First-order methods
Convexity

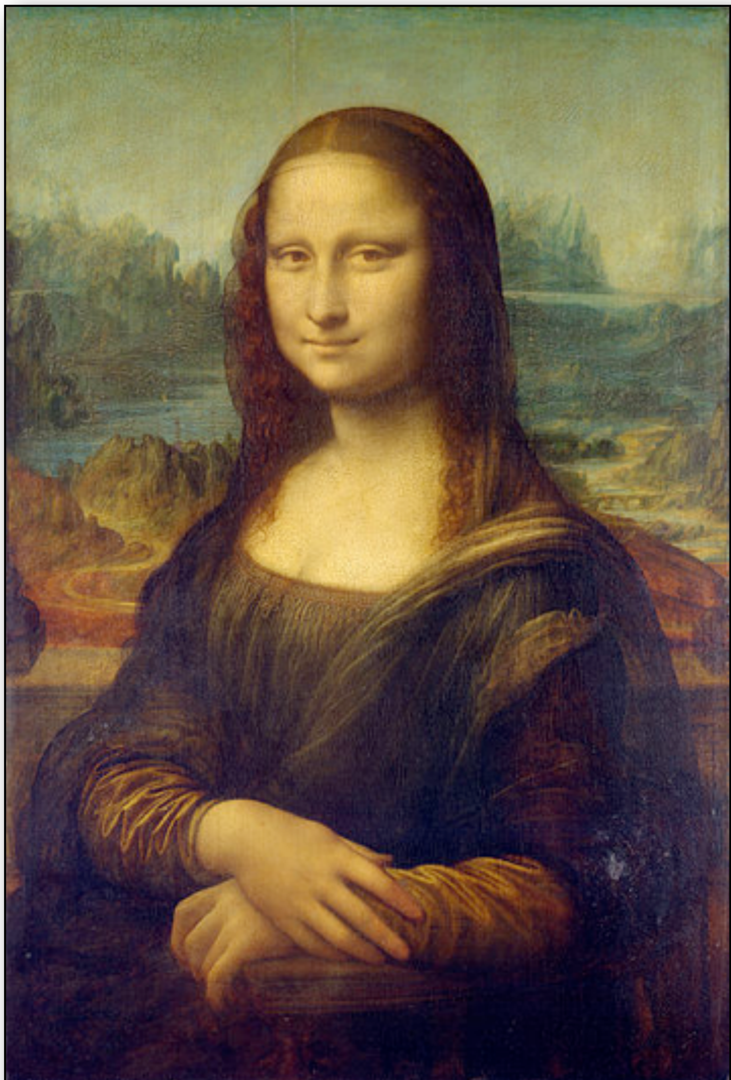
10-725 Optimization

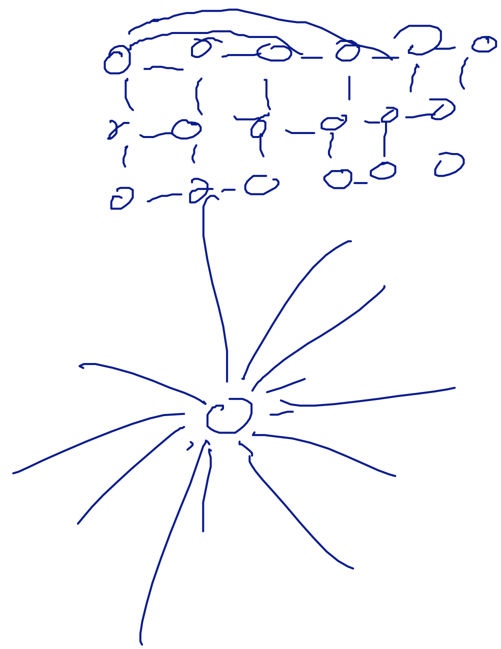
Geoff Gordon
Ryan Tibshirani

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



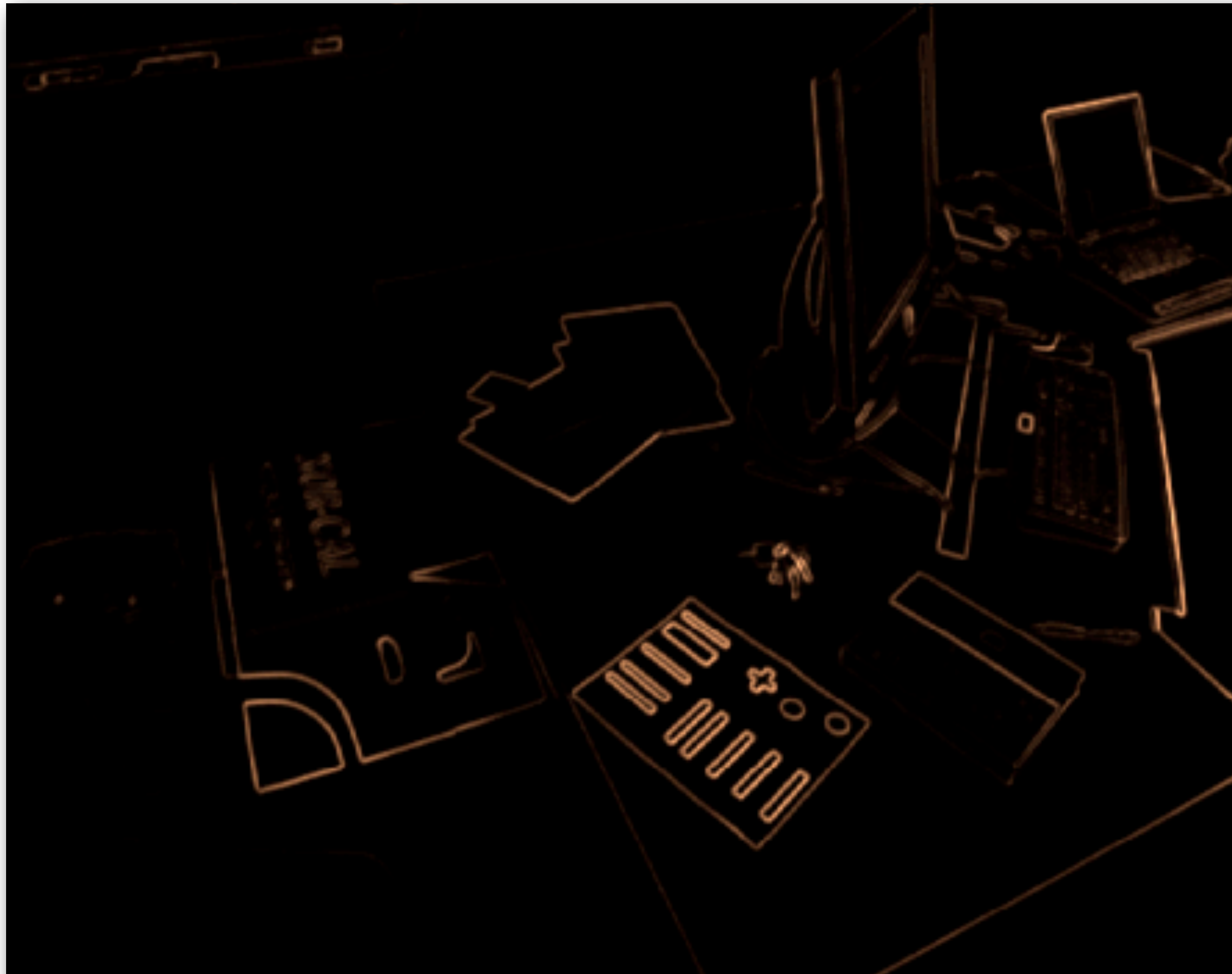


$x_i \in \{1, 2, \ldots 10\}$

$$\sum_{ij \in E} \phi(x_i, x_j)$$

