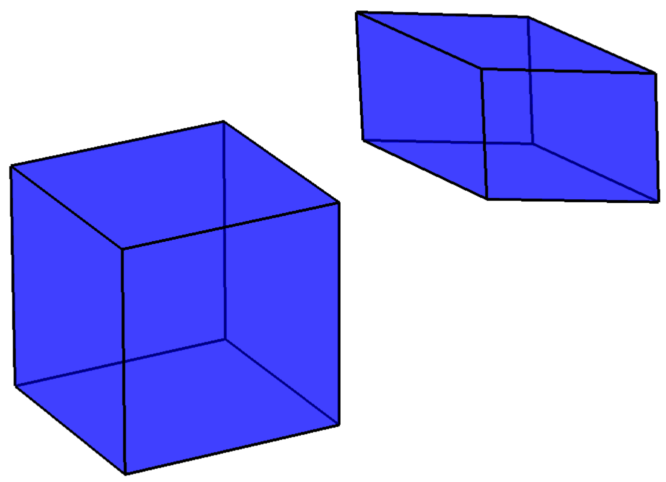
Newton's method

10-725 Optimization

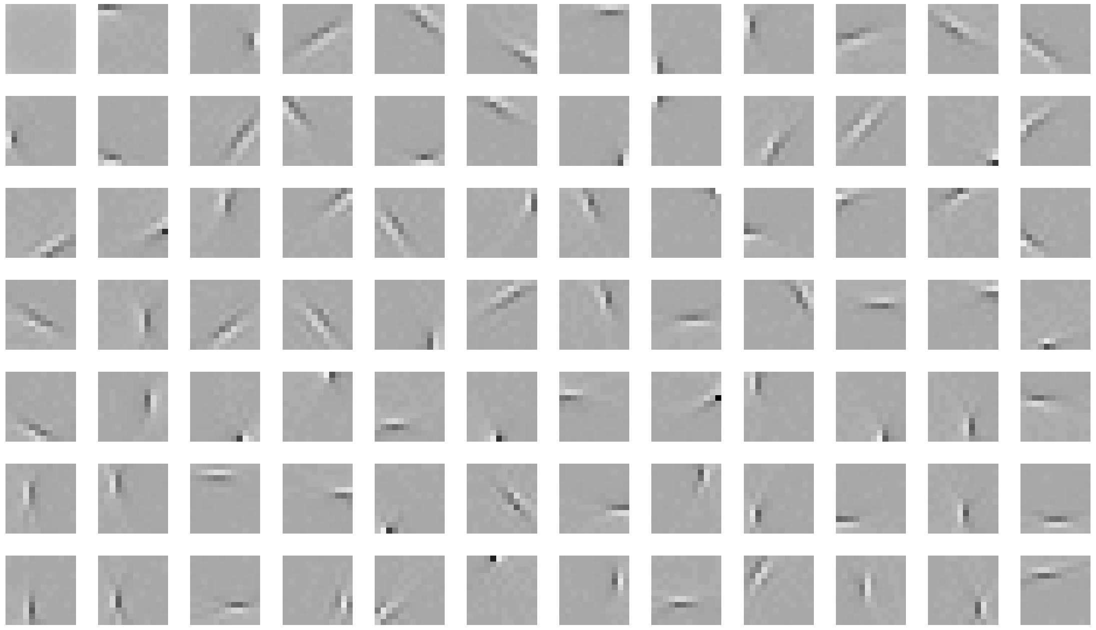
Geoff Gordon
Ryan Tibshirani

Review

- Volume rule



- Infomax ICA
  - matrix natural gradient

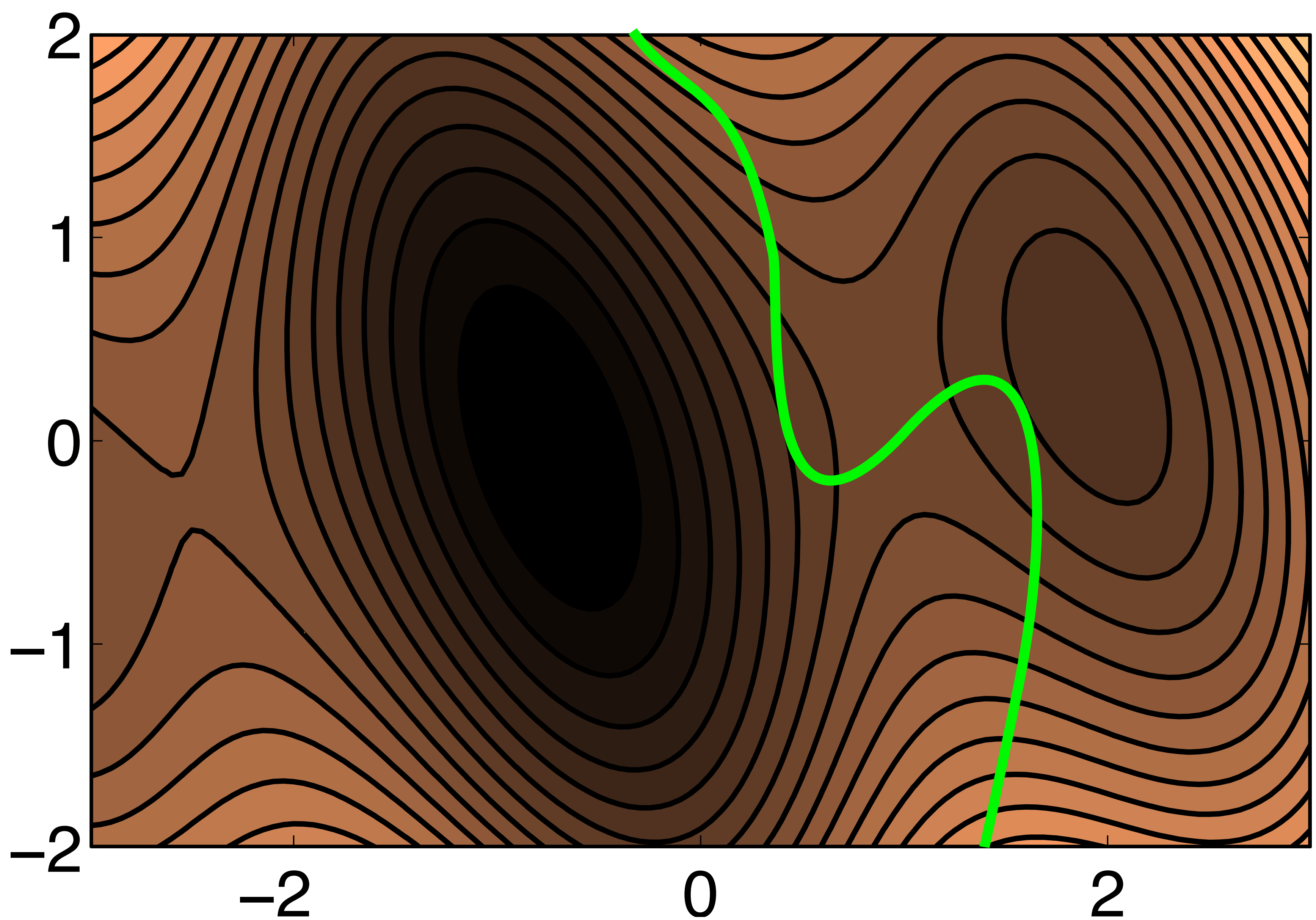


Review: Newton

- For solving nonlinear equations:
  - approx by linear ones, solve, update approx
  - $d = -J(x)^{-1}f(x)$
- For finding minima/maxima/saddles:
  - just use Newton on gradient $g(x) = f'(x) = 0$
  - $d = -H(x)^{-1}g(x)$
- Line search: Newton is a descent method
- (Often) quadratic convergence

