_*) &= f(x) + \Delta f(x)^{\top}\Delta x + \frac{1}{2}\Delta x^{\top}\nabla^2 f(y)\Delta x \\ &\geq f(x) + \nabla f(x)^{\top}\Delta x + \frac{1}{2}m||\Delta x||^2 \\ &\geq \min_{z\in\mathbb{R}^n}\left(f(x) + \nabla f(x)^{\top}z + \frac{1}{2}m||z||^2\right) \end{aligned}$$

The second inequality comes from the fact that $y \in B$ so $\nabla^2 f(y) \succcurlyeq m\mathbb{I}$ holds. We let m be the minimum eigenvalue of $\nabla^2 f(y)$ for any $y \in B$. Then, the last inequality is a quadratic convex function since $\nabla f(x) \succ 0$. Then it attains its minimum at point z_* that satisfies FOC. Therefore $z_* = -\nabla f(x)/m$ is plugged into the quadratic term and we get the following.

$$\begin{aligned} f(x_*) &\geq f(x) - \frac{||\nabla f(x)||^2}{2m} \\ &\geq f(x) - \frac{\epsilon^2}{2m} \end{aligned}$$

this completes the proof of the second bullet point. Now, let's complete the proof of the first bullet point. Again, we have that Taylor expansion as

$$\begin{aligned} f(x) &= f(x_*) + \nabla f(x_*)^\top \Delta x + \frac{1}{2}\Delta x^\top \nabla^2 f(\tilde{y})\Delta x \\ &= f(x_*) + \frac{1}{2}\Delta x^\top \nabla^2 f(\tilde{y})\Delta x \end{aligned} \tag{4}$$

and by the second bullet point, we have

$$f(x_*) \geq f(x) - \frac{\epsilon^2}{2m} \tag{5}$$

Add above Equation (4) and Inequality (5), then we have

$$\frac{\epsilon^2}{2m} \geq \frac{1}{2}m||\Delta x||^2$$

which we finally have

$$||\Delta x|| \leq \frac{\epsilon}{m}$$

where $\Delta x = x - x_*$. □

4 Lecture 4

So far, we are interested in finding that $\min f(x)$. We find its minimum by an iterative method,i.e. $\{x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots\}$ where $x^{(k+1)} = x^{(k)} + \alpha^{(k)}\Delta x^{(k)}$ holds. If we use the gradient method, i.e. $\Delta x^{(k)} = -\nabla f(x^{(k)})$, then our gradient descent algorithm becomes $x^{(k+1)} = x^{(k)} - \alpha^{(k)}\nabla f(x^{(k)})$. However, in most cases, we don't know the exact value of $\nabla f(x^{(k)})$. Let $g^{(k)}$ to be approximate gradient of $\nabla f(x^{(k)})$. Namely, we let the following expression,

$$g^{(k)} = \nabla f(x^{(k)}) + e^{(k)}$$

where $e^{(k)}$ is an approximation error term.

Relatively small error

Assume that $e^{(k)}$ is small relative to the gradient, namely for all k, $||e^{(k)}|| \leq ||\nabla f(x^{(k)}||$ holds. one downside of this algorithm is that as we proceed gradient gets smaller, and then the error must be small simultaneously.

Lemma 4.1. If $||e^{(k)}|| \le ||\nabla f(x^{(k)}||$ holds for $\forall k$, then $g^{(k)}$ is a decent direction, i.e. $\nabla f(x^{(k)})^\top(-g^{(k)}) < 0$.

Proof.

$$\begin{aligned}
\nabla f(x^{(k)})^\top(-g^{(k)}) &= -||\nabla f(x^{(k)})||^2 - \nabla f(x^{(k)})^\top e^{(k)} \\
&\le -||\nabla f(x^{(k)})||^2 + ||f(x^{(k)})|| \cdot ||e^{(k)}|| \\
&= -||\nabla f(x^{(k)})||(\underbrace{||f(x^{(k)})|| - ||e^{(k)}||}_{>0}) \\
&\le 0
\end{aligned}$$

Note that first equality holds by definition of $g^{(k)}$, and second inequality holds by Cauchy inequality. □

Bounded error

Now assume that $||e^{(k)}|| \le \delta$ holds for $\forall k$. One of our guesses on how the $x^{(k)}$s behave is that points might show erratic behavior in the neighborhood of x_*, an optimal point. However, if $x^{(k)}$ states outside of the neighborhood, it might show convergence behavior following a decent algorithm.

Lemma 4.2. Define neighborhood $\mathcal{D} = \{x \mid ||\nabla f(x)|| \le \delta\}$. Then the point $x^{(k)}$ outside of $\mathcal{D}$ converges as a decent direction. Namely, $-\nabla f(x^{(k)})^\top g^{(k)} \le 0$ holds for $x^{(k)} \notin \mathcal{D}$.

Proof.

$$\begin{aligned}
\nabla f(x^{(k)})^\top(-g^{(k)}) &= -||\nabla f(x^{(k)})||^2 - \nabla f(x^{(k)})^\top e^{(k)} \\
&\le -||\nabla f(x^{(k)})||^2 + ||f(x^{(k)})|| \cdot ||e^{(k)}|| \\
&\le -||\nabla f(x^{(k)})||^2 + ||f(x^{(k)})|| \cdot \delta \\
&= -||\nabla f(x^{(k)})||(\underbrace{||f(x^{(k)})|| - \delta}_{>0}) \\
&\le 0
\end{aligned}$$

□

Algorithm	gradient algorithm	perturbed gradient algorithm
Equation	$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$	$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \big(\nabla f(x^{(k)}) + \delta\big)$
Objective	$\min f(x)$	$\min f(x) + \delta^\top x$

SGD: Stochastic Gradient descent

Now, instead of saying error $e^{(k)}$ is relatively smaller than the gradient or bounded by a constant, let's regard it as a random variable. Let us assume $g^{(k)} = \nabla f(x^{(k)}) + e^{(k)}$ where $e^{(k)}$ is random. Then the objective function that we wanted to minimize was $\min f(x)$. However, since we have randomness in a function itself, the objective function should be $\mathbb{E}_w\left[F(x, w)\right]$ where w is an uncertainty. Therefore our gradient looks like $\nabla f(x) = \mathbb{E}_w\left[\nabla_x F(x, w)\right]$.

One problem for computing gradient is that it is hard to obtain in the real world since the expectation is over the infinite w space. One basic approach is approximating $\nabla f(x)$ with just one uncertainty sample. Namely, we have defined $\nabla f(x^{(k)})$ as $\mathbb{E}_w\left[\nabla_x F(x^{(k)}, w)\right]$ but we will approximate this with just using one sample $w^{(k)}$,i.e $\nabla f(x^{(k)}) = \mathbb{E}_w\left[\nabla_x F(x^{(k)}, w)\right] \approx \nabla_x F(x^{(k)}, w^{(k)})$. Now, let's define the error $e^{(k)}$ as follows.

$$e^{(k)} = \nabla_x F(x^{(k)}, w^{(k)}) - \mathbb{E}_w\left[\nabla_x F(x^{(k)}, w)\right]$$

Lemma 4.3. Since $\mathbb{E}_w(e^{(k)}) = 0$, SGD is descent algorithm on average.

Theorem 4.1 (SGD convergence)**.** Let us assume that $e^{(k)}$ is $i.i.d$ and zero mean and bounded. Assume that stepsize satisfy $\lim_{k\to\infty} \alpha^{(k)} \to 0$, $\sum_{k=0}^{\infty} \alpha^{(k)} = \infty$, and $\sum_{k=0}^{\infty} (\alpha^{(k)})^2 = \infty$. Then SGD converges to a stationary point.

Before moving to the next theorem, what is the limit point?

Example 4.2. If a sequence is $\{+1, -1, +1, -1, ...,\}$, then there is no convergence, and the sequence has two points: $+1, -1$.

Theorem 4.3. For given algorithm $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$. Assume that $\{\Delta x^{(k)}\}_{k=1}^{\infty}$ is gradient related. Let's say we compute stepsize $\alpha^{(k)}$ by among 1) exact line search, 2) limited line search, or 3) Armijo rule. Then every limit point of $\{x^{(k)}\}_{k=1}^{\infty}$ is a stationary point.

Proof. To be continued. □

4.1 Lipschitz continuity of gradient

Definition 4.4 (Lipschitz continuity of gradient)**.** Function f is Lipschitz continuity of gradient if there exist $L > 0$ that satisfy $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for $\forall x, y \in \mathbb{R}^n$.

Theorem 4.5 (Bounded by quadratic function)**.** Lipschitz continuity of gradient implies that $f(x+y) \leq f(x) + y^\top \nabla f(x) + \frac{L}{2}||y||^2$ for $\forall x, y$. This also brings up that when $x = 0$, $f(y) \leq f(0) + y^\top \nabla f(0) + \frac{L}{2}||y||^2$ holds which means function $f(x)$ is bounded by quadratic term $\frac{L}{2}||y||^2$.

Proof. Let $g(t) = f(x + ty)$ for all $t \in \mathbb{R}$. Then, we have the following expression.

$$
\begin{aligned}
f(x+y) - f(x) = g(1) - g(0) &= \int_0^1 \frac{dg(t)}{dt} dt \\
&= \int_0^1 y^\top \nabla f(x+ty) dt \\
&= \int_0^1 y^\top f(x) dt + \int_0^1 y^\top (\nabla f(x+ty) - \nabla f(x)) dt \\
&\leq \int_0^1 y^\top f(x) dt + \int_0^1 ||y|| \cdot ||\nabla f(x+ty) - \nabla f(x)|| dt \\
&\leq \int_0^1 y^\top f(x) dt + \int_0^1 ||y|| \cdot L||(x+ty) - x|| dt \\
&= y^\top f(x) + L||y||^2 \int_0^1 t dt \\
&= y^\top f(x) + \frac{L}{2}||y||^2
\end{aligned}
$$

□

5 Lecture 5

Theorem 5.1 (Capture theorem)**.** If x^0 is close enough to isolated local min x_*, then $\{x^{(k)}\} \to x_*$.

5.1 Convergence rate

Let's say you have a algorithm that $x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots \to x_*$ where x_* are bunch of points in $\mathbb{R}^n$. We assume x_* is a unique limit point. We want to measure its speed of convergence. We have some candidates to express how errors become smaller. $e(x) = ||x - x_*||$ or $e(x) = ||f(x) - f(x_*)||$. In most cases, we are interested in *asympototic convergence rate* behavior so the tail is important. However, just an asymptotic convergence rate assumption is not enough. Thinks about two algorithms 1 and 2 where algorithm 1 yields $\{e^{(k)}\} = \{1, 0.9, 0.8, 0.2, 10^{-2}\}$ and algorithm 2 yields $\{e^{(k)}\} = \{1, 0.7, 0.8, 0.6, 0.3, 10^{-3}\}$. Which algorithm is better is not an obvious question since the comparison of $e^{(2)}$ and $e^{(4)}$ is different.

This reminds us to set a baseline for convergence rate comparison among algorithms.

Definition 5.2 (Linear convergence)**.** Think about a sequence $\{1, \beta, \beta^2, \beta^2, \cdots\}$. We say $\{e^{(k)}\}$ converges linearly or geometrically with factor β is there exist $q > 0$ that for all $k \in \mathbb{N}$ $e(x^{(k)}), \leq q\beta^k$ satisfied. Note that $e(x^{(k)}) = e^{(k)}$.

Theorem 5.3. The error sequence $\{e^{(k)}\}$ converges linearly if $\lim_{k\to\infty} e^{(k+1)}/e^{(k)} < 1$ holds. In this case, a factor $\beta = \lim_{k\to\infty} e^{(k+1)}/e^{(k)}$.

The previous definition and theorem let us conclude that if β decreases, then the algorithm gets faster. Now, what if $\beta = 0$? This means that the sequence $\{e^{(k)}\}$ is faster than $\{\beta^k\}$ for every $0 < \beta < 1$. We say this as converging superlinearly.

Definition 5.4 (Superlinear convergence)**.** The sequence $\{e^{(k)}\}$ converges superlinearly with order p if $e(x^{(k)}) \leq q\beta^{p^k}$ is satisfied (for all k or for large enough k) for some $q > 0$ and $0 < \beta < 1$.

Theorem 5.5. The error sequence $\{e^{(k)}\}$ converges superlinearly with order p if $\lim_{k\to\infty} e^{(k+1)}/(e^{(k)})^p \leq \infty$ holds.

In the case of $p = 2$, we call this quadratic convergence. Note that the gradient algorithm yields an error sequence to satisfy linear convergence and Newton's algorithm makes it to satisfy quadratic convergence.

5.2 Condition number

Definition 5.6. For a positive definite matrix Q, we define its condition number as

$$\text{c.d.} = \frac{\max \operatorname{eig}(Q)}{\min \operatorname{eig}(Q)} = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \geq 1.$$

Now let's think about a function $f : \mathbb{R}^n \to \mathbb{R}$ and let x_* be strict (isolated) local minima where it satisfies SOC (sufficient). By a taylor expansion, we can rewrite $f(x)$ as $f(x) = f(x_*) + \nabla f(x_*)^\top (x - x_*) + 1/2(x - x_*)^\top \nabla^2 f(x_*) + \mathcal{O}(||x - x_*||^2)$. By SOC (sufficient) condition. We have $f(x_*) = 0, \nabla f(x_*)^\top = 0$. Now, if x is close enough to x_*, by capture theorem, $x - x_* \to x$, so the RHS of the taylor expansion get closes to $\frac{1}{2}x^\top Qx$ where $Q := \nabla^2 f(x_*)$. This observation provides us an insight that $\min f(x)$ is close to solving where $\min \frac{1}{2}x^\top Qx$, namely thinking $f(x) = \frac{1}{2}x^\top Qx$ within a bound that satisfies a capture theorem.

Within this sense, let's remind the gradient algorithm: $x^{(k+1)} = x^{(k)} - \alpha^{(k)}\nabla f(x^{(k)})$. Applying the above observation, we have a modified gradient algorithm as $x^{(k+1)} = x^{(k)} - \alpha^{(k)}Qx^{(k)} = (\mathbb{I} - \alpha^{(k)}Q)x^{(k)}$. In this scenarios, the error term $e(x^{(k)}) = ||x^{(k)} - x_*|| = ||x^{(k)}||$ since $x_* = 0$ in a quadratic function $f(x)$. Then we have $||x^{(k+1)}||^2 = (x^{(k)})^\top(\mathbb{I} -$

$\alpha^{(k)}Q)^2 x^{(k)} \leq \lambda_{\max}((\mathbb{I} - \alpha^{(k)}Q)^2)||x^{(k)}||^2$. Then finally we have

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \sqrt{\lambda_{\max}((\mathbb{I} - \alpha^{(k)}Q)^2)}.$$

As we know that smaller β brings about a faster convergence rate, it is advantageous to minimize the upper bound as much as we can. We have the freedom to choose stepsize $\alpha^{(k)}$. Then, what is optimal $\alpha^{(k)}$ that minimizes the upper bound?

Constant step size

Let the eigenvalues of Q as $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$. Then eigenvalues of $\mathbb{I} - \alpha^{(k)}Q$ are $\{1 - \alpha^{(k)}\lambda_1, 1 - \alpha^{(k)}\lambda_2, \cdots, 1 - \alpha^{(k)}\lambda_n\}$, and eigenvalues of $(\mathbb{I} - \alpha^{(k)}Q)^2$ are $\{(1 - \alpha^{(k)}\lambda_1)^2, (1 - \alpha^{(k)}\lambda_2)^2, \cdots, (1 - \alpha^{(k)}\lambda_n)^2\}$. Let $\lambda_{\max}(Q) = m$ and $\lambda_{\min}(Q) = M$. Then it is known (or easy to check) that $\lambda_{\max}\left((\mathbb{I} - \alpha^{(k)}Q)^2\right) = \max\{(1 - \alpha^{(k)}m)^2, (1 - \alpha^{(k)}M)^2\}$. Therefore, we have

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \max\left\{\left|1 - \alpha^{(k)}m\right|^2, \left|1 - \alpha^{(k)}M\right|^2\right\}.$$

Now, this upper bound provides an answer on how to choose $\alpha^{(k)}$. By simple math, it is easy to find that $\frac{2}{m+M} = \arg\min_{\alpha^{(k)}}\left(\max\left\{\left|1 - \alpha^{(k)}m\right|^2, \left|1 - \alpha^{(k)}M\right|^2\right\}\right)$ and setting $\alpha^{(k)} = \frac{2}{m+M}$ provides its upper bound to be $\frac{M-m}{M+m} = \frac{\text{c.d.}(Q)-1}{\text{c.d.}(Q)+1}$.

To wrap up, we approximate $f(x)$ to $\frac{1}{2}x^\top Qx$ within a small region and use a gradient algorithm for optimal stepsize, then we have the factor β that satisfies as follows,

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \frac{\text{c.d.}(Q) - 1}{\text{c.d.}(Q) + 1}.$$

It is easy to check that $0 < \beta = \frac{\text{c.d.}(Q)-1}{\text{c.d.}(Q)+1} < 1$, so this method is linear convergence that rate depends on condition number.

6 Lecture 6

Different step size

Let's use exact line search to find optimal $\alpha^{(k)}$. Again, we have a gradient algorithm: $x^{(k+1)} = x^{(k)} - \alpha^{(k)}\nabla f(x^{(k)})$. We can find optimal $\alpha^{(k)}$ that satisfies

$$\alpha^{(k)} = \arg\min_{\alpha \geq 0} f(x^{(k) - \alpha\nabla f(x^{(k)})}.$$

Since the function is convex with respect to α, the optimal α satisfies $0 = \frac{\partial f(x^{(k)} - \alpha\nabla f(x^{(k)}))}{\partial \alpha}$. Let's call $g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)}$. Then the optimal α should satisfy

$$-(g^{(k)})^\top \nabla f(x^{(k)} - \alpha g^{(k)}) = 0.$$

Since we know that $\nabla f(x^{(k)} - \alpha g^{(k)}) = 0 = Q(x^{(k)} - \alpha g^{(k)})$ holds. Then the optimal $\alpha^{(k)}$ would be expressed as

$$\alpha^{(k)} = \frac{(g^{(k)})^\top g^{(k)}}{(g^{(k)})^\top Q g^{(k)}}.$$

Then, by some computation, we have the following expression as

$$f(x^{(k+1)}) = \left(1 - \frac{\left((g^{(k)})^\top (g^{(k)})\right)^2}{\left((g^{(k)})^\top Q(g^{(k)})\right)\left((g^{(k)})^\top Q^{-1}(g^{(k)})\right)}\right) f(x^{(k)}).$$

Now, apart from the previous constant step size case, let's define the error $e(x^{(k)}) := f(x^{(k)}) - f(x_*)$. As we know, we are minimizing the quadratic function so by SOC condition, we have $f(x_*) = 0$. Then we have

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \le \left(\frac{\text{c.d.}(Q) - 1}{\text{c.d.}(Q) + 1}\right)^2$$

. Note that the constant stepsize case and different step size case are not directly comparable since how we define the error $e(x^{(k)}$ is different. However, one important lesson is that the factor β in both cases has a relationship with the condition number, c.d.(Q).

6.1 When $f(x)$ holds strong convexity

We are still touching down the problem that approximating the problem $\min f(x)$ to $\min \frac{1}{2}x^\top Qx$ where $Q = \nabla^2 f(x_*)$. Let's assume that there exists $m, M > 0$ such that $m\mathbb{I} \preccurlyeq \nabla^2 f(x) \preccurlyeq M\mathbb{I}$ holds for all x.

