1 Convergence related questions in optimization

Given an optimization problem:

- Convergence to a set of stationary points
  - Typical minimum requirement
- Convergence rate (recap)
  - Error function, e.g., \( e(x) = \|x - x^*\| \)
  - Asymptotic behavior (linear, sublinear, superlinear)
- Iteration complexity analysis
  - Number of iterations necessary to achieve \( \epsilon \)-optimal solution
    \[ e(x^r) \leq \epsilon \]

2 Convergence to stationary points

Convergence to a stationary point may not be trivial.
Example: Consider \( f(x) = x^2, x^r = (-1)^r(1 + \frac{1}{r}) \). Then, we have

\[
f(x^{r+1}) = \left(1 + \frac{1}{1+r}\right)^2 \leq \left(1 + \frac{1}{r}\right)^2 = f(x^r),
\]
but the algorithm will not converge to the stationary point. Therefore,
\[ f(x^{r+1}) \leq f(x^r) \nRightarrow x^r \to x^*, \nabla f(x^*) = 0. \]

### 2.1 Gradient related condition

For any subsequence \( \{x^r\}_{r \in \kappa} \) converging to a non-stationary point, the corresponding subsequence \( \{d^r\}_{r \in \kappa} \) is bounded and
\[
\lim_{r \to \infty, r \in \kappa} \sup \nabla f(x^r)^T d^r < 0.
\]

Example: \( d^r = -D^r \nabla f(x^r) \) with \( \bar{\gamma} I \succeq D^r \succeq \underline{\gamma} I \succ 0 \) \( \forall r \).

### 2.2 General result

**Lipschitz gradient:** \( \exists L > 0 \) s. t. \( \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|, \forall x, y \in \mathbb{R}^n. \)

**Lemma (descent lemma).**
\[
f(x + h) \leq f(x) + h^T \nabla f(x) + \frac{L}{2} \| h \|^2
\]

**Proof.** Define \( g(t) = f(x + th) \). Then,
\[
f(x + h) - f(x) = g(1) - g(0) = \int_0^1 g'(t) dt = \int_0^1 \langle h, \nabla f(x + th) \rangle dt
\]
\[
= \int_0^1 \langle h, \nabla f(x + th) - \nabla f(x) \rangle dt + \int_0^1 \langle h, \nabla f(x) \rangle dt
\]
\[
\leq \int_0^1 \| h \| \| \nabla f(x + th) - \nabla f(x) \| dt + \langle h, \nabla f(x) \rangle
\]
\[
\leq \int_0^1 L t \| h \|^2 dt + \langle h, \nabla f(x) \rangle = \frac{L}{2} \| h \|^2 + \langle h, \nabla f(x) \rangle. \quad \blacksquare
\]

**Theorem.** Assume
- \( x^{r+1} \leftarrow x^r + \alpha^r d^r \)
- \( d^r \) gradient related
- Lipschitz gradient

2
• One of the following step-size rules:
  – a) Diminishing: \( \alpha^r \to 0, \sum_r \alpha^r = \infty \)
  – b) Armijo
  – c) Small enough: \( 0 < \epsilon \leq \alpha^r \leq (2-\epsilon) \frac{\|\nabla f(x^r)\|^T d^r}{L\|d^r\|^2} \)

Then, every limit point of the iterates is a stationary point, i.e.,

\[ \{x^r\}_{r \in \kappa} \to \bar{x} \Rightarrow \nabla f(\bar{x}) = 0. \]

**Proof of part (a)**
Proof is by contradiction. Assume the contrary that \( \{x^r\}_{r \in \kappa} \to \bar{x} \) but \( \bar{x} \) is not stationary. Therefore, by the gradient related condition, we should have \( \nabla f(x^r)^T d^r \leq C < 0 \) for large enough \( r \). Furthermore, after some point, the step-size \( \alpha^r \) is small enough and by the descent lemma, we have \( f(x^{r+1}) - f(x^r) \leq \frac{\epsilon}{2} \alpha^r \|\nabla f(x^r)\|^T d^r \leq \frac{\epsilon}{2} \alpha^r C \leq 0 \). Summing up all inequalities, we have that

\[
\sum_{r=0}^{t} f(x^{r+1}) - f(x^r) \leq \sum_{r=0}^{t} \frac{\epsilon}{2} \alpha^r \|\nabla f(x^r)\|^T d^r \leq \sum_{r=0}^{t} \frac{\epsilon}{2} \alpha^r C \leq 0.
\]

\[
\Rightarrow f(x^{t+1}) - f(x^0) \leq \frac{\epsilon}{2} C \sum_{r=0}^{t} \alpha^r \leq 0.
\]

By letting \( t \to \infty \) we have,

\[
f(\bar{x}) - f(x^0) \leq \frac{\epsilon}{2} C \sum_{r=0}^{\infty} \alpha^r \leq 0
\]

from which we conclude \( \sum_{r=0}^{\infty} \alpha^r \) is upper bounded by a constant which contradicts the assumption that \( \sum_{r=0}^{\infty} \alpha^r = \infty \).