Equality constraints

- min f(x) s.t. h(x) = 0



Optimality w/ equality

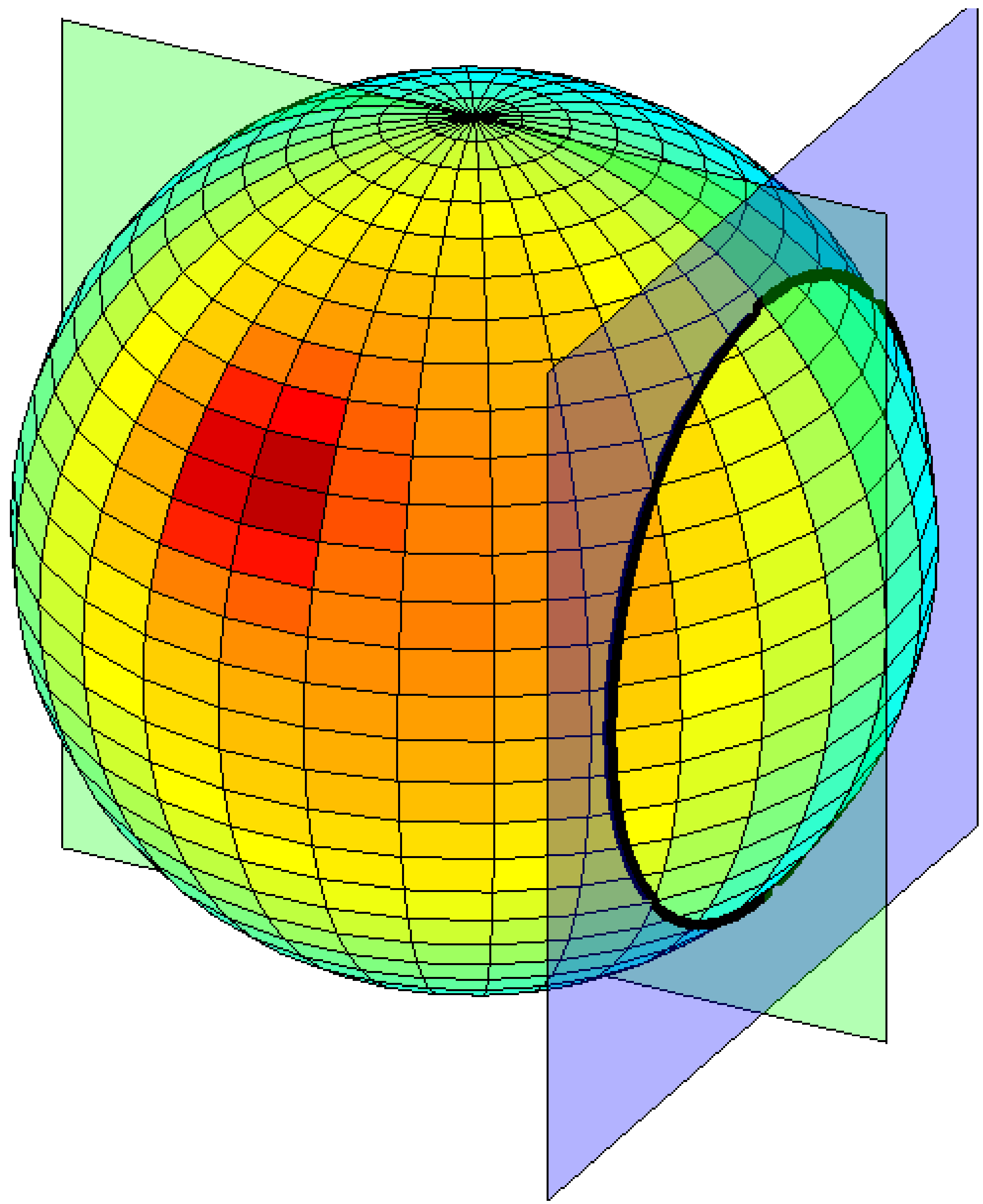
- min f(x) s.t. h(x) = 0
  - f: $R^d \to R$, h: $R^d \to R^k$ $(k \leq d)$
  - g: $R^d \to R^d$ (gradient of f)
- Useful special case: min f(x) s.t. Ax = 0

Picture

$$\max c^\top \begin{bmatrix} x \\ y \\ z \end{bmatrix} \text{ s.t.}$$

$$x^2 + y^2 + z^2 = 1$$

$$a^\top x = b$$



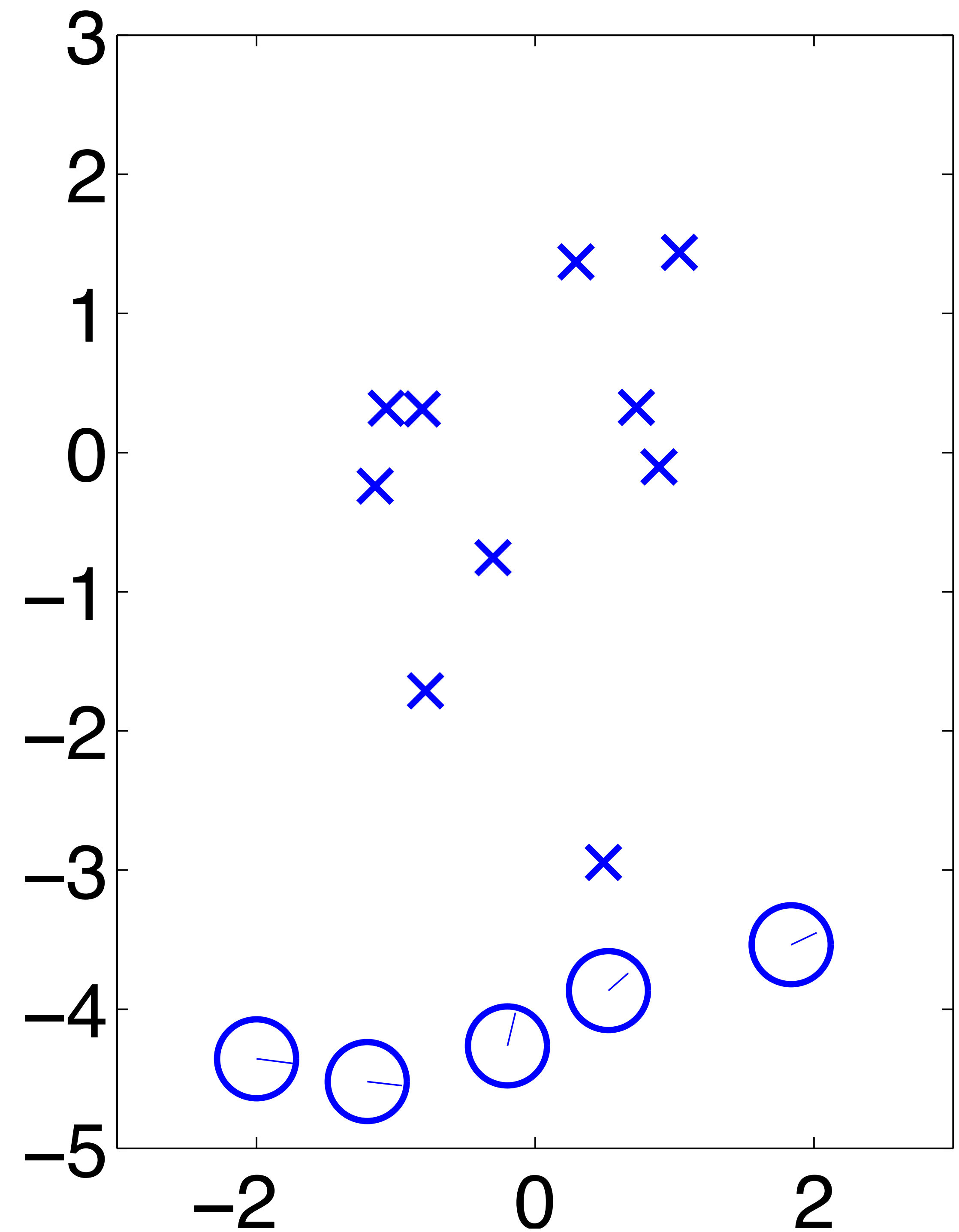
Optimality w/ equality

- min f(x) s.t. h(x) = 0
  - f: $R^d \rightarrow R$, h: $R^d \rightarrow R^k$ (k ≤ d)
  - g: $R^d \rightarrow R^d$ (gradient of f)
- Now suppose:
  - dg = dh =
- Optimality:

Example: bundle adjustment

- Latent:
  - Robot positions x_t, θ_t
  - Landmark positions y_k
- Observed: odometry, landmark vectors
  - $v_t = R_{\theta t}[x_{t+1} - x_t] + \textit{noise}$
  - $w_t = [\theta_{t+1} - \theta_t + \textit{noise}]_\pi$
  - $d_{kt} = R_{\theta t}[y_k - x_t] + \textit{noise}$

O = {*observed kt pairs*}



Bundle adjustment

$$\min_{x_t, u_t, y_k} \sum_t \|v_t - R(u_t)[x_{t+1} - x_t]\|^2 + \sum_t \|R_{w_t} u_t - u_{t+1}\|^2 + \sum_{(t,k)\in O} \|d_{k,t} - R(u_t)[y_k - x_t]\|^2$$

$$\text{s.t. } u_t^\top u_t = 1$$

- latent: Robot positions x_t, θ_t
  - $u_t = [\cos \theta_t; \sin \theta_t]$
- latent: Landmark positions y_k
- obs: $v_t = R_{\theta t}[x_{t+1} - x_t]$ + *noise*
- obs: $w_t = [\theta_{t+1} - \theta_t$ + *noise*$]_\pi$
- obs: $d_{kt} = R_{\theta t}[y_k - x_t]$ + *noise*

Ex: MLE in exponential family

$$L = -\ln \prod_k P(x_k \mid \theta)$$

$$P(x_k \mid \theta) =$$

$$g(\theta) =$$

MLE Newton interpretation

Convergence behavior

- $\min_x f(x)$ s.t. $Ax = b$
  - strictly convex $f(x)$, twice differentiable
  - some kind of bound on 3rd derivative
- Two phases
  - damped Newton—most of time here
    - step size < 1
  - quadratic convergence—a few final iterations to get accuracy very high
    - step size = 1

Convergence behavior

- Damped Newton
  - $f(x_{t+1}) \le f(x_t) - \Delta$ some fixed $\Delta > 0$
  - limit:
- Quadratic convergence
  - enter when $\text{error}_t \le \delta \le 0.5$
  - $\text{error}_{t+1} \le (\text{error}_t)^2$
  - limit:

Comparison

of methods for minimizing a convex function

	Newton	FISTA	(sub)grad	stoch. (sub)grad.
convergence				
cost/iter				
smoothness				

Variations

- Trust region
  - $[H(x) + tI]dx = -g(x)$
  - $[H(x) + tD]dx = -g(x)$
- Quasi-Newton
  - use only gradients, but build estimate of Hessian
  - in R^d, d gradient estimates at "nearby" points determine approx. Hessian (think finite differences)
  - can often get "good enough" estimate w/ fewer—even forget old info to save memory/time (L-BFGS)

Variations: Gauss-Newton

$$L = \min_{\theta} \sum_k \frac{1}{2} \|y_k - f(x_k, \theta)\|^2$$