What if such m, M doesn't exist for all x? We need to consider sub-level set $\{x \mid f(x) \le f(x^{(0)})\}$ and then we can define such m, M over this set. There are a few things that we can check

1. Since $m\mathbb{I} \preccurlyeq \nabla^2 f(x_*)$ holds, we can say SOC sufficient holds.
2. $\frac{M}{m}$ is an upper bound on c.d.(Q)
3. Recall that we have the inequality $f(x) - f(x_*) \le \frac{||\nabla f(x)||^2}{2m}$ where m says that the gradient is small, then we are close to a solution.

We want to check the convergence rate within this case, with 1) exact line search and 2) backtracking.

Exact line search

Let's start with how the exact line search shapes the gradient algorithm

$$\begin{aligned}
f(x^{(k+1)}) &= f(x^{(k)} - \alpha^{(k)}\nabla f(x^{(k)})) \\
&= \min_{\alpha \ge 0} f(\underbrace{x^{(k)}}_{\text{nomial point}} - \underbrace{\alpha^{(k)}\nabla f(x^{(k)})}_{\text{perturbation}}) \\
&= \min_{\alpha \ge 0} \left(f(x^{(k)}) + \nabla f(x^{(k)})^\top(-\alpha\nabla f(x^{(k)})) + \frac{1}{2}(-\alpha\nabla f(x^{(k)}))^\top \nabla^2 f(\underbrace{z^{(k)}}_{\text{new point}})(-\alpha\nabla f(x^{(k)})) \right)
\end{aligned}$$

Note that Taylor expansion of the second equation provides the third equation and we can replace the higher order terms larger than the third order by introducing a new point $z^{(k)}$ (mean value theorem). Now, note that by assumption, we have $\nabla^2 f(z^{(k)}) \le M\mathbb{I}$ holds. So we have the following

$$\begin{aligned} f(x^{(k+1)}) &\leq \min_{\alpha \geq 0} \left(f(x^{(k)}) + \left(-\alpha + \frac{M\alpha^2}{2} \right) \left|\left| \nabla f(x^{(k)}) \right|\right|^2 \right) \\ &\leq f(x^{(k)}) - \frac{1}{2M} \left|\left| \nabla f(x^{(k)}) \right|\right|^2 \end{aligned}$$

Note that in the second line inequality, the $\arg\min$ is obtained at $\alpha = 1/M$. Now we use the fact that $-2m(f(x^{(k)}) - f(x_*)) \geq -||\nabla f(x^{(k)})||^2$ holds. We define the error $e(x^{(k)}) := f(x^{(k)}) - f(x_*)$. Then we have the following inequality.

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq 1 - \frac{m}{M}$$

So, in case of exact line search, the error sequence **linearly converges** with a factor $\beta = 1 - m/M = 1 - 1/\text{c.d.}$

When to stop?

So far, we have investigated the algorithm's convergence rate. Then, if we know the algorithm we have come up with converges, it is natural to ask what would be a good k to stop the iteration. Specifically, if we are given a task $\min_x f(x)$, then the sequence $\{x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots \to x^{(k)} \to \cdots\}$ goes infinity and we need to stop at some point k.

Let's say we stop at point k if $f(x^{(k)}) - f(x_*) \leq \epsilon$ is satisfied. Note that ϵ is a hyperparameter that a user can choose before executing iteration. Since in the above example, we have defined the error term $e(x^{(k)}) := f(x^{(k)}) - f(x_*)$, then we stop when $e(x^{(k)}) \leq \epsilon$ holds. We have computed that the error sequence satisfies $e(x^{(k)}) \leq (1 - \frac{m}{M})^k e^{(0)}$, let's finish our iteration at time k when $(1 - \frac{m}{M})^k e^{(0)} \leq \epsilon$ is satisfied. Then, we have

$$k \geq \frac{\log \left(\frac{f(x^{(0)}) - f(x_*)}{\epsilon} \right)}{\log \left(1 - \frac{m}{M}\right)^{-1}}.$$

We say the **number of iteration** we need is

$$k = \mathcal{O}(\log \frac{1}{\epsilon}).$$

Also the number of iterations is proportional to log of initial optimality gap $f(x^{(0)}) - f(x_*)$. One remark that we can make through this is that if the denominator term ($\log \left(1 - \frac{m}{M}\right)^{-1}$) takes a huge impact on lower bound, then we use the method *heavy ball method.*

Backtracking

Recall that backtracking is given by $\alpha \to \alpha\beta \to \alpha\beta^2 \to \cdot \to \alpha\beta^{i_k}$ where setting $\alpha^{(k)} = \alpha\beta^{i_k}$ satisfy the Armijo rule $f(x^{(k)} + \alpha^{(k)}\Delta x^{(k)}) - f(x^{(k)}) \leq \sigma \nabla f(x^{(k)})^\top (\alpha^{(k)} \Delta x^{(k)})$ for $0 > \sigma < 1$. Since we are dealing with gradient descent algorithm, i.e. $\Delta x^{(k)} = -\nabla f(x^{(k)})$, Armijo rule is modified as

$$f\,(\underbrace{x^{(k)} - \alpha^{(k)}\nabla f(x^{(k)})}_{x^{(k+1)}}) - f(x^{(k)}) \leq -\sigma\alpha^{(k)}||\nabla f(x^{(k)})||^2.$$

For any non-negative number j, we can do taylor expansion on $f(x^{(k)} + \alpha\beta^j\Delta x^{(k)}) = f(x^{(k)}) + \nabla f(x^{(k)})^\top(\alpha\beta^j\Delta x^{(k)}) + \frac{1}{2}(\alpha\beta^j\Delta x^{(k)})^\top\nabla^2 f(z^{(k)})(\alpha\beta^j\Delta x^{(k)})$. Since we know we are using gradient descent,i.e. $\Delta x^{(k)} = -\nabla f(x^{(k)})$ and by assumption we have $\nabla f(z^{(k)}) \leq M\mathbb{I}$. So we have the following Taylor expansion with assumption inequality.

$$f(x^{(k)} + \alpha\beta^j\Delta x^{(k)}) - f(x^{(k)}) \leq -\left(1 - \frac{M\alpha\beta^j}{2}\right)(\alpha\beta^j)\left|\left|\nabla f(x^{(k)})\right|\right|^2.$$

Now, let's take a look at the coefficient of RHS of the above two inequalities: σ and $1-\frac{M\alpha\beta^j}{2}$. We know that $0 < \sigma < 1$ holds and as $j \to \infty$, $1 - \frac{M\alpha\beta^j}{2} \to 1$. This means there exist j that satisfies $1 - \frac{M\alpha\beta^j}{2} \geq \sigma$. Let $\mu \in \mathbb{N} \cup \{0\}$ to be the smallest nonnegative integer such that $1 - \frac{M\alpha\beta^\mu}{2} \geq \sigma$ holds. Then we have two scenarios,

1. If $\mu = 0$, then α is small
2. if $\mu > 0$ then $1 - \frac{M\alpha\beta^{\mu-1}}{2} < \sigma$ is satisfied.

By armijo rule, $\alpha^{(k)} \geq \alpha\beta^\mu$ holds. Then Armijo rule provides the following inequalities

$$\begin{aligned} f(x^{(k+1)}) - f(x^{(k)}) &\leq -\sigma\alpha^{(k)}||\nabla f(x^{(k)})||^2 \\ &\leq -\sigma\alpha\beta^\mu||\nabla f(x^{(k)})||^2 \end{aligned}$$

Now, we apply the assumption $-||\nabla f(x^{(k)})||^2 \leq -2m(f(x^{(k)}) - f(x_*))$ and really that we have defined the error as $e(x^{(k)}) := f(x^{(k)}) - f(x_*)$. Then we have

$$e(x^{(k+1)}) \leq e(x^{(k)})\left(1 - \sigma\alpha\beta^\mu(2m)\right).$$

Let's say scenario 2 happens ($\mu > 0$), which means $1 - \frac{m\alpha\beta^{\mu-1}}{2} \leq \sigma$ holds. This leads us to come up with

$$-\alpha\beta^\mu < \frac{2(1-\sigma)\beta}{M}.$$

Plugging the above inequality into the error inequality, then we finally have

$$\frac{e^{(k+1)}}{e^{(k)}} \leq 1 - \left(4\sigma\,(1-\sigma)\,\beta\left(\frac{m}{M}\right)\right).$$

One important remark we can come up with is that if $\sigma \to 1/2$ and $\beta \to 1$ make the backtracking method similar to the exact linear search method. Both case's β is goes to $1 - m/M$.

6.2 When $f(x)$ holds convexity

Convergence rate

So far, we have looked up the convergence rate when $f(x)$ holds strong convexity, i.e. existence of $m, M > 0$ such that $m\mathbb{I} \preccurlyeq \nabla^2 f(x) \preccurlyeq M\mathbb{I}$ holds for all x. However, what if m

does not exist? Let's focus on the case where $\nabla^2 f(x) \geq 0$ holds for all x.

Definition 6.1 (little o notation)**.** A sequence $\{e^{(k)}\}$ satisfies $\lim_{k\to\infty} \frac{e^{(k)}}{1/k} = 0$, then we say $e^{(k)} = o(\frac{1}{k})$. This could be interpreted as $1/k$ converges faster than a sequence $\{e^{(k)}\}$. We say if the

Theorem 6.2 (Sublinear convergence of convexity function)**.** Assume that $\nabla^2 f(x) \succcurlyeq 0$ holds. We define X_* as a set of global solutions and assume X_* is non-empty and bounded. We also have the following additional three assumptions as follows:

1. $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$.
2. $\exists c > 0$ such that $\nabla f(x^{(k)})^\top \Delta x^{(k)} \leq -c||\nabla f(x^{(k)})||^2$ holds in case of gradient algorithm $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$.
3. $\alpha^{(k)} \in [\epsilon, (2-\epsilon)\bar{\alpha}^{(k)}]$

Then, within above assumptions, all limit points of $\{x^{(k)}\}$ are optimal, and $e^{(k)} := f(x^{(k)}) - f(x_*) = o(\frac{1}{k})$

Proof. Based on the preview theorme, we know that $\{e(x^{(k)})\} \to 0$ as $k \to \infty$ □

When to stop?

One implication of above theorem is that $e^{(k)} = o(\frac{1}{k})$ holds. This means there exists $q > 0$ such that $e^{(k)} \leq \frac{q}{k}$ satisfies. Then, let's say we want to stop the iteration when $e^{(k)} \leq \epsilon$ holds. We can compute the iteration k that satisfies $\frac{q}{k} \leq \epsilon$, where $k = \mathcal{O}(\frac{1}{\epsilon})$

$m > 0$	$m = 0$
$k = \mathcal{O}(\log 1/\epsilon)$(Subsection 6.1)	$k = \mathcal{O}(1/\epsilon)$ (Theorem 6.2)

Table 1: Iteration number of strongly convex and convex case

Let's say we want to guarantee the accuracy of L digits, i.e. $\epsilon = 10^{-L}$. Then $m > 0$ provides a complexity to a linear in number of digits (L) and $m = 0$ provides a complexity to an exponential in number of digits (L).

7 Lecture 7

So far, we have talked about gradient method $x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$. However, could we do better than that? Let's think about the following two *acceleration* methods: The heavy-ball method and Nestrov's acceleration method.

7.1 Heavy ball method

Let's think about the following descent method.

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}) + \beta^{(k)} (\underbrace{x^{(k)} - x^{(k-1)}}_{\text{momentum}})$$

where we could think the term $x^{(k)} - x^{(k-1)}$ as momentum. If the stepsize $\alpha^{(k)}$ and $\beta^{(k)}$ are constant as α and β, then we call this **Heavy-ball method**.

7.1.1 Casestudy: quadratic function optimization

Let;s think about the problem of min $\frac{1}{2}x^\top Qx$ where $Q \succ 0$. Note that for the PD matrix, its minimum eigenvalue $m > 0$. What we have shown in equation (5.2) is that using constant step size for gradient algorithm yields the error bound as

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \frac{\text{c.d.}(Q) - 1}{\text{c.d.}(Q) + 1}$$

. For an acceleration method (Equation (7.1)), we will show that the error bound holds as

$$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \frac{\sqrt{\text{c.d.}(Q)} - 1}{\sqrt{\text{c.d.}(Q)} + 1}$$

for some choices of α and β. Comparing the upper bound between equation (5.2), which comes from the gradient method, and equation (6), which comes from the acceleration method, provides an insight that the acceleration method could do better when the condition number of Q is ill-conditioned. For example, if $\text{c.d.}(Q) = 10^4$, then the acceleration method provides a much faster convergence rate.

Recall how we start the derivation of the gradient method in the Equation (5.2). Then it is easy to come up with that this sort of acceleration method is good when $\nabla^2 f(x_*) > 0$. Note that the iteration number concerning error bound ϵ does not change. This is because when we compute the iteration number of k, we are doing $\left(\frac{\text{c.d.}(Q)-1}{\text{c.d.}(Q)+1}\right)^k \leq \epsilon$ for gradient method and $\left(\frac{\sqrt{\text{c.d.}(Q)}-1}{\sqrt{\text{c.d.}(Q)}+1}\right)^k \leq \epsilon$ for acceleration method. Therefore, in both cases, the number of iterations is the same as $k = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$. What the acceleration method improves is the constant of $\mathcal{O}$.

7.2 Nestrov's acceleration method

Let's think about the following alternative method

$$\begin{cases} y^{(k)} = x^{(k)} + \beta^{(k)} \left(x^{(k)} - x^{(k-1)}\right) \\ x^{(k+1)} = y^{(k)} - \alpha \nabla f(y^{(k)}) \end{cases} \tag{6}$$

The first equation of Equation (6) is using intermediate parameter $y^{(k)}$ and the second equation of Equation (6) is applying gradient algorithm to an intermediate point $y^{(k)}$. Assume $\beta^{(k)} \to 1$ as $k \to \infty$.

Theorem 7.1.

8 Lecture 8

8.1 Newton method

Theorem 8.1 (Newton-like method)**.** Consider the algorithm $x^{(k+1)} = x^{(k)} + \alpha^{(k)}\Delta x^{(k)}$. We have the following assumptions.

1. Assume that $\{x^{(k)}\} \to x_*$ where $\nabla f(x_*) = 0$ (FOC) and $\nabla^2 f(x_*) \succcurlyeq 0$ (SOC) holds.
2. Assume that for $\nabla f(x^{(k)}) \neq 0$ for all k and
$$\lim_{k\to\infty} \frac{||\Delta x^{(k)} - (\nabla^2 f(x_*))^{-1}\nabla f(x^{(k)})||}{||\nabla f(x^{(k)})||} = 0$$
.

Let's apply Armijo rule by $\alpha = 1, 0 < \beta < 1, 0 < \sigma < \frac{1}{2}$. Then the two following holds.

1. It holds that
$$\lim_{k\to\infty} \frac{||x^{(k+1)} - x_*||}{||x^{(k)} - x_*||} = 0$$
which means $\{x^{(k)}\}$ converges superlinearly.
2. It holds that there exists $\bar{k} \geq 0$ such that $\alpha^{(k)} = 1$ for all $k \geq \bar{k}$, which means there is no reduction by Armijo rule after some time.

Applying the capture theorem to Theorem (8.1) guarantees the assumption of Theorem (8.1) is satisfied. To be specific, recall **capture theorem** (Theorem (5.1)). It guarantees that if $x^{(0)}$ is in a neighborhood of x_*, then $\{x^{(k)} \to x_*\}$, which guarantees the first assumption of Theorem (8.1) is satisfied. Now let's apply$\Delta x^{(k)} = \nabla^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$, which is a **newton method**. Then, the second assumption of Theorem (8.1) is also satisfied. Then we can conclude the following theorem

Theorem 8.2 (Superlinear convergence of Newton method)**.** Let's assume $x^{(-)}$ is close to x_* (Capture theorem holds). Then the Newton method has a superlinear convergence.

The superlinear convergence of Theorem 8.2 does not mean that $\{x^{(k)}\}$ converges to x_* as *extremely fast*. It's just *slightly faster* than linear. However, under some extra assumptions, we can guarantee that Newton's method has quadratic convergence as the following Theorem 8.3.

8.2 Newton's method when the initial point is close to the optimal point.

We first show when the initial point $x^{(0)}$ is within a set $S_\delta := \{x|||x - x_*|| \le \delta\}$.

Theorem 8.3 (Quadratic convergence of Newton's method)**.** Assume that there exist $L > 0, m > 0, \delta > 0$ that the followings hold

1. (Existence of tensor) $||\nabla^2 f(x) - \nabla^2 f(y)||_2 \le L||x - y||_2$ for $\forall x, y \in S_\delta$.
2. (Strongly convex) $\nabla^2 f(x) \ge m\mathbb{I}$ for $\forall x \in S_\delta$.
3. (Small δ) $\frac{L\delta}{2m} < 1$.

where $S_\delta := \{x|||x - x_*|| \le \delta\}$. Then if the initial point is $x^{(0)}$ is inside a set S_δ, i.e. $x^{(0)} \in S_\delta$, then the following holds.

1. (Invariant set) $x^{(k)} \in S_\delta$.
2. (Quadratic convergence) $||x^{(k+1)} - x_*|| \le \frac{L}{2m}||x^{(k)} - x_*||^2$.

Proof. proof by induction. Let's first assume that 2. (Quadratic convergence) holds. Then, it is easy to show that 1. (Invariant set) holds. Then, all we need to show is 2. (Quadratic convergence). □

8.3 Newton's method for arbitrary initial point

We then show for arbitrary initial point $x^{(0)}$, how the convergence happens with perspective as set $D_\eta := \{x|||\nabla f(x)|| < \eta\}$

Theorem 8.4 (Quadratic convergence of Newton's method)**.** Assume that there exist $L > 0, m > 0, M > 0$ that the following holds

1. (Strong convexity) $m\mathbb{I} \preccurlyeq \nabla^2 f(x) \preccurlyeq M\mathbb{I}$ for $\forall x$.
2. (Existence of tensor) $||\nabla^2 f(x) - \nabla^2 f(y)||_2 \le L||x - y||_2$ for $\forall x$.

Now, run Newton's method with the Armijo rule such that $\alpha = 1, 0 < \beta < 1, 0 < \sigma < 1/2$. Define two constants $\eta = 3(1 - 2\sigma)\frac{m^2}{L}$ and $\gamma = \sigma\beta\eta^2\frac{m}{M^2}$. Now let's define a set $D_\eta := \{x|||\nabla f(x)|| < \eta\}$. Then the following holds

1. If $x^{(k)} \notin D_\eta$, then $f(x^{(k+1)}) - f(x^{(k)}) < -\gamma$.
2. If $x^{(k)} \in D_\eta$, then $\alpha^{(k)} = 1$ have quadratic convergence.

Remark 8.5 (Insight from Theorem 8.4)**.** One important lesson from Theorem 8.4 is that σ takes a role of the tradeoff between set size and iteration from $x^{(0)}$ to $x^{(k)}$. If σ is large, then η becomes small and γ becomes small. Note that from $x^{(0)}$ to $x^{(k)}$, it took iteration of $\frac{f(x^{(k)}-f(x^{(0)})}{\gamma}$. This means as γ becomes small, then it takes a large iteration to get to the set D but the set size η becomes large.

8.4 Stopping criterion for Newton's method

For Newton's method, the stopping criterion is not $||\nabla f(x^{(k)}|| \leq \epsilon$. Recall what we have done to decide on stopping criteria in Subsection 6.1 and Subsection 6.2. Instead, we will use the following criteria:

$$\frac{1}{2}\nabla f(x^{(k)})^\top \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)}) \leq \epsilon \tag{7}$$

Let's take a look at where Equation (7) comes from. **Recall that the fundamental idea of Newton's method is just approximating $f(x)$ with quadradic function.** Think about the Taylor expansion as follows.

$$f(x^{(k)} + \Delta x^{(k)}) = f(x^{(k)}) + f(x^{(k)})^\top \Delta x^{(k)} + \frac{1}{2}(\Delta x^{(k)})^\top \nabla^2 f(x^{(k)})(\Delta x^{(k)})$$

Then if we take $f(x^{(k)} + \Delta x^{(k)}) - f(x^{(k)})$ and recall that Newton's method uses $\Delta x^{(k)} := -(\nabla^2 f(x))^{-1}\nabla f(x)$. Then we get Equation (7) $= f(x^{(k)} + \Delta x^{(k)}) - f(x^{(k)})$. This means **Equation** (7) **is a difference between** $f(x)$ **and its quadratic approximation**.

8.5 Number of iterations for Newton's method

Theorem 8.6.

$$\begin{aligned}\text{Number of iteration} &= \text{iteration get to set } D + \text{iteration of converge inside } D \\ &\leq \frac{f(x^{(0)} - f(x_*)}{\gamma} + \log_2 \log_2 \left(\frac{2m^3/L^2}{\epsilon}\right) \\ &= \mathcal{O}\left(\log \log \frac{1}{\epsilon}\right)\end{aligned}$$

In Theorem 8.6, note that $\log\log$ provides almost constant value regardless of ϵ.

8.6 Newton's method for nonconvex problem and arbitrary initial point

Recall what we have done in Subsection 8.2 and Subsection 8.3. We have assumed that strong convexity holds, i.e. $\nabla^2 f(x) \succcurlyeq m\mathbb{I}$. Now, what if $f(x)$ is **nonconvex**? Suppose that $\nabla^2 f(x) \succ 0$ does not hold. We can approach this with the following two methods:

Method1: Design corrected direction

Design $\Delta^{(k)}$ such that $\Delta^{(k)} + \nabla^2 f(x^{(k)}) \succ 0$ holds and use it as

$$\Delta x^{(k)} = -\left(\Delta^{(k)} + \nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$$

then we can show that $\Delta x^{(k)} \to 0$ as $k \to \infty$. However, the problem is we don't know how to design $\Delta^{(k)}$.

Method2: Trust region optimization

Recall that we can approximate the $f(x^{(k+1)}) = f(x^{(k)} + \Delta x^{(k)})$ as follows,

$$f(x^{(k)} + \Delta x^{(k)}) \approx f(x^{(k)}) + \nabla f(x)^\top \Delta x^{(k)} + \frac{1}{2}(\Delta x^{(k)})^\top \nabla^2 f(x^{(k)}) \Delta x^{(k)}.$$

Then rather than determining $\Delta x^{(k)}$ that minimizes $f(x^{(k)} + \Delta x^{(k)})$, we solve minimizing the approximation value as follows.

$$\min_{\Delta x} f(x^{(k)}) + \nabla f(x^{(k)})^\top \Delta x + \frac{1}{2}\Delta x^\top \nabla^2 f(x^{(k)}) \Delta x$$

We know that if $\nabla^2 f(x^{(k)}) \not\succeq 0$, the objective value have it minimum as $-\infty$. This fact reminds us that we should do a **restriction to a local region, named trust region**. Within this sense, we would like to recall "S-lemma"

Lemma 8.1 (S-lemma)**.** If the minimization problem is quadratic, and constraints are quadratic functions, then the problem has zero duality gap.

Let's utilize Lemma 8.1. Think about the following quadratic optimization problem.

$$\begin{aligned} \min_{\Delta x} \quad & f(x^{(k)}) + f(x^{(k)})^\top \Delta x + \frac{1}{2}\Delta x^\top \nabla^2 f(x^{(k)}) \Delta x \\ \text{s.t.} \quad & ||\Delta x|| \leq \gamma^{(k)} \end{aligned} \tag{8}$$

If we think the constraint as $(\Delta x)^\top \Delta x \leq (\gamma^{(k)})^2$, then we could say the problem 8 satisfies Lemma 8.1. So let's move the constraint up to the objective function. Then we can say problem 8 is equivalent to the following problem.

$$\min_{\Delta x} \quad f(x^{(k)}) + f(x^{(k)})^\top \Delta x + \frac{1}{2}\Delta x^\top \left(\nabla^2 f(x^{(k)}) + 2\lambda^{(k)}\mathbb{I}\right) \Delta x \tag{9}$$

Then we have the following optimal solution for problem 9 as follows,

$$\Delta x_* = -\left(\nabla^2 f(x^{(k)}) + 2\lambda^{(k)}\mathbb{I}\right)^{-1} \nabla f(x^{(k)}).$$

Theorem 8.7 (Trust region optimization)**.** If $\gamma^{(k)}$ is small enough, then

$$f(x^{(k)} + \Delta x_*) < f(x^{(k)}).$$

Note that $\alpha^{(k)} = 1$ (no step size).

8.7 Intermediate Wrap-up

- Newton's methods is second-order method since $x^{(k)}$ depends on second derivatives.
- Comparison between second-order methods and first-order methods.

Method	Second-order	First-order
Convergence rate	fast ($\log\log 1/\epsilon$)	slow ($1/\epsilon$ or $\log 1/\epsilon$)
Reference	Theorem 8.6	Table 6.2

8.8 Second-order is always better than first-order?

When we think about the complexity, it is composed of as following

$$\text{Complexity} = \text{Iteration number } \times \text{ Complexity per iteration}$$

Note that computing Newton's method is much more expensive than the gradient method due to Hessian. Note that computing a graident has $\mathcal{O}(n)$ complexity but computing a hessian has $\mathcal{O}(n^3)$ complexity. This leads us to an important research question. Can we design a first-order method that mimics the behavior of a second-order method? We call this **Quasi-Newton** method.

9 Lecture 9

So far we have dealt with **Newton's method**

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}).$$

9.1 Quasi Newton method

The problem was that we needed a heavy computation to compute hessian. Then our natural question would be **can we approximate the hessian?**. One way to estimate is by doing as follows.

$$\nabla^2 f(x^{k+1})(x^{(k+1} - x^{(k)}) \approx \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$$

One new algorithm we could come up with is that let's say we estimate the inverse of Hessian as $D^{(k)}$ and find the matrix $D^{(k)}$as the following equation.

$$\underbrace{D^{(k+1)}}_{\text{matrix}} \underbrace{(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}))}_{\text{vector}} = \underbrace{x^{(k+1)} - x^{(k)}}_{\text{vector}}$$

Is this a well-defined problem? Let's take a look at some cases:

1. If $n = 1$, then we can solve $D^{(k+1)}$.
2. If $n > 1$, we have n equations and $n(n+1)/2$ variables.

So we cannot find $D^{(k+1)}$ uniquely. Then what if we find an optimal $D^{(k+1)}$? Some facts are

- $D^{(k+1)}$ can't be too far away from $D^{(k)}$. This means that we have *smoothness in Hessian.* n n

This lets us compute $D^{(k)}$ as an optimization problem as follows.

$$\begin{aligned} \min_{D} \; & ||D - D^{(k)}||_Q \\ \text{s.t. } & D = D^\top \\ & D(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})) = x^{(k+1)} - x^{(k)} \end{aligned} \tag{10}$$

Note that 10 have a closed-form solution where $D^{(k+1)}$ is a function of $D^{(k)}$. Let $q^{(k)} := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$ and $p^{(k)} := x^{(k+1)} - x^{(k)}$. Then for any arbitrary $D^{(0)} \succ 0$, the closed form solution is

$$D^{(k+1)} = D^{(k)} + \left(1 + \frac{(q^{(k)})^\top D^{(k)} q^{(k)}}{(p^{(k)})^\top q^{(k)}}\right) \frac{p^{(k)}(p^{(k)})^\top}{(p^{(k)})^\top p^{(k)}} - \frac{D^{(k)} q^{(k)} (p^{(k)})^\top + p^{(k)} (q^{(k)})^\top D^{(k)}}{(p^{(k)})^\top q^{(k)}} \tag{11}$$

The equation (11) is derived by **BFGS rule** and we call this method as **Quasi Newton method**. A different way to understand quasi-newton (like the half-Newton method) is that since finding D is an optimization problem, it's not a unique estimate of $(\nabla^2 f)^{-1}$. Therefore, we can say the following does not converge.

$$\lim_{k\to\infty} D^{(k)} \not\to \lim_{k\to\infty} \nabla^2 f(x^{(k)})^{-1}.$$

However, we can say that since we have started with $x^{(k+1)} - x^{(k)}$ to estimate Hessian, the method just worked through directional derivative as follows,

$$\lim_{k\to\infty} -D^{(k)} \nabla f(x^{(k)}) \to \lim_{k\to\infty} - \underbrace{\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})}_{\text{Newton's direction}}.$$

9.2 Intermediate wrap-up

So far, what has dealt with **Gradient method**, **Quasi-newton method**, **Newton method**? Note that the following **only holds for strongly convex case**.

Method	Algorithm	Convergence rate
1st order	gradient alg.	linear
1st order	quasi-newton alg.	sublinear
2nd order	Newton alg.	quadratic

10 Lecture 10

11 Lecture 11

11.1 Algorithms for constrained optimization

$$\begin{aligned}&\min_{x\in\mathbb{R}^n} f(x)\\&\text{s.t. } x\in X\end{aligned} \tag{12}$$

11.2 Convex function

Definition 11.1 (Convex function)**.** $f(x)$ is a **convex function** if it satisfies (all equivalent definitions):

- Zero-th order convexity condition: $f(\alpha x+(1-\alpha)y)\leq \alpha f(x)+(1-\alpha)f(y)$ for $\forall x,y\in\mathbb{R}^n, \forall\alpha\in[0,1]$.
- First-order convexity condition: $f(y)\geq f(x)+\nabla f(x)^\top(y-x)$ for $\forall x,y\in\mathbb{R}^n$.
- Second-order convexity condition: $\nabla^2 f(x)\succcurlyeq 0$ for $\forall x\in\mathbb{R}^n$.

11.3 $m-$Strongly convex function

Definition 11.2 (m-strong convex function)**.** $f(x)$ is a **m-strong convex** function if it satisfies (all equivalent definitions):

- $\nabla^2 f(x)\succcurlyeq m\mathbb{I}$ for $\forall x\in\mathbb{R}^n$ where $m>0$.
- $f(y)\geq \underbrace{f(x)+\nabla f(x)^\top(y-x)+\frac{m}{2}||x-y||_2^2}_{\text{global quadratic under-estimator}}$ for $\forall x,y\in\mathbb{R}^n$.
- $f(\alpha x+(1-\alpha)y)+\underbrace{\frac{1}{2}m\alpha(1-\alpha)||x-y||^2}_{\text{quadratic function}}\leq \alpha f(x)+(1-\alpha)f(y)$
- $\underbrace{(\nabla f(x)-\nabla f(y))^\top(x-y)}_{\text{monotone operator}}\geq m||x-y||^2$ for $\forall x,y$.

For second bullet point, note that $f(x)+\nabla f(x)^\top(y-x)+\frac{m}{2}||x-y||_2^2$ is a global quadratic under-estimator, namely, the function has some curvature depending on m. For the fourth bullet point, note that its definition means $\nabla f(\cdot)$ is a monotone operator.

11.4 Convex set

Definition 11.3 (Convex set)**.** We define a set S as a convex set if $\alpha x+(1-\alpha)y\in S$ for any $x,y\in S$ and any $\alpha\in[0,1]$. Namely, its segment is also in the set.

11.5 $m-$Strongly convex set

Definition 11.4 ($m-$Strongly convex set)**.** We define a set S is a m-strongly convex set if $\alpha x + (1-\alpha)y + \frac{1}{2}\alpha(1-\alpha)||x-y||^2 z \in S$ for any $x, y \in S$, any $\alpha \in [0,1]$, and any z where $||z|| \leq 1$.

Note that the convex set is a segment and the m-strongly convex set is a boundary that has a curvature. One equivalent definition of the m-strongly convex set is given as follows

Definition 11.5 (Equivalent definition of $m-$Strongly convex set)**.** For $\forall x, y \in S$ and for a$\forall\alpha \in [0,1]$, the ball $B(\underbrace{\alpha x + (1-\alpha)y}_{\text{center}}, \underbrace{\frac{m}{2}\alpha(1-\alpha)||x-y||^2}_{\text{radius}}) \in S$

11.6 When f is convex, X is convex

Recall that in Lecture 2, we solve *unconstrained* optimization problem and derive FOC (necessary), and SOC (necessary, sufficient). In this subsection, we derive similar property as follows.

Theorem 11.6. Consider the problem $\min f(x)$ such that $x \in X$ where $f(\cdot)$ is a convex function and X is a convex set. Then every local min is a global min.

Proof. to be continued. □

11.7 When f is arbitrary, X is convex

Recall that in Lecture 2, we solve *unconstrained* optimization problem and derive FOC (necessary), and SOC (necessary, sufficient). In this subsection, we derive similar properties for *constrained* problem when f is an arbitrary function and X is a convex set.

11.7.1 FOC necessary condition

Theorem 11.7 (FOC necessary)**.** If X is a convex set and $f(x)$ is arbitrary, and if x_* is a local min, then $\nabla f(x_*)^\top(x - x_*) \geq 0$ for all $x \in X$.

Proof. to be continued. □

Note that in an arbitrary function f, we have a new term $x - x_*$ (Recall Theorem 2.1 for unconstrained optimization problem). Therefore, for further analysis, let's define new terms as follows. Before, let's recall the definition of cone K.

Definition 11.8 (Cone)**.** We define K as a cone if $\alpha x \in K$ for $\forall x \in K$ and $\forall\alpha \geq 0$.

Cone of feasible direction, $F_X(x_*)$.

Definition 11.9 (Cone of feasible direction)**.** For given convex set X and a point x_*, we define $F_X(x_*) := \{\Delta x \mid \Delta x = \alpha(x - x_*)\}$ for some $\alpha \geq 0$ and $x \in X$.

Tangent cone at x_*, $T_X(x_*)$.

Definition 11.10 (Tangent cone at x_*)**.** For given convex set X and a point x_*, we define $T_X(x_*) := \{\Delta x \mid \Delta x = 0$ or $\exists \{x^{(k)}\}_{k=1}^{\infty} \subset X$ s.t. $x^{(k)} \neq x_*,\ \lim_{k\to\infty} x^{(k)} = x_*,\ \lim_{k\to\infty} \frac{x^{(k)} - x_*}{||x^{(k)} - x_*||} = \frac{\Delta x}{||\Delta x||}\}$.

Normal cone at x_*, $N_*(x_*)$

Definition 11.11 (Normal cone at x_*, $N_X(x_*)$.)**.** For given convex set X and a point x_*, we define $N_X(x_*) := \{\Delta y \mid \Delta y^\top \Delta x \leq 0,\ \forall \Delta x \in T_X(x_*)\}$.

Within using the above definition, let's take a look at how the FOC could be rewritten

$$\begin{aligned}
&\nabla f(x_*)^\top \underbrace{(x - x_*)}_{\times\alpha} \geq 0,\ \forall x \in X \\
\iff& \nabla f(x_*)^\top \Delta x \geq 0,\ \forall \Delta x \in F_X(x_*) \\
\iff& \nabla f(x_*)^\top \qquad \underbrace{\frac{x - x_*}{||x - x_*||}}_{\text{take limit for a sequence } \{x^{(k)}\}} \qquad \geq 0 \\
\iff& \nabla f(x_*)^\top \Delta x \geq 0,\ \forall \Delta x \in T_X(x_*) \\
\iff& -\nabla f(x_*) \in N_X(x_*)
\end{aligned}$$

Note that for unconstrained problem, i.e. $X = \mathbb{R}^n$, then $N_X(x_*) = 0$, so $\nabla f(x_*) = 0$. Now, using the definitions of cones above, let's come up with a geometric intuition of FOC.

" There is no feasible descent direction". "The optimality implies that there is no feasible descent direction".

Note that if X is non-convex, then the above statement is not true. there exists x that satisfies $\nabla f(x_*)^\top (x - x_*) < 0$. This violates FOC. So $\nabla f(x_*)^\top (x - x_*) \geq 0$ for $\forall x \in X$ does not work.

However, it is easy to check that $\nabla f(x_*)^\top \Delta x \geq 0$ works for $\forall \Delta x \in T_X(x_*)$. This observation leads us to SOC's necessary condition.

11.7.2 SOC necessary condition

Theorem 11.12 (SOC necessary condition)**.** Assume X is a context set and f is a **arbitrary function**. If x_* is a local min, then $\Delta x^\top \nabla^2 f(x_*) \Delta x \geq 0$ holds for $\forall \Delta x$ such that $\Delta x \in F_X(x_*)$ and $\nabla f(x_*)^\top \Delta x = 0$ holds. **??**

Note that the above condition could be regarded as restricted Hessian.

Theorem 11.13 (SOC sufficient condition)**.** Assume X is a convex set and $f(x)$ is **arbitrary function**. x_* is a local min if it satisfies

1. FOC
2. $\Delta x^\top \nabla^2 f(x_*)\Delta x > 0$ for $\forall \Delta x$ such that $\Delta x \neq 0$, $\nabla f(x_*)^\top \Delta x = 0$, $\Delta x \in T_X(x_*)$.

??

12 Lecture 12

12.1 How to solve a constrained optimization problem?

Since this lecture started, we have been interested in solving the unconstrained problem $\min f(x)$. So we derived FOC, SOC condition in Lecture 11. Then we came up with how to solve it by the gradient method, Newton method, and coordinate gradient method. In the same sense, starting from lecture 11, we have dealt with how to solve *constrained* optimization problem. Then, we have come up with corresponding FOC, and SOC conditions in Lecture 11. How we are concerned with how to solve it.

First, recall the problem setting as follows.

$$\begin{aligned} &\min\ f(x) \leftarrow \text{arbitrary function} \\ &\text{s.t. } x \in X \leftarrow \text{convex set} \end{aligned}$$

We are utilizing the descent algorithm as

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$$

. Note that $\Delta x^{(k)}$ is a **feasible direction**, i.e. $\Delta x^{(k)} \in \boldsymbol{F_X(x^{(k)})}$. Recall the definition in Lecture 11. Therefore, to choose a good direction, note that $\Delta x^{(k)}$ should satisfy following two properties

1. (Feasible direction) $\Delta x^{(k)} \in F_X(x^{(k)})$
2. (Descent direction) $\Delta f(x^{(k)})^\top \Delta x^{(k)} < 0$

One possible choice that we can make is $\Delta x^{(k)} = \bar{x}^{(k)} - x^{(k)}$ for any point $\bar{x}^{(k)} \in X$. Since $\bar{x}^{(k)} \in X$, then we can conclude that $\Delta x^{(k)} \in F_X(x^{(k)})$ holds.

Then, we have $x^{(k+1)} = x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)}) = (1 - \alpha^{(k)})x^{(k)} + \alpha^{(k)}\bar{x}^{(k)}$. Since $x^{(k)}, \bar{x}^{(k)} \in X$ then $x^{(k+1)} \in X$ since X is a convex set. Now, how to find $\alpha^{(k)}$?

1. Limited line search : find $\alpha^{(k)}$: $\min_\alpha f(x^{(k)} + \alpha \Delta x^{(k)})$ such that $0 \leq \alpha \leq 1$.
2. Constant stepsize $\alpha^{(k)} = 1$
3. Armijo rule

12.2 Frank-wolfe

Frank-Wolfe method is a conditional gradient method. It finds the most descent direction which is feasible direction. It finds the next point $x^{(k+1)}$ by solving the following optimization problem.

$$\begin{aligned} &\min_x \; f(x^{(k)})^\top (x - x^{(k)}) \\ &\quad \text{s.t. } x \in X \leftarrow \text{compact set} \end{aligned} \tag{13}$$

then we let the optimal solution x^* as $\bar{x}^{(k)}$ and let $\Delta x^{(k)} = \bar{x}^{(k)} - x^{(k)}$. Note that we need a compact set assumption to make sure $\bar{x}^{(k)}$ is not infinity.

- insert figure

12.3 Remarks on Frank-Wolfe's complexity

However, the problem is that to find $\bar{x}^{(k)}$, we need to solve additional optimization subproblems. Please note that the most expensive part of the iteration update is solving for $\bar{x}^{(k)}$. If solving the suboptimization problem (13) takes a long time, then this method is not helpful. Hopefully, note that the suboptimization problem (13)'s objective function is linear to x. This means that X has a nice structure that allows us to compute the closed-form solution.

Example 12.1. Let we are solving a $\min f(x)$ over a simplex where $X := \{x \mid x \geq 0, \sum_{i=1}^n x_i = r$. Then suboptimization problem (13) is given as

$$\begin{aligned} &\min_x \; \frac{\partial f(x^{(k)})}{\partial x_i}(x_i - x_i^{(k)}) \\ &\quad \text{s.t. } \sum_{i=1}^n x_i = r, x_1, \cdots, x_n \geq 0 \end{aligned}$$

The closed form solution of above problem is $\bar{x}_1^{(k)}, \cdots, \bar{x}_{j-1}^{(k)}, \bar{x}_{j+1}^{(k)}, \cdots, \bar{x}_n^{(k)} = 0$ and $\bar{x}_j^{(k)} = r$ where $j = \arg\min_{i=1,\cdots,n} \frac{\partial f(x^{(k)})}{\partial x_i}$.

12.4 Convergence

Now, let's talk about convergence. Recall we are handling $x^{(k+1)} = x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)})$ where $\bar{x}^{(k)} \in X$. We need to make sure that $\bar{x}^{(k)} - x^{(k)}$ that is descent is not asymptotically orthogonal to $\Delta f(x^{(k)})$. This means that $\{\bar{x}^{(k)} - x^{(k)}\}$ should be **gradient related** (Definition 3.3). Just recall the definition of gradient-related. That means for any subsequence $\{x^{(k)}\}k \in K$ that converges to a point that satisfying FOC,

- The corresponding sequence $\{\bar{x}^{(k)} - x^{(k)}\}_{k \in K}$ is bounded
- $\limsup_{k \to \infty, k \in K} \nabla f(x^{(k)})^\top (\bar{x}^{(k)} - x^{(k)}) < 0$

Theorem 12.2. If $\{\bar{x}^{(k)} - x^{(k)}\}$ is gradient related and $\alpha^{(k)}$ is designed based on limited line search or Armijo rule, then every limit point of $\{x^{(k)}\}$ satisfies FOC. We call that point a stationary point.

12.5 Frank-worlfe: gradient related

Theorem 12.3. Frank-Wolfe method guarantees gradient-related directions.

Proof. suppose $\{x^{(k)}\}_{k\in K}$ converges to a non-stationary point $\tilde{x}$. Recall the gradient-related definition (Definition 3.3). We need to prove the following:

1. $\limsup_{k\in\infty,k\in K} ||\bar{x}^{(k)} - x^{(k)}|| < \infty$
2. $\limsup_{k\to\infty,k\in K} \nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) < 0$

1st point is true since $\bar{x}^{(k)}, x^{(k)} \in X$ and X is a compact set. For the 2nd point. Note that $\bar{x}^{(k)}$ is the optimal point. so $\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) \leq \nabla f(x^{(k)})^\top(x - x^{(k)})$ holds for $\forall x \in X$. Now, take a limit on $k \to \infty, k \in K$, then we got

$$\limsup_{k\to\infty,k\in K} \nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) \leq \nabla f(\tilde{x})^\top(x - \tilde{x}),\ \forall x \in X.$$

Since $\tilde{x}$ is not a stationary point, there exists y such that

$$\nabla f(\tilde{x})^\top(y - \tilde{x}) < 0$$

holds. Therefore, we got

$$\limsup_{k\to\infty,k\in K} \nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) < 0$$

12.6 Frank-wolfe: find stationary point

Theorem 12.4. Assume $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ holds for $\forall x, y \in X$. We don't need to use the Armijo rule or Line search. Just pick $\alpha^{(k)} = \min\left\{1, \frac{\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)})}{L||\bar{x}^{(k)} - x^{(k)}||}\right\}$. Then, every limit point of $\{x^{(k)}\}$ is a stationary point.

Proof. To be continue □

□

12.7 Frank-wolfe when f is convex

Before, let's define how hard the problem is to solve. Intuitively, it's related to how large our compact set X is. So we define the diameter of set X as follows.

$$D := \max_{x,y\in X} ||x - y||$$

Theorem 12.5. We are solving a constrained problem: $\min f(x)$ where $x \in X$ that X is a compact set. We are using an iterative method as $x^{(k+1)} = x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)})$ and compute $\bar{x}^{(k)}$ by an Frank-wolfe method. Assume

- f is convex
- $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for $x, y \in X$
- Let $D := \max_{x,y \in X} ||x - y||$

Suppose we are solving unconstrained proNow, set stepsize $\alpha^{(k)} = \frac{2}{2+k}$. Then we have

$$f(x^{(k)}) - f_* \leq \frac{2LD^2}{k+2}, \ \forall k$$

which means to guarantee $f(x^{(k)}) - f_* \epsilon$, we need

$$f(x^{(k)}) - f_* = \mathcal{O}\left(\frac{1}{\epsilon}\right)$$

complexity.

Proof. First, use that $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for $x, y \in X$ imply that the function is upper bounded by the quadratic function which its coefficient is L. That is

$$\begin{aligned} f(x^{(k+1)}) &= f\left(x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)})\right) \\ &\leq f(x^{(k)}) + \nabla f(x^{(k)})^\top \left(\alpha^{(k)}(\bar{x}^{(k)} - x^{(k)})\right) + \frac{L}{2}\left|\left|\alpha^{(k)}(\bar{x}^{(k)} - x^{(k)})\right|\right| \end{aligned} \tag{14}$$

holds. Also, since the $\bar{x}^{(k)}$ is the minimizer of $\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)})$ for $x \in X$, the following

$$\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) \leq \nabla f(x^{(k)})^\top(x_* - x^{(k)}) \tag{15}$$

holds where x_* is the optimal solution. Also, due to the convexity of $f(x)$, we have

$$\nabla f(x^{(k)})^\top(x_* - x^{(k)}) \leq f(x_*) - f(x^{(k)}) \tag{16}$$

Now, combine Equations (14),(15), and (16), we have the following

$$\begin{aligned} f(x^{(k+1)}) &\leq f(x^{(k)}) + \alpha^{(k)}(f_* - f(x^{(k)})) + \frac{L}{2}(\alpha^{(k)})^2\left|\left|\bar{x}^{(k)} - x^{(k)}\right|\right|^2 \\ &\leq f(x^{(k)}) + \alpha^{(k)}(f_* - f(x^{(k)})) + \frac{L}{2}(\alpha^{(k)})^2 D^2 \end{aligned}$$

Arrange the above inequality as follows.

$$f(x^{(k+1)}) - f_* \leq (1 - \alpha^{(k)})(f(x^{(k)}) - f_*) + \frac{L}{2}(\alpha^{(k)})^2\left|\left|\bar{x}^{(k)} - x^{(k)}\right|\right|^2.$$

Now, choose $\alpha^{(K)} = \frac{2}{2+k}$. By induction, we have

$$f(x^{(k)}) - f_* = \mathcal{O}\left(1/k\right)$$

or iteration complexity as $\mathcal{O}(1/\epsilon)$ □

13 Lecture 13

Let's wrap up what we have done so far.

	$\min f(x)$	$\min f(x)$ s.t. $x \in X$
f convex	$\mathcal{O}(1/\epsilon)$ (Theorem 6.2)	$\mathcal{O}(1/\epsilon)$ (Theorem 12.5)
f strongly convex	$\mathcal{O}(\log(1/\epsilon))$ (Subsection 6.1)	?

One guess that the iteration complexity for the contained problem when strongly convexity holds is that it can't be better than $\mathcal{O}(\log(1/\epsilon))$. The following theorem tells us the lower bound of iteration complexity exists for this case.

13.1 Frank-wolfe: when f is strongly convex

Theorem 13.1. There is a class of problems where $f(x)$ is quadratic & strongly convex, and set X is described by linear inequalities such that the sequence generated by Frank-Wolfe satisfies

$$f(x^{(k)}) - f_* \geq \frac{1}{k^{1+\epsilon}}, \ \forall \epsilon > 0$$

and infinitely many values of k.

Then, how can we attain some reasonable iteration complexity? In the case of X being a strongly convex set, we will show that Frank-Wolfe attains $\mathcal{O}(\log(1/\epsilon))$ complexity in the following theorem.

13.2 Frank-wolfe: when f is convex & X is strongly convex

Theorem 13.2. Assume

1. f is convex.
2. $\min_{x\in X} ||\nabla f(x)|| > 0$
3. $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for $x, y \in X$
4. X is a strongly convex set.

Then, if the stepsize is constant and small, we could guarantee linear convergence for the Frank-Wolfe method. Namely, the iteration complexity is

$$\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

Remark 13.3 (Remark on Theorem 13.2)**.** Theorem 13.2 doesn't require $f(x)$ to be strongly convex but an optimization sub-problem should be solved $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ times.