**Proof of part (b)**
Proof by contradiction: assume the contrary that \( \{x^r\} \to \bar{x} \) and \( \bar{x} \) is not a stationary point. Let us assume that \( \alpha^r = \beta^{i_r} \) with \( \beta \) is as defined in the definition of Armijo rule. If \( i_r = 0, \forall r \geq r_0 \) then the proof is the same as the proof in part (c). Therefore, let us assume that \( i_r \geq 1 \) for a subsequence
of iterates. By restricting to a subsequence if necessary, we can write

$$f(x^r) - f(x^r + \alpha^r d^r) \geq -\sigma \alpha^r \nabla f(x^r)^T d^r \geq 0,$$

and

$$f(x^r) - f(x^r + \frac{\alpha^r}{\beta} d^r) < -\sigma \frac{\alpha^r}{\beta} \nabla f(x^r)^T d^r.$$  \hspace{1cm} (2)

Defining $p^r = \frac{d^r}{\|d^r\|}$ and $\bar{\alpha}^r = \alpha^r \frac{\|d^r\|}{\beta}$, (2) can be rewritten as

$$\frac{f(x^r) - f(x^r + \bar{\alpha}^r p^r)}{\bar{\alpha}^r} < -\sigma \nabla f(x^r)^T p^r.$$  \hspace{1cm} (3)

Since $\{p^r\}$ belongs to a compact set (the surface of unit ball), it has a limit point $\bar{p}$. On the other hand, letting $r \to \infty$ in (1) implies that $\alpha^r d^r \to 0$. Taking the limit $r \to \infty$ in (3) leads to

$$-\nabla f(\bar{x})^T \bar{p} \leq -\sigma \nabla f(\bar{x})^T \bar{p}.$$  

Since $\sigma < 1$, we conclude that $\nabla f(\bar{x})^T \bar{p} \geq 0$ which contradicts the gradient related condition.

\textbf{Proof of Part (c)}

Proof by contradiction: We have $f(x^r + \alpha^r d^r) - f(x^r) \leq \alpha^r \nabla f(x^r)^T d^r + \frac{L}{2} (\alpha^r)^2 \|d^r\|^2$ from Descent Lemma. Suppose $\{x^r\}_{r\in\kappa} \to \bar{x}$ but $\bar{x}$ is not stationary.

$$f(x^r + \alpha^r d^r) - f(x^r) \leq \alpha^r \nabla f(x^r)^T d^r + \frac{L}{2} \alpha^r \|d^r\|^2 (2 - \epsilon) \frac{\|\nabla f(x^r)^T d^r\|}{L\|d^r\|^2}$$

$$= \alpha^r \nabla f(x^r)^T d^r \left(1 - \frac{2 - \epsilon}{2}\right)$$

$$= \frac{\epsilon}{2} \alpha^r \nabla f(x^r)^T d^r \leq 0$$  \hspace{1cm} (4)

which implies that $f(x^{r+1}) \leq f(x^r)$. If $x^r \to \bar{x}$ then $f(x^r) \to f(\bar{x})$. Therefore by taking limit as $r \to \infty$ in (4), we obtain $\lim_{r \to \infty} \nabla f(x^r)^T d^r = 0$ which contradicts gradient related condition in slide 7.
3 Non-singular Convex Quadratic Programming

Given the following optimization problem:

\[
\min_x f(x) = \frac{1}{2} x^T Q x
\]

We will study the convergence analysis of the gradient descent algorithm:

\[
x^{r+1} = x^r - \alpha^r \nabla f(x^r)
\]

Note that \( f(x) \) is a non-singular convex function, so \( Q > 0 \iff x^T Q x > 0 \ \forall \ x \neq 0 \). Thus the optimal solution is \( x^* = 0 \).

\[
\|x^r+1\|^2 \|x^r\|^2 = \|x^r - \alpha^r Q x^r\|^2 (\text{Using } \nabla f(x^r) = Q x^r)
\]

\[
= \frac{(x^r - \alpha^r Q x^r)^T (x^r - \alpha^r Q x^r)}{\|x^r\|^2}
\]

\[
= \frac{(x^r)^T (x^r) - 2 \alpha^r (x^r)^T Q x^r + (\alpha^r)^2 (x^r)^T Q^T Q x^r}{\|x^r\|^2}
\]

\[
= \frac{(x^r)^T (I - \alpha^r Q) x^r}{\|x^r\|^2}
\]

\[
\leq \lambda_{\max}((I - \alpha^r Q)^2) \quad (\text{where } \lambda_{\max} \text{ is the maximum eigenvalue function})
\]

Let the eigenvalues of \( Q \) be: \( m = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = M \), then eigenvalues of \( (I - \alpha^r Q)^2 \) are: \( (1 - \alpha^r \lambda_1)^2, (1 - \alpha^r \lambda_2)^2, \ldots, (1 - \alpha^r \lambda_n)^2 \).

We get \( \lambda_{\max}((I - \alpha^r Q)^2) = \max\{|1 - \alpha^r m|^2, |1 - \alpha^r M|^2\} \).

Let \( g(\alpha) = \max\{|1 - \alpha m|, |1 - \alpha M|\} \), then \( \alpha^* = \arg \min_\alpha g(\alpha) \) occurs when \( 1 - \alpha m = -1 + \alpha M \).
This implies:

\[ 1 - \alpha^* m = -1 + \alpha^* M \iff \alpha^* = \frac{2}{M + m} \]

\[ g(\alpha^*) = 1 - \frac{2m}{M + m} = \frac{M - m}{M + m} = \frac{M/m - 1}{M/m + 1} = \frac{\kappa - 1}{\kappa + 1} \quad (\text{where } \kappa = M/m) \]

Thus using the optimal step-size \( \alpha^* = \frac{2}{M + m} \), we get

\[ \frac{||x^{r+1} - x^*||^2}{||x^r - x^*||^2} \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^2 \]

Similarly,

Using the optimal step-size \( \alpha^* = \frac{2}{M + m} \), we get

\[ \frac{f(x^{r+1}) - f(x^*)}{f(x^r) - f(x^*)} \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^2 \]

**Observations:**

- We have a linear rate of convergence.

- This can be generalized to most locally non-singular optimization problems

- Define the conditional number to be \( \kappa \triangleq \frac{M}{m} \), then with a large conditional number we get a Zig-Zag behavior. Thus the number of iterations before convergence is higher.
Similar to the above analysis, we can show

\[
\frac{f(x^{r+1}) - f(x^*)}{f(x^r) - f(x^*)} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2
\]

\[
\iff f(x^r) - f(x^*) \leq (f(x^{r-1}) - f(x^*))\left(\frac{\kappa - 1}{\kappa + 1}\right)^2
\]

\[
\iff f(x^r) - f(x^*) \leq (f(x^{r-2}) - f(x^*))\left(\frac{\kappa - 1}{\kappa + 1}\right)^4
\]

\[\vdots\]

\[
\iff f(x^r) - f(x^*) \leq (f(x^0) - f(x^*))\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2r}
\]

Let \( D = f(x^0) - f(x^*) \), then \( \varepsilon = f(x^r) - f(x^*) \approx \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2r} D \)

\[
\iff \log(D/\varepsilon) = -2r \log\left(\frac{\kappa - 1}{\kappa + 1}\right)
\]

\[
\iff \log(D/\varepsilon) = 2r \frac{2}{\kappa} \text{ for large } \kappa
\]

\[
\iff r = O(\kappa \log(D/\varepsilon))
\]

Thus for \( \varepsilon \)-optimal solutions, we need \( O(\kappa \log(D/\varepsilon)) \) iterations where \( D \triangleq f(x^0) - f(x^*) \).

Note if \( \kappa \) is 100 times more, then we need 100 times more iterations. However, if \( D \) is 100 times more, then we need 2 times more iterations. Thus \( \kappa \) is a more important factor.

4 General Strongly Convex Objective

Definition:

1. Let \( f \) be continuously differentiable, then \( f \) is strongly convex with modulus \( \sigma > 0 \iff f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}||y - x||^2, \forall y, x \)

2. Let \( f \) be twice continuously differentiable, then \( f \) is strongly convex with modulus \( \sigma > 0 \iff \nabla^2 f(x) \geq \sigma I \ \forall x \)

Note that if \( f \) is twice continuously differentiable, then the Lipschitz con-
stant of the gradient $L$ gives:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y|| \quad \forall x, y \iff LI \geq \nabla^2 f(x) \quad \forall x$$

Therefore, Lipschitz constant of the gradient and the strongly convex modulus are related to the eigenvalues of Hessian. More precisely, $\lambda_{\text{max}}(\nabla^2 f(x)) = L$ and $\lambda_{\text{min}}(\nabla^2 f(x)) = \sigma$. The ratio between these two numbers is defined as the condition number, i.e.,

$$\kappa \triangleq \frac{L}{\sigma}.$$

Notice that this definition is consistent with our previous definition of condition number for quadratic problems.