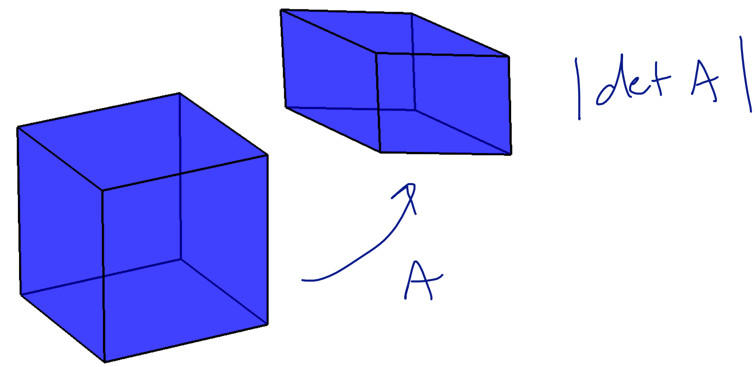
Newton's method

10-725 Optimization
Geoff Gordon
Ryan Tibshirani

Review

- Volume rule



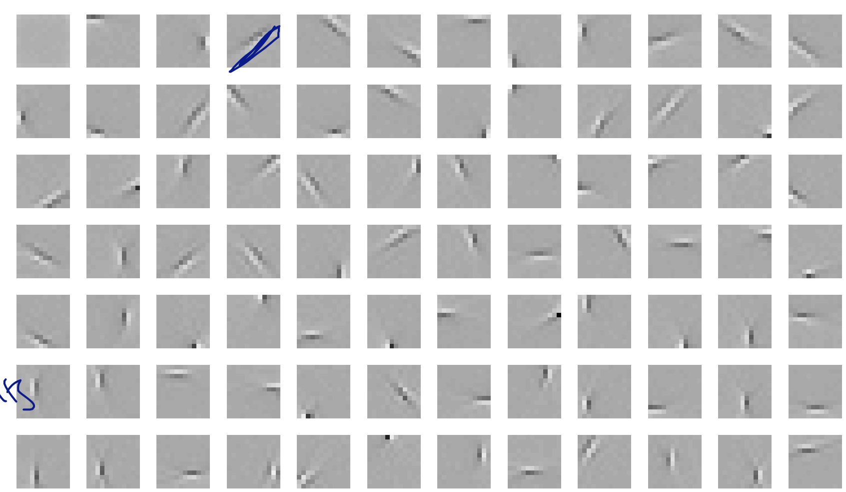
- Infomax ICA
  - matrix natural gradient

$y_i = g(W x_i)$

parameter

$x_i = W^{-1} g^{-1}(y_i)$

indep. components



Review: Newton

$\mathbb{R}^d \to \mathbb{R}^d$

$O(d^3)$

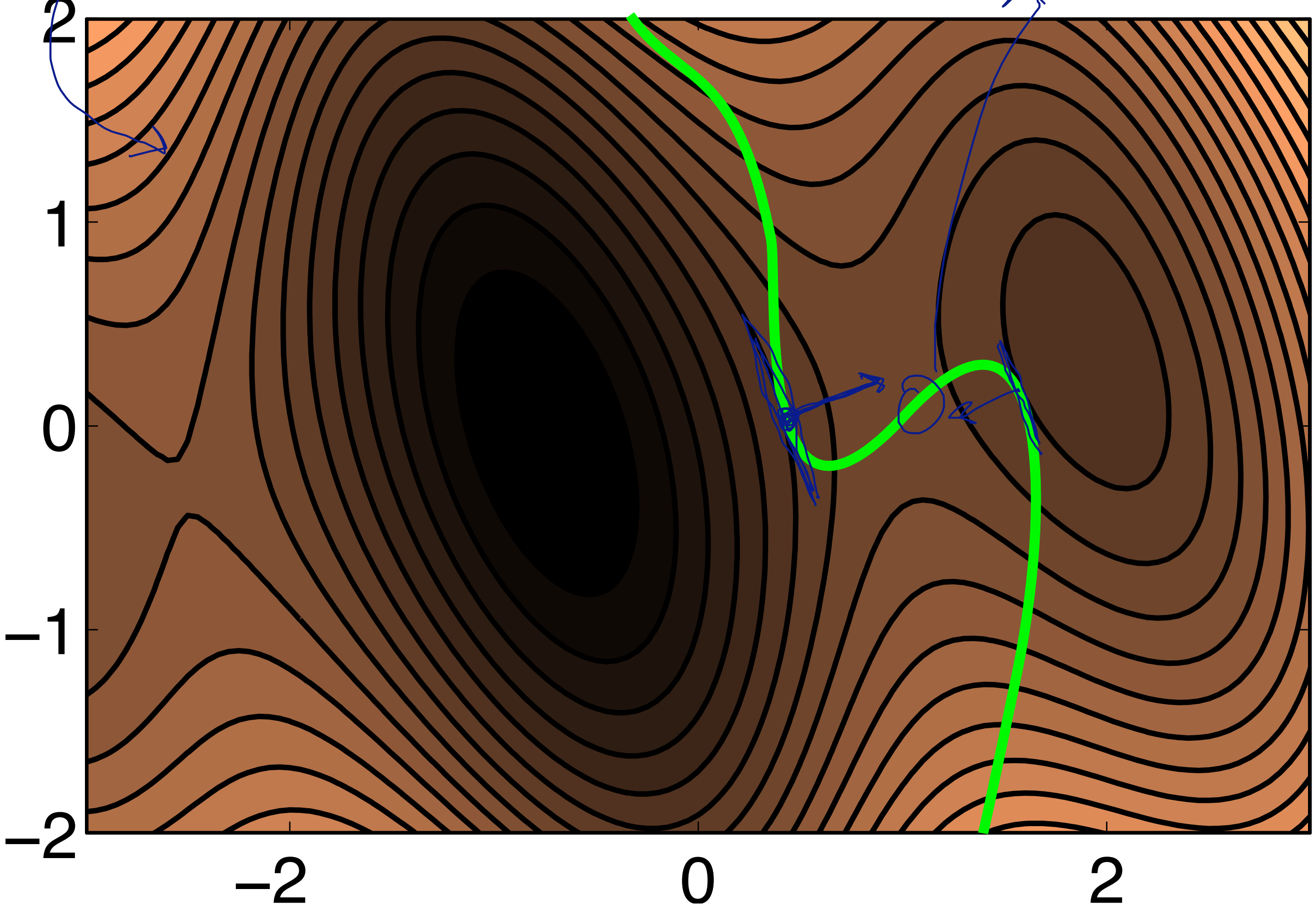
- For solving nonlinear equations: $f(x) = 0$
  - approx by linear ones, solve, update approx
  - $d = -J(x)^{-1}f(x)$
- For finding minima/maxima/saddles: $f : \mathbb{R}^d \to \mathbb{R}$ $g : \mathbb{R}^d \to \mathbb{R}^d$
  - just use Newton on gradient $g(x) = f'(x) = 0$
  - $d = -H(x)^{-1}g(x)$
- Line search: Newton is a descent method $\to f$ strictly convex
- (Often) quadratic convergence $\ln 1/\epsilon = O(k^2)$

$(\text{FISTA } 1/\epsilon = O(k^2))$

Equality constraints

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

local max



$g(x) = f'(x)$

$f'(x) = \lambda h'(x)$

grad.

constr normal

Optimality w/ equality

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

$C = \{x \mid Ax = 0\}$

$g(x) \perp C$

$g = f'$

$z \perp C \iff z^T x = 0 \quad \forall x \in C$

$z = A^T \lambda \qquad z^T x = \lambda^T (Ax) = 0$

$\{z \mid A^T \lambda = z\}$ $\exists \lambda$

$g(x) = A^T \lambda$

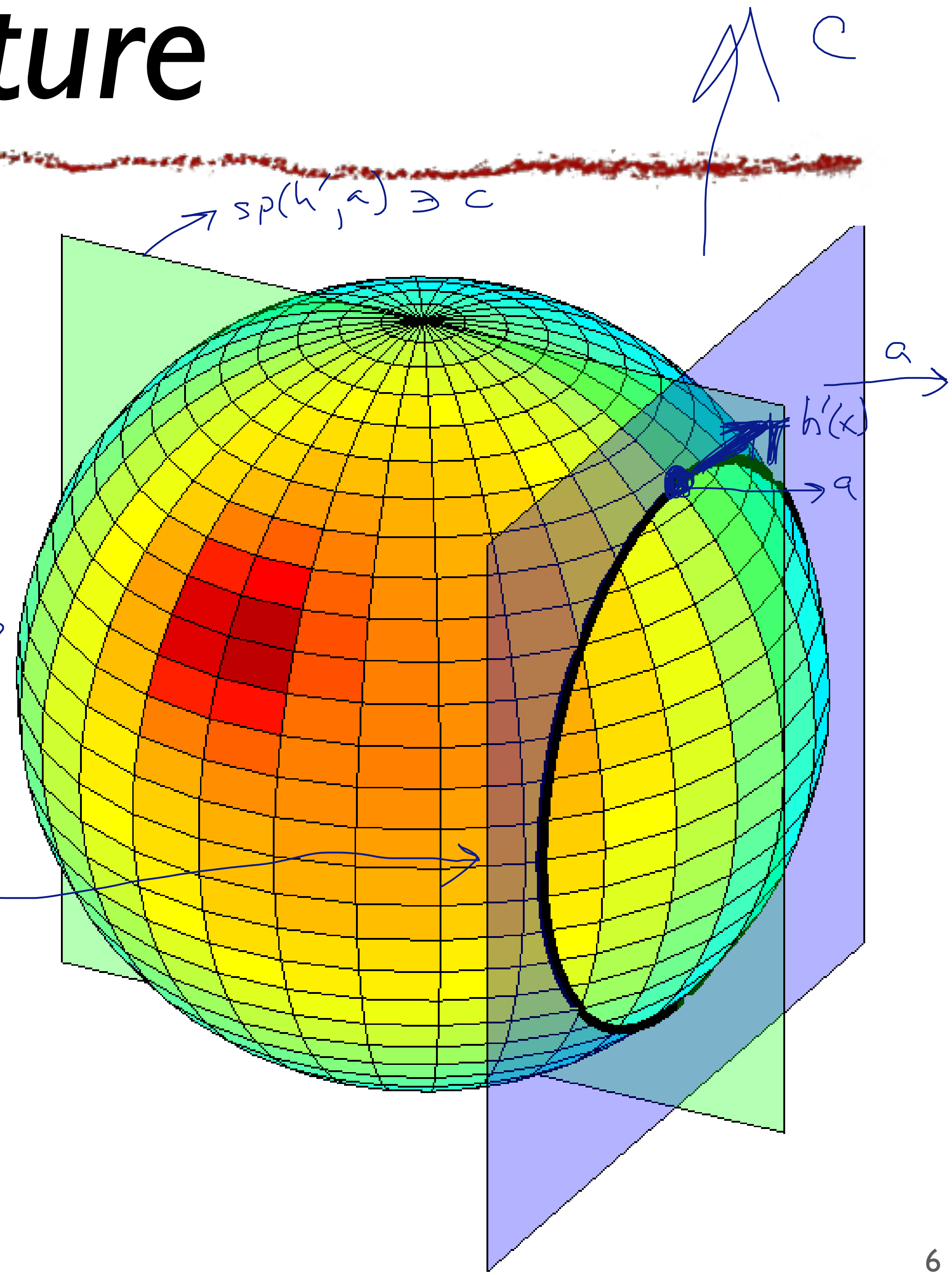
$g(x) = h'(x)^T \lambda$

Picture

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

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

$$a^\top x = b$$



Optimality w/ equality

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

$k{+}d \times k{+}d$

$$N = \begin{pmatrix} H & J^T \\ J & 0 \end{pmatrix}$$

- Now suppose:
  - dg = $\overset{d \times d}{H}(x)\,dx$ dh = $\overset{k \times d}{J}(x)\,dx$
- Optimality:

$$H(x)\,dx + J(x)^T \lambda = 0$$

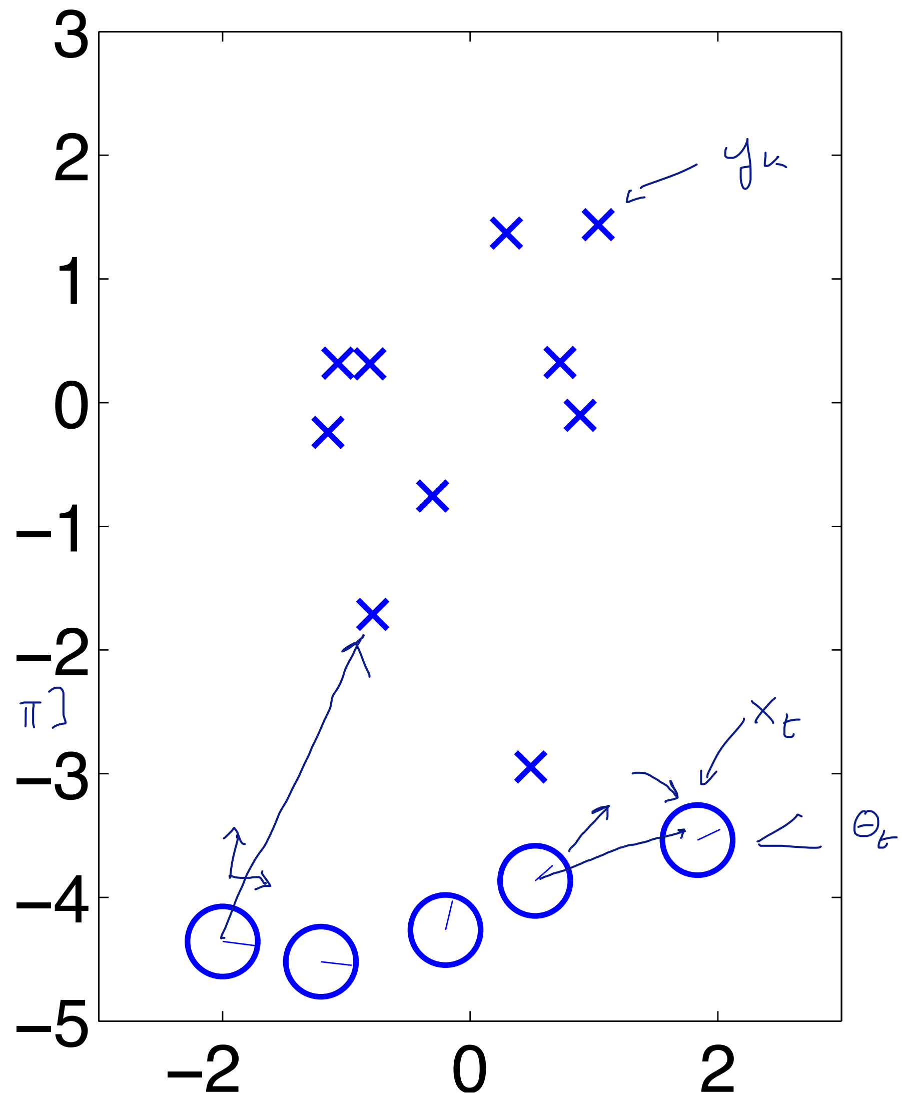
$$J(x)\,dx = -h(x)$$

$$N \begin{pmatrix} dx \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ -h \end{pmatrix}$$

Example: bundle adjustment

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

O = {observed kt pairs}



Bundle adjustment

$\begin{pmatrix} u_1 & u_2 \\ -u_2 & u_1 \end{pmatrix}$

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

$$\sum_{(t,k)\in O} \|d_{k,t} - R(u_t)[y_k - x_t]\|^2$$

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

H sparse

→ solve d quickly

- 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

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

$$dL = -\sum_k d(x_k \cdot \theta - g(\theta))$$

$$= -\sum_k (x_k \cdot d\theta - g'(\theta) \cdot d\theta) = \underbrace{-\sum_k s_k^T d\theta}$$

$$P(x_k \mid \theta) = \exp(x_k \cdot \theta - g(\theta))$$

$$g(\theta) = \ln \int_x \exp(x \cdot \theta)\, dx$$

$$s_k = x_k - g'(\theta)$$

$$d - \sum_k s_k = N g''(\theta)\, d\theta$$

MLE Newton interpretation

$$\int_{\mathcal{X}} \exp(\theta^T x - g(\theta))\, dx = 1$$

$$0 = \int_{\mathcal{X}} \nabla \exp(\theta^T x - g(\theta))\, dx$$

$$= \int_{\mathcal{X}} \exp(\theta^T x - g(\theta))[x - g'(\theta)]\, dx$$

$$g'(\theta) = \int_{\mathcal{X}} x\, P(x \mid \theta)\, dx = \mathbb{E}(x \mid \theta)$$

$$g''(\theta) = \int_{\mathcal{X}} \nabla [x \exp(\theta^T x - g(\theta))]\, dx$$

$$= \int_{\mathcal{X}} x \exp(\theta^T x - g(\theta))[x^T - g'(\theta)]\, dx$$

$$= \mathbb{E}(xx^T) - \mathbb{E}(x)\,\mathbb{E}(x)^T$$

$$= Var(x \mid \theta)$$

$$Var(x \mid \theta)^{-1} [\bar{x} - \mathbb{E}(x \mid \theta)]$$

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
  - $f(x_{t+1}) \leq f(x_t) - \Delta$ some fixed $\Delta > 0$
  - limit: $[f(x_0) - f(x_*)]/\Delta$ iters
- Quadratic convergence "error_t"
  - enter when $\text{error}_t \leq \delta \leq 0.5$
  - $\text{error}_{t+1} \leq (\text{error}_t)^2$
  - limit: 6 iters

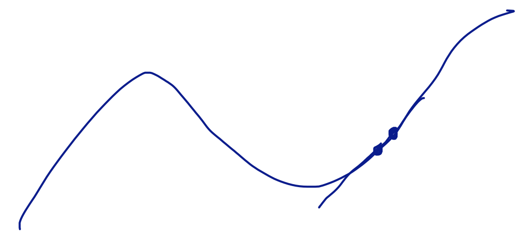
Comparison

of methods for minimizing a convex function

	Newton	FISTA	(sub)grad	stoch. (sub)grad.
convergence	*****	***	* / ** / ****½	*
cost/iter	$$$$$	$$$	$$	$
smoothness	+++	++	+ / ++ / +++	+

Variations

- Trust region $D = H \circ I$
  - $[H(x) + tI]dx = -g(x)$
  - $[H(x) + tD]dx = -g(x)$



- 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

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

$$L \approx \sum_k \frac{1}{2} \| y_k - [f(x_k, \theta) + J_k \, d\theta] \|^2$$

$$df = J_k \, d\theta$$

$$L \approx \sum_k \frac{1}{2} \| r_k - J_k \, d\theta \|^2$$

$$r_k = y_k - f(x_k, \theta)$$

$$g = - \sum J_k^T r_k \leftarrow$$

$$H = N J_k^T J_k \leftarrow$$

trust regions + G-N = Levenberg-Marquardt