Proof of Theorem 13.2. Recall the definition of a strongly convex set (Definition 11.3) (**Assumption 4**). Then set $\alpha = 1/2$ where α is defined in Definition 11.3. Then we can say

$$\underbrace{\frac{1}{2}x^{(k)} + \frac{1}{2}\bar{x}^{(k)} + \frac{1}{2}m\frac{1}{2}\left(1 - \frac{1}{2}\right)\left|\left|x^{(k)} - \bar{x}^{(k)}\right|\right|^2 z}_{y^{(k)}} \in X,\ ||z|| \leq 1 \tag{17}$$

holds. We let the LHS of (17) as $y^{(k)}$. Also since we are using the Frank-Wolfe method, we can say

$$\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) \leq \nabla f(x^{(k)})^\top(y^{(k)} - x^{(k)}) \tag{18}$$

holds by its definition that $\bar{x}^{(k)}$ is the optimal point. Now, pick

$$z = -\frac{\nabla f(x^{(k)})}{||\nabla f(x^{(k)})||}.$$

Let use define the constant $C := \min_{x\in X} ||\nabla f(x)||$ (**Assumption 2**). Insert $y^{(k)}$ of Equation (17) into the Equation (18), then we got

$$\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) \leq \frac{1}{2}\nabla f(x^{(k)})^\top\left(\bar{x}^{(k)} - x^{(k)}\right) - \frac{mC}{8}\left|\left|\bar{x}^{(k)} - x^{(k)}\right|\right|^2 \tag{19}$$

Also, note that the following inequality holds.

$$\begin{aligned}\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) &\leq \nabla f(x^{(k)})^\top(y^{(k)} - x^{(k)}) &(20)\\ &\leq f(x_*) - f(x^{(k)}) &(21)\end{aligned}$$

Inequality (20) comes from the definition of the Frank-Wolfe method and Inequality (21) comes from the f convexity assumption (**Assumption 1**). Combine Equation (19) and Inequality (21), then we have the following inequality.

$$\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)}) \leq \frac{1}{2}\left(f(x_*) - f(x^{(k)})\right) - \frac{mC}{8}\left|\left|\bar{x}^{(k)} - x^{(k)}\right|\right|^2 \tag{22}$$

Also, by **Assumption 2**, the function value $f(x^{(k+1)})$ is bounded by quadratic function (quadratic over-estimator) as follows.

$$f(x^{(k+1)}) - f_* \leq f(x^{(k)}) - f_* + \nabla f(x^{(k)})^\top\left(\alpha^{(k)}\right)\left(\bar{x}^{(k)} - x^{(k)}\right) + \frac{L}{2}\left|\left|\alpha^{(k)}\left(\bar{x}^{(k)} - x^{(k)}\right)\right|\right| \tag{23}$$

Finally, combine Inequalities (22) and (23), then we have

$$\underbrace{f(x^{(k+1)}) - f_*}_{e^{(k+1)}} \leq \underbrace{\left(f(x^{(k)}) - f_*\right)}_{e^{(k)}}\left(1 - \frac{\alpha^{(k)}}{2}\right) - \left|\left|x^{(k)} - \bar{x}^{(k)}\right|\right|^2\alpha^{(k)}\left(\underbrace{\frac{-L\alpha^{(k)}}{2} + \frac{mC}{8}}_{\text{Term}I}\right). \tag{24}$$

Note that Term I of Equation (24) is positive if $\alpha^{(k)} = \alpha$(constant)=small. Then, if Term$I$ is positive, we could say

$$e^{(k+1)} \leq e^{(k)} \left(1 - \frac{\alpha}{2}\right)$$

where $1 - \frac{\alpha}{2} \in (0, 1)$. So this is **linear convergence** (Definition 5.2 and Theorem 5.3). Recall that if linear convergence holds, then it guarantees $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ (See Subsection 6.1 how linear convergence is related with *log* iteration). □

13.3 Gradient Projection method

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \alpha^{(k)}(\bar{x}^{(k)} - x^{(k)}) \\ \bar{x}^{(k)} &= \mathcal{P}_X\left(x^{(k)} - s^{(k)}\nabla f(x^{(k)})\right) \end{aligned} \tag{25}$$

Note that $\alpha^{(k)}, s^{(k)}$ are step sizes, $\mathcal{P}_X$ is a projection operator. We define projection operator $\mathcal{P}_X$ as

$$\mathcal{P}_X(y) := \arg\min_{x \in X} ||x - y||^2.$$

Basically, $\mathcal{P}_X$ conducts a minimization over a quadradic function over X. The basic idea is that we use the gradient method on $x^{(k)}$ to find a better point with respect to $f(\dot{)}$. It may not be feasible, so we project it onto X to get a feasible point $\bar{x}^{(k)}$, and then get a feasible direction $\bar{x}^{(k)} - x^{(k)}$.

Note that

$$\begin{aligned} \bar{x}^{(k)} &= \arg\min_{x \in X}\left(\left|\left|x - \left(x^{(k)} - s^{(k)}\nabla f(x^{(k)})\right)\right|\right|^2\right) && (26) \\ &= \arg\min_{x \in X}\left(\left|\left|x - x^{(k)}\right|\right|^2 + 2s^{(k)}\nabla f(x^{(k)})^\top\left(x - x^{(k)}\right) + \left(s^{(k)}\right)^2\left|\left|\nabla f(x^{(k)})\right|\right|^2\right) && (27) \\ &= \arg\min_{x \in X}\left(\frac{1}{2s^{(k)}}\left|\left|x - x^{(k)}\right|\right|^2 + \nabla f(x^{(k)})^\top\left(x - x^{(k)}\right)\right) && (28) \end{aligned}$$

Look at the Equation 28. This allows us to interpret the Gradient Projection method is actually a **Regularized Frank-Wolfe** version. To be specific. Compare two methods as follows.

$$\begin{cases} \text{Frank-Wolfe} : \min_{x \in X}\left(\nabla f(x^{(k)})^\top\left(x - x^{(k)}\right)\right) \\ \text{Gradient Prjection method} : \min_{x \in X}\left(\nabla f(x^{(k)})^\top\left(x - x^{(k)}\right) + \frac{1}{2s^{(k)}}\left|\left|x - x^{(k)}\right|\right|^2\right) \end{cases}$$

13.4 Remarks on Gradient Prjection method's complexity

However, there is also one problem (as we have talked in previous subsection 12.3, to make Gradient Prjection method to be useful, the suboptimizatio problem (solving quadratic optimization problem over compact set X) should be much easier than solving original problem.

Example 13.4. Let a compact set $X := \{x \mid a_i \leq x_i \leq b_i,\ i = 1, \cdots, n\}$. Then it is easy to check that projection operator $\mathcal{P}_X(y)$ has a closed form solution as

$$i^{\text{th}} \text{ entry} = \begin{cases} a_i \text{ if } y_i \leq a_i \\ b_i \text{ if } y_i \geq a_i \\ y_i \text{ if } a_i \leq y_i \leq b_i \end{cases}$$

13.5 Special case: Projected Gradient method

Recall the method (25). When $\alpha^{(k)} = 1$, then we have $x^{(k+1)} = \bar{x}^{(k)} = 1$ which makes shrinkage to the following method

$$x^{(k+1)} = \mathcal{P}_X\left(x^{(k)} - s^{(k)}\nabla f(x^{(k)})\right) \tag{29}$$

What this implies is that we are using regular gradient method, but whenever the point gets outside of the sets, the algorithm projects it back into X.

13.6 Properties of projection operator

To do some convergence analysis, let us first investigate some mathematical properties of projection operator $\mathcal{P}_X(y)$.

Proposition 13.5 (properties of projection operator $\mathcal{P}_X(y)$)**.** The projection operator $\mathcal{P}_X(y) := \arg\min_{x\in X} ||x - y||$ where X is a compact set has the following properties:

1. $\mathcal{P}_X(y)$ is unique.
2. z is projection of y on X iff $(x - z)^\top(y - z) \leq 0$ for $\forall x \in X$.
3. For $y_1, y_2 \in \mathbb{R}^n$, it holds that
$$||\mathcal{P}_X(y_1) - \mathcal{P}_X(y_2)|| \leq ||y_1 - y_2||$$
4. FOC of constrained optimization problem (Theorme 11.7) is same as
$$\mathcal{P}_X\left(x_* - s\nabla f(x_*)\right) = x_*$$

Let's prove the 4th property of Proposition 13.5. Recall the FOC of constarined problem is that for a point x_*, $\nabla f(x_*)^\top (x - x_*) \geq 0$ holds for $\forall x \in X$. This means

$$-s\nabla f(x_*)^\top (x - x_*) \leq 0$$

holds for $\forall x \in X$ and any arbitrary $s > 0$. This could be rewrited as follows.

$$\left(\left(x_* - s\nabla f(x_*)^\top\right) - x_*\right)(x - x_*) \leq 0$$

Then by 2nd property of Proposition 13.5, regard $y = \left(x_* - s\nabla f(x_*)^\top\right)$ and $z = x_*$. Then we have

$$\mathcal{P}_X\left(x_* - s\nabla f(x_*)\right) = x_*$$

Lemma 13.1. By property 4 of Proposition 13.5, the Gradient projection methods stop if and only if the algorithm finds a stationary point.

13.7 Step size of Gradient projection method

As we have discussed in Subsections 2.3 and 3.1, stepsizes of gradient projection method also could be found by following method. Just note that we have additional step size $s^{(k)}$ beside $\alpha^{(k)}$.

1. Limited line search
   - $s^{(k)} = s = \text{constant}$
   - $\alpha^{(k)} : \min_{\alpha\in[0,1]} f(x^{(k)} + \alpha(\bar{x}^{(k)} - x^{(k)}))$.
2. Armijo rule along the feasible direction
   - $s^{(k)} = s = \text{constant}$
   - $\alpha^{(k)} : 1 \to \beta \to \beta^2 \to \cdots \text{ s.t. } f(x^{(k+1)}) - f(x^{(k)}) \leq \sigma\alpha^{(k)}\nabla f(x^{(k)})^\top(\bar{x}^{(k)} - x^{(k)})$