$$|\{i \mid x_i = 3\}| \approx |\{i \mid x_i = 4\}|$$

Edge detectors

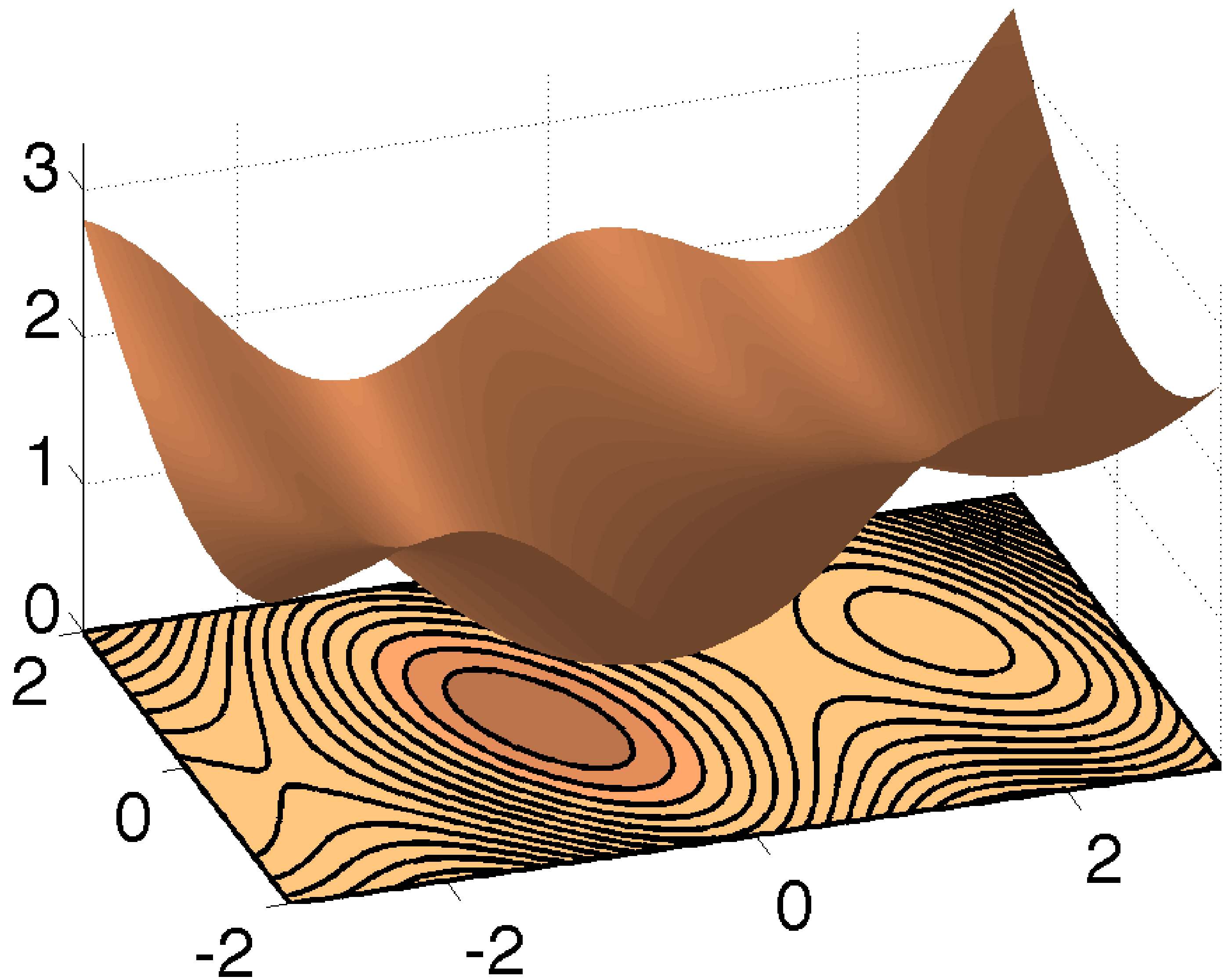


Gradient descent

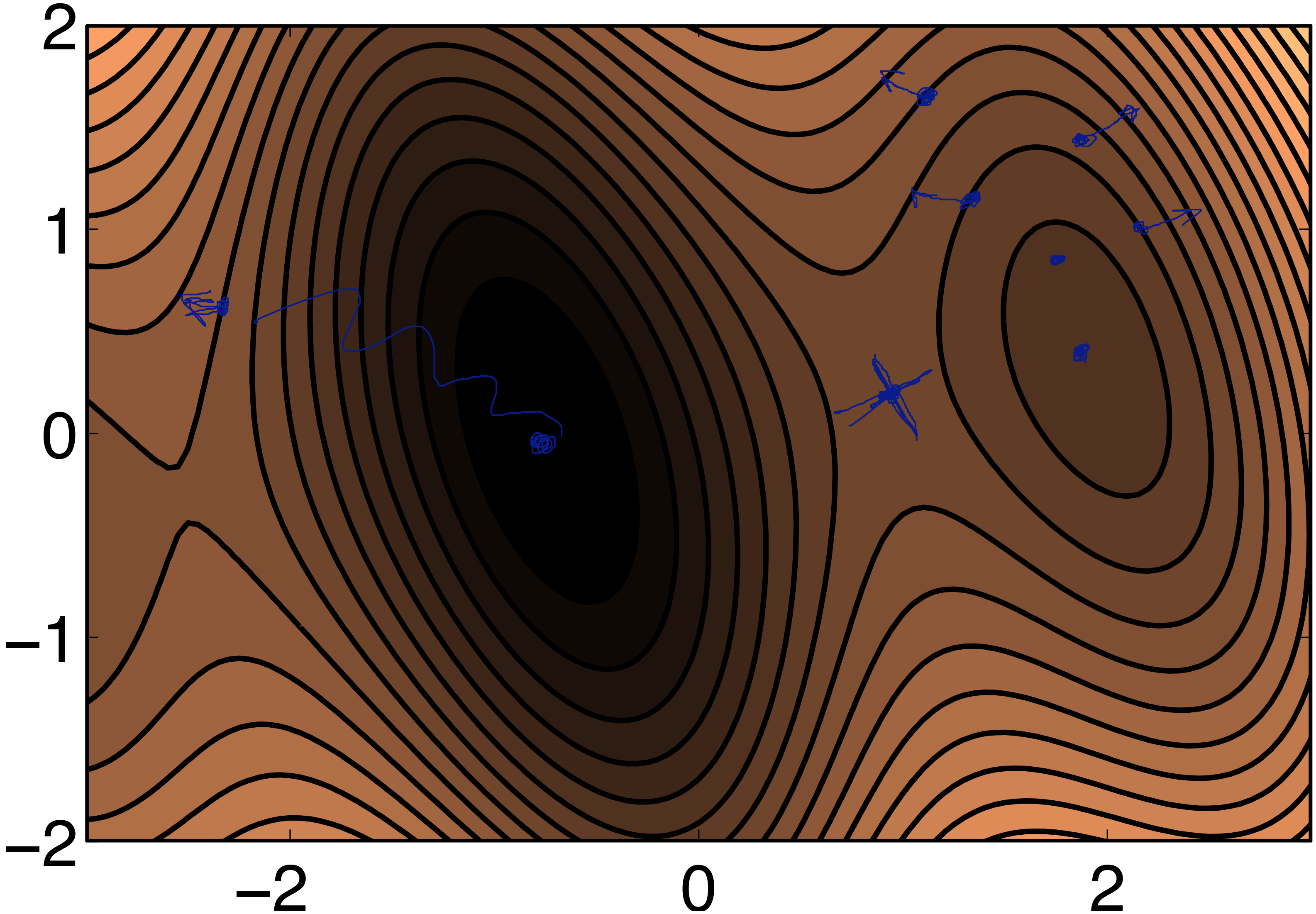
$\min_x \quad f(x)$

- for k = 1, 2, …
  - $g_k \leftarrow \nabla f(x_k)$ $= \frac{d}{dx} f \Big|_{x = x_k}$
  - $x_k \leftarrow x_{k-1} - t_k g_k$
- Choices: x_0, t_k, when to stop

Gradient descent: example



