Gradient Descent Complexity: Note the difference:

\[ \text{Strict Convexity: } f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y) \]

\[ \text{Strong Convexity: } \exists m > 0 \text{ s.t. } \nabla^2 f(x) \succeq mI \text{ for all } x \in S \]

\( e^x \) is strictly convex but not strongly convex.

Big Picture Results for Gradient Descent from last lecture, assuming \( f \) strongly convex:

- **Exact line search** at least as good as \( t = \frac{1}{M} \)

Which gives:

\[
f(x^{(t)}) - p^* \leq (1 - \frac{1}{K}) (f(x^{(0)}) - p^*)
\]

where:

\[ K = M/m \]

\[ m = \inf_{x \in S} \lambda_{\min} (\nabla^2 f(x)) \]

\[ M = \sup_{x \in S} \lambda_{\max} (\nabla^2 f(x)) \]

\[ \Rightarrow \text{ #iter for any } \epsilon = o(\frac{1}{\log \epsilon}) \]

Hence: BACKTRACKING LINE SEARCH

- \( \alpha \in (0, \frac{1}{2}) \) ARMijo Cond.
- \( \beta \in (0, 1) \) Backtrack Param

When \( \beta = \frac{1}{2}, M \geq \frac{1}{2} \), set

\[
f(x^{(t)}) - p^* \leq (1 - \frac{1}{K}) (f(x^{(0)}) - p^*)
\]
Gradient Descent, cont'd.

Nesterov (Eqn 2.1.15) also gives another complexity result for

\[ t = \frac{2}{m + M} \]

namely

\[ \| x^{(k)} - x^* \| \leq \left( \frac{1 - \frac{1}{\sqrt{k}}}{1 + \frac{1}{\sqrt{k}}} \right)^k \| x^{(0)} - x^* \| \]

which leads to

\[ f(x^{(k)}) - p^* \leq K \left( \frac{1 - \frac{1}{\sqrt{k}}}{1 + \frac{1}{\sqrt{k}}} \right)^{2k} (f^{(0)} - p^*) \]

where \( x^* \) is minimizer, \( p^* = f(x^*) \).

Both BV and Nesterov also give results for the non-strongly convex case — these are much weaker.
Newton's Method

\[ \Delta x \text{ is solution of } \nabla^2 f(x^{(k)}) \Delta x = -\nabla f(x^{(k)}) \]

Motivation: minimize "quadratic model"

\[ q(v) = \nabla f(x^{(k)})^T v + \frac{1}{2} v^T \nabla^2 f(x^{(k)}) v \]

To solve equation, use CHOLESKY FACTORIZATION

\[ \nabla^2 f(x^{(k)}) = LL^T \]

\[ L L^T \Delta x = \begin{bmatrix} -g \\ y \end{bmatrix} \]

1) forward solve for \( y \)
2) backward solve for \( \Delta x \)

Cost: \( \frac{1}{6} n^3 \) adds +.mults

Use same backtracking line search.

(Newton used this in finding zeros of polynomials, not minimization; particularly, root of \( p(\lambda) = \lambda^2 - c \) i.e. square roots)
Convergence Analysis of Newton's Method.

As before, suppose that $S = \{x : f(x) \leq f(x_0) \}$ is compact and $M \geq \nabla^2 f(x) \geq mI$ a.s., $m > 0$.

Now also need

$$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L ||x - y|| \quad \forall x, y \in S$$

i.e. $\nabla^2 f$ is Lipschitz.

Turns out (BV §3.5) that $\exists \eta > 0$, $\epsilon > 0$ s.t.

$$\begin{cases}
\text{if } ||\nabla f(x^{(k)})|| \geq \eta, \text{ then B.T.H.S. returns } t_k = 2M/||\nabla f(x^{(k)})|| \\
\text{while } ||\nabla f(x^{(k)})|| < \eta, \text{ B.T.H.S. returns } t_k = 1 \\
\text{with } \frac{1}{2m^2} \frac{L}{\eta^2} \leq \left( \frac{1}{2m^2} \frac{L}{\eta^2} \right)^2 \left( \frac{1}{2m^2} \frac{L}{\eta^2} \right)^2 (*)
\end{cases}$$

\[ \text{"QUADRATIC CONVERGENCE".} \]

Consequence: If $m \leq \frac{\eta^2}{L}$ and, in some $x^*, ||\nabla f(x^{(k)})|| \leq \eta$ then
\[ \left\| \nabla f(x^{(k+1)}) \right\| \leq \frac{L}{2m^2} \frac{\eta^2}{\eta^2} \leq \frac{1}{2} \eta \]

so this applies recursively and hence \((*)\) holds for all \(k \geq K\), so

\[
\frac{1}{2m^2} \left\| \nabla f(x^{(k)}) \right\| \leq \left( \frac{1}{2m^2} \left\| \nabla f(x^{(k)}) \right\| \right)^{l-K} \leq \left( \frac{1}{2} \right)^{l-K} \quad (\dagger)
\]

\(l=K\) \hspace{1cm} \(l=K+1\) \hspace{1cm} \(l=K+2\) \hspace{1cm} \(l=K+3\)

\(K+5 \quad \frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{16} \quad \frac{1}{256} \quad \text{quad. conv.}\)

But what are \(\eta, \gamma\)? Turns out (BV p. 489-491) that

we always have

\[ t_k \geq \frac{\beta m}{M} \]

and that consequently

\[ f(x^{(k+1)}) - f(x^{(k)}) \leq -\alpha \beta \frac{\eta^2 m}{M^2} \]

so \((\dagger)\) holds if we set \(\eta \geq \frac{M}{M^2} \beta \alpha \gamma\).

It also turns out that if \(\eta \leq 3(1-2\alpha) \frac{m^2}{M^2} \)

\hline
Then \(t_k = 1\), i.e. \(f(x^{(k)} + \alpha x^{(k)})\) satisfies the Newton step

"sufficient decrease" condition in the BTLSS.

We will now show that \((*)\) holds as a consequence.

\(\text{QUAD CONTR.}\)
Proof of (\textcolor{red}{\star}) (quadratic contraction) assuming \( t_k = 1 \).

\[
\| \nabla f(x^{(k)} + \Delta x_{NT}) \| = \| \nabla f(x^{(k)} + \Delta x_{NT}) - \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)}) \Delta x_{NT} \| \\
\text{ZERO BY DEF.}
\]

\[ \int_0^1 \nabla^2 f(x^{(k)} + s \Delta x_{NT}) \Delta x_{NT} \, ds \\
\text{find the calc.}
\]

\[
= \left\| \int_0^1 (\nabla^2 f(x^{(k)} + s \Delta x_{NT}) - \nabla^2 f(x^{(k)})) \Delta x_{NT} \, ds \right\|
\]

\[
\leq \int_0^1 \| \nabla^2 f(x^{(k)} + s \Delta x_{NT}) \| \| \Delta x_{NT} \| \, ds
\]

\[
= \frac{L}{2} \| \Delta x_{NT} \|^2
\]

\[
= \frac{L}{2} \| (\nabla^2 f(x^{(k)}))^{-1} \nabla f(x^{(k)}) \|^2
\]

\[
\leq \frac{L}{2m} \| \nabla f(x^{(k)}) \|^2 \equiv (\textcolor{red}{\star})
\]

Note: In order to apply recursively, also need \( \eta \leq \frac{m^2}{L} \) as explained in NM2 (bottom). So, need

\[
\eta = \min \{ 1, 3(1-2\varepsilon) \} \frac{m^2}{L}.
\]
Total # iterations:

Initial phase with \( \| \nabla f(x) \| \geq \gamma \)

\[ \# \text{steps} \leq \frac{f(x_0) - p^*}{\gamma} \text{ immediately in (1).} \]

Quadratically convergent phase with \( \| \nabla f(x) \| < \gamma \)

\[ f(x^{(k)}) - p^* \leq \frac{1}{2m} \| \nabla f(x^{(k)}) \|^2 \leq \frac{1}{2m} \frac{4m^4}{L^2} \left( \frac{1}{2} \right)^{2^{l-K+1}} \]

\[ = \frac{2m^3}{L^2} \left( \frac{1}{2} \right)^{2^{l-K+1}} \]

If want RHS \( \leq \varepsilon \), set LHS \( \leq \varepsilon \), need

\[ \left( \frac{1}{2} \right)^{2^{l-K+1}} \leq \frac{\varepsilon L^2}{2m^3} \]

\[ 2^{l-K+1} \geq \frac{\varepsilon_0}{\varepsilon} \text{ where } \varepsilon_0 = \frac{2m^3}{L^2} \]

\[ l - K + 1 \geq \log_2 \frac{\varepsilon_0}{\varepsilon} \]

\[ \# \text{steps} = l - K + 1 \geq \log_2 \log_2 \frac{\varepsilon_0}{\varepsilon} \]

* e.g. \( \frac{\varepsilon_0}{\varepsilon} = 10^{-15} \)

\[ \log_2 10^{-15} = \log_2 2^{50} = 50 \]

\[ \log_2 50 \leq 6 \]

Very few steps until quadratic's converged starts.
Lower Complexity Bounds
(Mirzov - Sec 2.14)

Assume as before that \( f \) is strongly convex \( \mathcal{C}^2 \) with
\[
\begin{align*}
\mathcal{L}_I & \leq \nabla^2 f(x) \leq \mathcal{L}_I \quad \text{for all } x \in \mathcal{S},
\end{align*}
\]
\( \mathcal{L}_I \) in Mirror, \( \mathcal{L}_I \) in Mirror.

Assume that at each point \( x^{(k)} \), a "first-order oracle" or "black box" computes \( f(x^{(k)}) \) and \( \nabla f(x^{(k)}) \).

Assume also that for \( k = 1, 2, \ldots \)

\textbf{(ASSUMPTION)} \quad x_k \in \text{lin span} \{ \nabla f(x_0), \ldots, \nabla f(x_{k-1}) \}.

In simplicity, assume dom \( f = \mathbb{R}^\infty = \mathbb{R}_2^\infty \),
\[
\begin{align*}
\{ x = (x_i)_{i=1}^\infty : \|x\|^2 = \sum_{i=1}^\infty x_i^2 < \infty \}
\end{align*}
\]

Now we define a "difficult" function \( F \) by
\[
F(x) = \max \left\{ \frac{M-m}{8} \left[ (x_1)^2 + \sum_{i=1}^{\infty} (x_i - x_{i+1})^2 - 2x_1 \right] \right\}
+ \frac{m}{2} \|x\|^2.
\]

We have
\[
\begin{align*}
\frac{\partial F}{\partial x_1} &= \frac{M-m}{8} \left( 2x_1 + 2(x_1 - x_2) - 2 \right) + mx_1,
\end{align*}
\]
\[
\begin{align*}
\frac{\partial F}{\partial x_i} &= \frac{M-m}{8} \left( 2(x_i - x_{i+1}) - 2(x_{i-1} - x_i) \right) + mx_i,
\end{align*}
\]
\[
\begin{align*}
&= \frac{M-m}{8} \left( 4x_i - 2x_{i+1} - 2x_{i-1} \right) + mx_i.
\end{align*}
\]
\[ \nabla^2 F(x) = \frac{M-m}{4} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} + mI \]

with

\[ \lambda_{\max}(\nabla^2 F(x)) \leq M \quad \text{as required} \]

\[ \lambda_{\min}(\nabla^2 F(x)) \geq m \]

and

\[ \nabla F(x) = \left(\frac{M-m}{4} \mathbf{T} + mI\right)x - \frac{M-m}{4} e_i \begin{bmatrix} 1 \\ 0 \end{bmatrix} \]

The solution \( x^* \) is given by \( \nabla F(x^*) = 0 \):

\[ \frac{M-m}{4} \mathbf{x}^* \begin{bmatrix} 2x_1 - x_2 \\ 2x_1 - x_2 + \frac{4}{M-m} \mathbf{M} x_1 \end{bmatrix} = \frac{M-m}{4} \]

\[ 2x_1 - x_2 + \frac{4}{M-m} \mathbf{M} x_1 = \frac{M-m}{4} \]

\[ 2x_1 - x_2 = \frac{M-m}{4} \mathbf{M} x_1 = 1 \]

\[ x_2 - \frac{2(M-m) + 4m x_1 + 1}{M-m} = 0 \]

\[ x_2 = 2 \frac{M+m}{M-m} x_1 + 1 = 0 \]

and, for \( j = 2, 3, \ldots \):

\[ x_{j+1} + 2 \frac{M+m}{M-m} x_j + x_{j-1} = 0 \]

This difference equation can be solved by plugging in \( x_j = q^j \) and solving for \( q \).
\[ q^{j+1} - 2 \frac{M+m}{M-m} q^j + q^{j-1} = 0 \]

\[ q^2 - 2 \frac{M+m}{M-m} q + 1 = 0 \]

Claim: Roots are \( M+m \pm 2\sqrt{Mm} \)

\[ \frac{M+m}{M-m} \]

check: sum of roots is then \( 2 \frac{M+m}{M-m} \) \( \checkmark \)

Product of roots: \( \left( M+m \right)^2 - 4Mm = 1 \)

\[ \frac{(M+m)^2 - 4Mm}{(M-m)^2} \]

Smaller root is

\[ q_1 = \frac{M+m - 2\sqrt{Mm}}{M-m} = \frac{(\sqrt{M} - \sqrt{m})^2}{(\sqrt{M} - \sqrt{m})(\sqrt{M} + \sqrt{m})} \]

\[ = \sqrt{M} - \sqrt{m} \]

\[ \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \]

\[ = 1 - \sqrt{1/k} \]

where \( k = \sqrt{m} \)

\[ \frac{1}{1 + \sqrt{1/k}} \]

NOTE THE SQRT.
We get Theorem 5.4 for any $x^{(0)} \in \mathbb{R}^n$ and any $m > 0$.

$$M > m, \quad \exists \text{ function } F \text{ with } m \mathbf{I} \preceq \nabla^2 F \preceq M \mathbf{I}$$

(Quadratic)

With (under assumption)

$$\mathbf{A}$$

$$\|x^{(k)} - x^*\|^2 \geq \left(1 - \frac{1}{\sqrt{1/k}}\right) \|x^{(0)} - x^*\|^2$$

where $x^*$ minimizes $F$ and $k = M/m$, and hence

$$f(x^{(k)}) - f^* \geq \frac{m}{2} \|x^{(0)} - x^*\|^2$$

**Proof.** We can take $x^{(0)} = 0$. Then, we use $F_{00}$ defined already, getting:

$$\|x^{(0)} - x^*\|^2 = \sum_{i=1}^{\infty} (x^*_i)^2 = \sum_{i=1}^{\infty} \frac{z_i^2}{1 - q^2} = \frac{z_{1(1)}}{1 - q^2}$$

Since $T$ is tridiagonal, $T^T = T$, we can show by induction that $x_j \in \text{span}(e_1, e_2, \ldots, e_j)$,

$$\|x^{(k)} - x^*\|^2 \geq \sum_{i=j+1}^{\infty} (x^*_i)^2 = \sum_{i=j+1}^{\infty} \frac{z_{i}^2}{1 - q^2} = \frac{z_{j+1}^2}{1 - q^2} = \frac{z_{j}^2}{1 - q^2} \|x^{(0)} - x^*\|^2$$

with $q = 1 - \frac{1}{\sqrt{1/k}}$. (Last inequality follows from our original (1), just Taylor theorem)

(BV (9.8))
In comparison, the gradient method with \( t = \frac{1}{M} \)
gave us

\[
f(x^{(k)}) - f^* \leq \left( 1 - \frac{1}{k} \right)^k (f(x^{(0)}) - f^*)
\]

Compare \( \frac{1 - \sqrt{k}}{1 + \sqrt{k}} \approx (1 - \sqrt{k})^2 \approx 0.999 \)

if \( k = 10^6 \), while \( 1 - \frac{1}{k} \approx 0.9999999 \).

So lower bound indicates we may be able to do much better. The rest of Nesterov’s
Chapter 2 derives "optimal gradient" method, but the argument is very
complicated! In the end, the simplest is:

**NESTEROV OPTIMAL GRADIENT ALGORITHM** (p. 81)

Choose \( y^{(0)} = x^{(0)} \in \mathbb{R}^n \)

For \( k = 0, 1, 2, \ldots \):

\[
\begin{align*}
&\text{let } x_{k+1} = y_k - \frac{1}{M} \nabla f(y_k) \\
&\text{let } y_{k+1} = x_{k+1} + q_k (x_{k+1} - x_k)
\end{align*}
\]

where \( q_k = \frac{(-\sqrt{k})}{1 + \sqrt{k}} \)