Notes

- Extra class this Friday 1-2pm
- If you want to receive emails about the course (and are auditing) send me email

From last time

- Steepest Descent:
  - Start with guess $x^{(0)}$
  - Until converged:
    - Find direction $d^{(k)} = -\nabla f\left(x^{(k)}\right)$
    - Choose step size $\alpha^{(k)}$
    - Next guess is $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$
- **Line search**: keep picking different step sizes until satisfied
  - (Reduce a multidimensional problem to 1D)

Simplest line search

- Start with initial step-size α
  - E.g. suggested by user, or from last step of algorithm
- Check if it reduces *f* "enough":

$$f\left(x+\alpha d\right) \le f\left(x\right) - \varepsilon\alpha ?$$

- If not, halve the step and try again

$$\alpha \leftarrow \frac{1}{2}\alpha$$

  (Also, if first guess works for a few iterations in a row, try increasing step size)
- Not enough to guarantee convergence, but often does OK in practice

Convergence of Steepest Descent

- We'll use a model problem:

$$f\left(x\right) = \tfrac{1}{2} x^T A x$$

- Here A is symmetric positive definite, so 0 is the unique minimum (f is strictly convex)
- Gradient is: $\nabla f\left(x\right) = Ax$
- We can further simplify: change to eigenvector basis, A becomes diagonal

Convergence of SD cont'd

- For the benefit of the doubt, assume line-search is "perfect": picks the step to exactly minimize $f\left(x+\alpha d\right)$

$$\frac{\partial}{\partial\alpha} f\left(x+\alpha d\right) = 0$$

$$d^T\left(A(x+\alpha d)\right) = 0$$

$$\alpha = -\frac{d^T A x}{d^T A d}$$

Diagnosis

- Problem occurs for ill-conditioned A
- Quite soon the bulk of the error is along the "shallow" directions, not the steepest (gradient)
- Typical improved strategy:
  pick smarter directions
  - Conjugate Gradient (later in the course): avoid repeating the same directions
  - Newton and Quasi-Newton: try to pick the optimal direction

Newton's Method

- Use the Hessian: second derivatives
- Model the objective function as a quadratic

$$f(x+\Delta x) \approx f(x) + \frac{\partial f}{\partial x}\Delta x + \frac{1}{2}\Delta x^T \frac{\partial^2 f}{\partial x^2}\Delta x$$
$$= f(x) + g \bullet \Delta x + \tfrac{1}{2}\Delta x^T H \Delta x$$

- Minimize the model (solve a linear system)

$$0 = g + H\Delta x$$
$$\Delta x = -H^{-1} g$$

Newton's Method

- Now need to evaluate Hessian and gradient, and solve a linear system
  - Perhaps too expensive…
- But, can get quadratic convergence (# significant digits doubles each iteration)
- But can also fail in more ways
  - Hessian might be singular, indefinite, or otherwise unhelpful
  - Higher-order nonlinearity might cause divergence
  - Some of these problems can occur even if f is strictly convex

Making Newton more robust

- Modify Hessian to make it positive definite (e.g. add scaled identity): mixes in SD to guarantee descent direction ("regularization")
- Line search methods: use Newton direction, but add line search to make sure step is good
- Trust region methods: only use quadratic model in a small "trust region", don't overstep bounds (and steadily shrink trust region to convergence)

Quasi-Newton

- Attack problems with Hessian (expense and possible ill-conditioning): build approximation to Hessian from information from gradients:

$$H\Delta x \approx g(x+\Delta x) - g(x)$$

- Example: BFGS (use this for low rank updates to approximation of H)