Gradient descent: example



In ML & stats

- Often have f(x) = $E_{p(i)}[f_i(x)]$
  - where i ~ $p(i)$
- E.g., linear regression: $E_{a,b}(a \cdot x - b)^2$

  $a \in \mathbb{R}^2 \quad b \in \mathbb{R}$

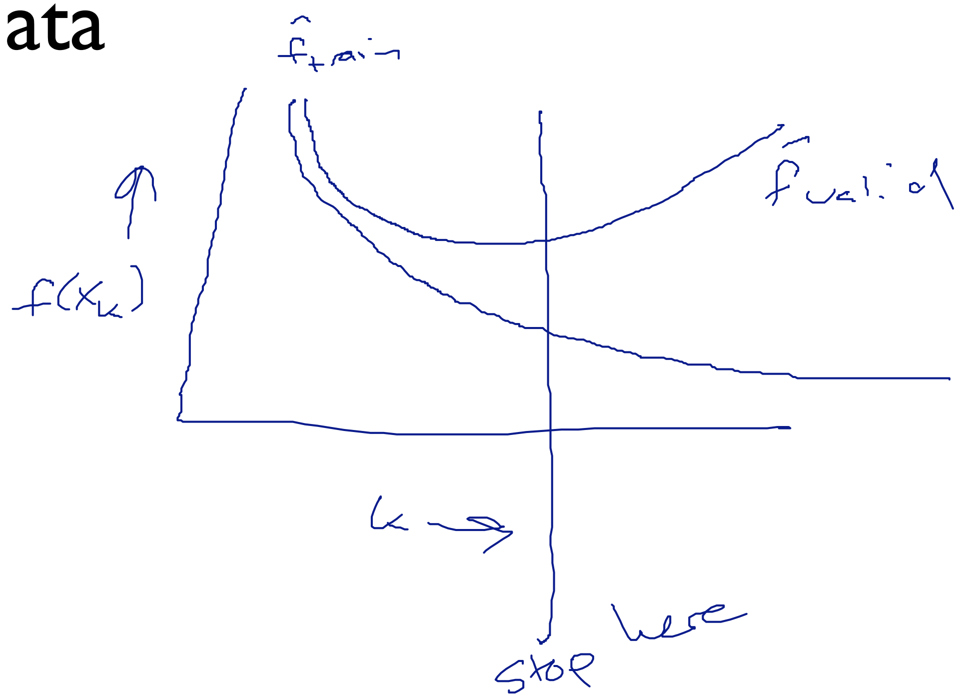
- Let: I = i.i.d. sample $\sim p(i)$
  - then $\hat{f}(x) = \sum_{i \in I} f_i(x) / |I|$

When do we stop?

- ML/stats: held out data

I_{train} $\quad$ I_{valid}

$\hat{f}_{train}$ $\quad$ $\hat{f}_{valid}$



- Early stopping
  - regularization
  - why bother?

When do we stop?

$\inf_x f(x)$

- Using convergence bounds (see below)
  - usual form is:

$$k \geq \underline{(f(x_0) - f^*)}\,(\text{fn of } 1/\epsilon)\,\underline{K_f}$$

  - need estimates of: