Lecture 20:

(1) More detailed convergence proof:

- Damped Newton phase:
  By strong convexity $\nabla^2 f(x) \succeq M I + x \in \mathcal{E}$:
  
  \[
  f(x + t \Delta x) \leq f(x) + t \nabla f(x) \Delta x + \frac{M}{2} \| \Delta x \|^2 \leq f(x) - t \lambda(x) - \frac{t^2 \lambda(x)^2}{2} - \frac{t \lambda(x)}{2} \|
  \]

  Noting step size $t = \frac{m}{M}$ satisfies backtracking line search, then line search must return $t \geq \frac{m}{M}$, thus

  \[
  f(x^*) - f(x) \leq -\frac{t}{2} \lambda(x) \leq -\frac{t \lambda(x)}{2} \|
  \]

  \[
  \leq -\frac{M}{m^2} \| \nabla f(x) \|^2 \leq -\frac{m^2}{M^2} \|
  \]

  Thus choose $\gamma = \frac{m^2}{M^2} \| f(x) \|^2$ satisfies the damped phase.

- Quadratic phase:
  For this phase we can always use step size $t = 1$ in backtracking search. We skip the proof for this case. Now $x^* = x + \Delta x$ and

  \[
  \| \nabla f(x^*) \|^2 = \| \nabla f(x + t \Delta x) - \nabla f(x) - \nabla^2 f(x) \Delta x \| \|^2 \leq \| \int_0^t \nabla^2 f(x + s \Delta x) \cdot \Delta x dt - \nabla^2 f(x) \Delta x \|^2 \|
  \]

  \[
  \leq \frac{1}{2} \| \Delta x \|^2 = \frac{1}{2} \| \nabla^2 f(x) \|^2 \|
  \]

  \[
  \|
  \frac{M}{m^2} \| \nabla f(x) \|^2
  \]

  Can show $\gamma = \min \int_0^1 \| 3(1 - 2x)^2 \| \frac{m^2}{2}$.}

---

- Damped Newton phase:
  By strong convexity $\nabla^2 f(x) \succeq M I + x \in \mathcal{E}$:
  
  \[
  f(x + t \Delta x) \leq f(x) + t \nabla f(x) \Delta x + \frac{M}{2} \| \Delta x \|^2 \leq f(x) - t \lambda(x) - \frac{t^2 \lambda(x)^2}{2} - \frac{t \lambda(x)}{2} \|
  \]

  Noting step size $t = \frac{m}{M}$ satisfies backtracking line search, then line search must return $t \geq \frac{m}{M}$, thus

  \[
  f(x^*) - f(x) \leq -\frac{t}{2} \lambda(x) \leq -\frac{t \lambda(x)}{2} \|
  \]

  \[
  \leq -\frac{M}{m^2} \| \nabla f(x) \|^2 \leq -\frac{m^2}{M^2} \|
  \]

  Thus choose $\gamma = \frac{m^2}{M^2} \| f(x) \|^2$ satisfies the damped phase.

- Quadratic phase:
  For this phase we can always use step size $t = 1$ in backtracking search. We skip the proof for this case. Now $x^* = x + \Delta x$ and

  \[
  \| \nabla f(x^*) \|^2 = \| \nabla f(x + t \Delta x) - \nabla f(x) - \nabla^2 f(x) \Delta x \| \|^2 \leq \| \int_0^t \nabla^2 f(x + s \Delta x) \cdot \Delta x dt - \nabla^2 f(x) \Delta x \|^2 \|
  \]

  \[
  \leq \frac{1}{2} \| \Delta x \|^2 = \frac{1}{2} \| \nabla^2 f(x) \|^2 \|
  \]

  \[
  \|
  \frac{M}{m^2} \| \nabla f(x) \|^2
  \]

  Can show $\gamma = \min \int_0^1 \| 3(1 - 2x)^2 \| \frac{m^2}{2}$.}
Summary for Newton's method:

- Convergence rapid, quadratically near $x^*$.
- Affine invariant, thus insensitive to choice of coordinate or condition number of sublevel sets.
- The condition number of sublevel sets (or the Hessian) only affects the numerical inversion of the Hessian, but has little effect on the rate of convergence. Newton's method can tolerate large condition numbers of up to $10^{10}$, whereas gradient descent tolerates far smaller numbers (practically useless for $K > 1000$).
- Newton's method scales well with problem size. For example, for problems in $IR^n$ and $IR^{1000}$ the number of iterations can be comparable.

The difficulty of the Newton's method is the cost of forming (computing) and storing the Hessian $\nabla^2 f(x)$ and the cost of computing the Newton step, requires to solve $\nabla^2 f(x) \Delta x = -\nabla f(x)$.

Roughly, the cost of computing the inverse of an $n \times n$ Hessian matrix is of order $n^3$, but for some problem we can exploit the structure of the Hessian to reduce this cost.

Some variants of Newton's method:

The goal is to reduce the computational complexity of Newton's.

- Quasi-Newton methods: Replace $\nabla^2 f(x)$ by approximation $H$.
  Many update rules for $H$ all satisfies:

  - $H = H^T > 0$
  - Secant condition: $\nabla f(x^t + 1) - \nabla f(x^t) = H^t (x^t + 1 - x^t)$
  - $H^t \nabla f(x^t)$ is more easily computed than $\nabla^2 f(x^t) \nabla f(x^t)$

- Broyden-Fletcher-Goldfarb-Shanno (BFGS): (most common)

  $y = \nabla f(x^t + 1) - \nabla f(x^t)$

  then $H^t = H + \frac{yy^T}{y^T s} - \frac{H s s^T H}{s^T H s} \rightarrow O(n^2)$
Self-concordance

- Motivation:
  - Newton's method is affine invariant, but the convergence analysis is not.
  - Often do not know constants \( m, M, L \) in practice.
  - Constants \( m, M, L \) can depend on starting point.

Self-concordance condition (Nesterov & Nemirovski):
allows a new analysis of Newton's method.

- Is affine invariant.
- Involves no unknown constants.
- Is valid for many function \( f \), including the logarithmic barrier function (where on this latter in interior-point methods).

Self-concordance condition:

Convex \( f : \mathbb{R} \rightarrow \mathbb{R} \) is self-concordant if
\[
|f'''(x)| \leq 2f''(x)^{3/2} \quad \forall x \in \text{dom} f.
\]

Examples:
  - \( f(x) = -\log x \) are SC.
  - \( f(x) = x \log x - \log x \)
  - \( f(x) = -\log \det x \)

Some simple properties:

- Affine invariant:
  \( f(Ax) \text{ SC } \Rightarrow g(x) = f(A^2 + b) \text{ is SC.} \)

- Sum and scaling:
  \( f, f \text{ SC } \Rightarrow f + f \text{ SC} \)
  \( f \text{ SC } \Rightarrow af \text{ SC.} \)

Thus:

- \( \sum_i \log (b_i - \alpha_i^TX) \text{ is SC.} \)
- \( \log \det (F_0 + X_1 F_1 + \ldots + X_n F_n) \text{ is SC.} \)
+ Analysis of Newton's method for SC functions.
   Can show that with backtracking or exact line search:
   \[ \text{# iterations} \leq \frac{\| \nabla f(x) \|- \| \nabla f(x^*) \|}{\gamma_2} + \log_2 \log_2 \left( \frac{\| \nabla f(x) \|}{\varepsilon} \right) \]
   where \( \gamma_2 \) depends only on backtracking linesearch parameters:
   \[ \gamma_2 = \beta \frac{\alpha(\gamma_2 - \alpha)^2}{5 - 2\alpha} \]

+ SC functions allow a more explicit analysis of convergence and complexity.
  It is not known if SC functions are easier to minimize than non-SC functions.