First-order methods

Convexity

10-725 Optimization
Geoff Gordon
Ryan Tibshirani
Administrivia

• Schedule posted:
  ‣ Time for poster session: 3:30–6:30, Wed, Dec 12
  ‣ Midterm: Tue, Nov 6 (in class)
  ‣ HW1 will be released: hopefully Tue, Sep 4
  ‣ First recitations: next week

• How to do scribing:
  ‣ http://www.cs.cmu.edu/~aarti/Class/10704/lecs.html

• In case of mishaps with scribe signup sheet
Worked ex: image understanding
Edge detectors
Gradient descent

\[
\min_x f(x)
\]

- for \( k = 1, 2, \ldots \)
  - \( g_k \leftarrow \nabla f(x_k) = \frac{df}{dx} \bigg|_{x=x_k} \)
  - \( x_k \leftarrow x_{k-1} - t_k g_k \)

- Choices: \( x_0, t_k, \) when to stop
Gradient descent: example
Gradient descent: example
In ML & stats

- Often have $f(x) = \mathbb{E}_{p(i)} \left[ f_i(x) \right]$
  - where $i \sim p(i)$

- E.g., linear regression:
  $$\min_{a,b} \mathbb{E}_{a \sim \mathcal{N}(0,1)} \left[ (a \cdot x - b)^2 \right]$$

- Let: $I = \text{i.i.d. sample } \sim p(i)$
  - then $\hat{f}(x) = \frac{\sum_{i \in I} f_i(x)}{|I|}$
When do we stop?

• ML/stats: held out data
  - $f_{\text{train}}$ vs $\hat{f}_{\text{valid}}$

• Early stopping
  - Regularization
  - Why bother?
When do we stop?

- Using convergence bounds (see below)
  - usual form is:
    \[ \kappa \geq \frac{(f(x_0) - f^*)}{\text{fun of } f} \]
  - need estimates of: