Today…

- Thus far, focused on formulating convex problems and gradient methods (first-order)
  - Now: second order and interior point methods
  - Plan: 200 pages of book (Part III) in two lectures

- Focus:
  - Convex functions
  - Twice differentiable

- Overview
  - Unconstrained
  - Equality constraints
  - General convex constraints

Good luck!
Solving unconstrained problems

- Unconstrained problem
- Sequence of points:
  \[ x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots \]
  \[ f(x^{(k)}) \rightarrow \min \]
- Exactly: Stop when
  \[ f(x^{(k)}) = \rho^* \]
- Approximately: Stop when
  \[ f(x^{(k)}) - \rho^* \leq \varepsilon \]

Descent methods

- \[ x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \]
  - Want: \[ f(x^{(k+1)}) < f(x^{(k)}) \]
- From convexity:
  \[ f(y) \geq f(x^{(k)}) + \nabla f(x^{(k)})^\top (y - x^{(k)}) \]
  if positive
- Thus: \[ \nabla f(x^{(k)})^\top (y - x^{(k)}) \geq 0 \]
  if we pick \( y \) such that \[ f(y) > f(x^{(k)}) \]
- Therefore, pick \( \Delta x \) such that:
  \[ \Delta x = y - x^{(k)} \]
  \[ \nabla f(x^{(k)})^\top (y - x^{(k)}) < 0 \]
  intuively
Generic descent algorithm

- Start from some $x$ in $\text{dom } f$
- Repeat
  - Determine descent direction $\Delta x$
  - **Line search** to choose step size $t$
  - Update: $x \leftarrow x + t \Delta x$
- Until stopping criterion

Good stopping criterion:

- In gradient descent, $\Delta x = -\nabla f(x)$
- $\|\nabla f(x)\|_2 \leq \epsilon$

Exact line search

- Find best step size $t$:
  
  \[ t = \min_{s \geq 0} f(x^{(k)} + s \Delta x) \]

- Problem is
  - $g(t) = f(x^{(k)} + s \Delta x) \in \mathbb{R}$ convex
  - Sometimes easy to solve in closed form
  - Other times can take a long time
Backtracking line search

- From convexity, lower bound on $f(x+t\Delta x)$:
  \[ f(x+t\Delta x) \geq f(x) + t \nabla f(x)^T \Delta x \]
  - Can’t really hope to achieve ideal decrease of
  - Instead pick some $\alpha \in (0,0.5)$
    - And achieve:
      \[ f(x^{(\alpha)} + t\Delta x) \leq f(x^{(\alpha)}) + \alpha \nabla f(x^{(\alpha)})^T \Delta x \]

Choosing $t$:
- $\beta \in (0,1)$

Boyd & Vandenberghe: pick
- $\alpha$ in $[0.01,0.3]$
Analysis of gradient descent

- Linear convergence rate:
  - Geometrically decreasing
    - \[ f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*) \]
  - Geometrically decreasing in log plot, error decreases below a line...

- Rate \( c \) related to “condition number” of Hessian
  - \( c \equiv 1 - 1/\text{condition number} \)

- For quadratic problem:
  - Condition number is \( \lambda_{\text{max}}/\lambda_{\text{min}} \)

- Gradient descent bad when condition number is large

Observations about descent algorithms

- Observe linear convergence in practice
- Boyd & Vandenberghe: difference often not significant in large dimensional problems
  - May not be worth implementing exact LS when complex

- Condition number can greatly affect convergence
Solving quadratic problems is easy

- Quadratic problem:
  $$\min_x \frac{1}{2} x^T P x + q^T x$$

- Solving equivalent to solving linear system:
  $$\nabla f(x) = 0; \quad \nabla^2 f(x) = P x + q = 0 \implies P x = -q$$

  - If system has at least one solution: done!
  - If system has no solutions: problem is unbounded

- Usually don’t have simple quadratic problems, but…

Newton’s method

- Second order Taylor expansion:
  $$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

  - Descent direction, solution to linear system
    $$\nabla^2 f(x) \Delta x = -\nabla f(x)$$
      solve for $\Delta x$

- Nice property:
  - We wanted: $\nabla^2 f(x) \Delta x < 0$
  - We get:
    $$\nabla f(x)^T \Delta x + = 0$$
      by convexity $\nabla f(x) \neq 0$
Newton’s method – alg.

- Start from some $x$ in $\text{dom} \ f$
- Repeat
  - Determine descent direction $\Delta x_{nt}$
    - solving a linear system $\nabla^2 f(x) \Delta x = -\nabla f(x)$
  - Line search to choose step size $t$
  - Update: $x \leftarrow x + t \Delta x_{nt}$
- Until stopping criterion

- Good stopping criterion:
  $$\frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \leq \epsilon$$
  - $f(x+\Delta x) > f(x) + \nabla f(x)^T \Delta x$ by convexity

Convergence analysis for Newton’s

- Two phases:
  - Gradient is large
    - Damped Newton Phase
      - Step size $t\leq 1$
      - Linear convergence
  - Gradient is small
    - Pure Newton Phase
      - Step size $t=1$
      - Quadratic convergence
    - $\mathcal{O}(\epsilon^2)$
    - Only lasts 6 steps

(Really see book for details.)
Summary on Newton’s

- Converges in very few iterations, especially in quadratic phase
- Invariant to choice of coordinates or affine scaling
  - Very useful property!
- Performs well with problem size, not very sensitive to parameter choices
- Can prove even cooler things when function is smooth
  - E.g., “self-concordance,” see book
  - Many implementation tricks (see book)

But…
- Forming and storing Hessian is quadratic \( O(n^3) \)
  - Can be prohibitive
- Solving linear system can be really expensive \( O(n^2) \)
- Use quasi-Newton methods
  - BFGS and others