First-order methods
Convexity

10-725 Optimization

Geoff Gordon
Ryan Tibshirani

Administrivia

- HW1 out, due 9/20
  - in class, at ***beginning*** of class—no skipping lecture to keep working on it ;-)
  - if you use late days, hand in to course assistant before 1:30PM on due date + k days
- Reminder: think about project teams and project proposals (due 9/25)

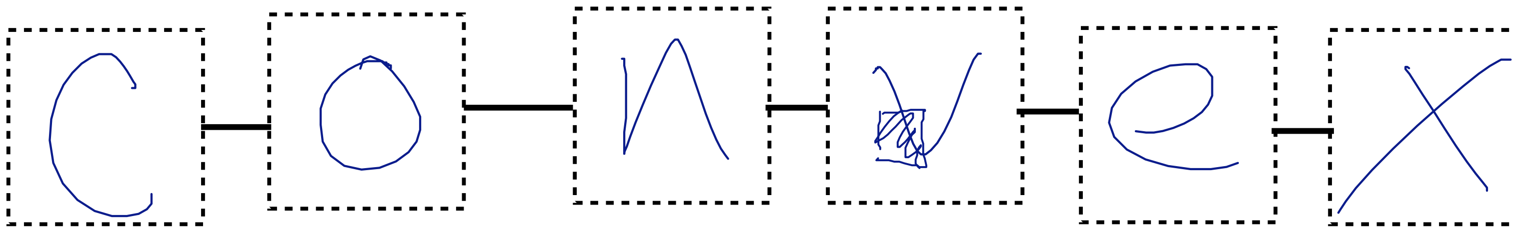
Review

- Convex sets
  - primal (convex hull) vs. dual (intersect hyperplanes)
  - supporting, separating hyperplanes
  - operations that preserve convexity
    - affine fn; perspective
  - open/closed/compact

Review

- Convex functions
  - epigraph
  - domain
  - sublevel sets; quasiconvexity
  - first order, 2nd order conditions
  - operations that preserve convexity
    - perspective; minimization over one argument

Ex: structured classifier



x_i pixels of char i

y_i $a \dots z$

$\varphi_j(x_i)$ feature of a char

$\psi_{ijk}(x_i, y_i)$ $\quad \phi_j(x_i)\,\delta(y_i = k)$ $\leftarrow w_{ijk}$

$\chi_{ikl}(y_i, y_{i+1})$ $\quad \delta(y_i = k)\,\delta(y_{i+1} = l)$ $\leftarrow v_{ikl}$

$L(x,y;v,w) = \sum_{ijk} \psi_{ijk}(x_i, y_i)\, w_{ijk} + \sum_{ikl} \chi_{ikl}(y_i, y_{i+1})\, v_{ikl}$

Classifier: $y = \arg\max_y L(x,y;v,w)$

Learning structured classifier

- Get it right if: $L(x, y; v, w) > L(x, y', v, w)$ $\forall y' \neq y$
- So, want: $L(x, y; v, w) \geq \max_{y'} \left(L(x, y'; v, w) + \pi(y, y') \right)$
- Where $\pi(y,y') = \{$ 0 $y = y'$, > 0 $y \neq y'$
- RHS: convex in v, w
- RHS – LHS: convex
- Train: lots of pairs (x^t,y^t) $\min_{v,w} \sum_t (RHS^t - LHS^t) + C(\|v\|^2 + \|w\|^2)$

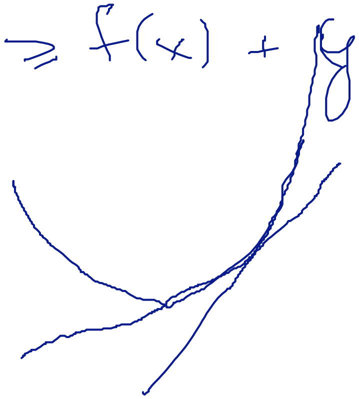
Strict convexity; strong convexity

$x, y \in \mathrm{dom}\, f$ $\quad t \in (0,1)$ $\quad x \neq y$

- Strictly convex:
  - $f(tx + (1-t)y) < t f(x) + (1-t) f(y)$

$\ln(1+e^x)$

- k-strongly convex:
  - $f(y) \geq f(x) + (y-x) f'(x) + k (y-x)^2/2$



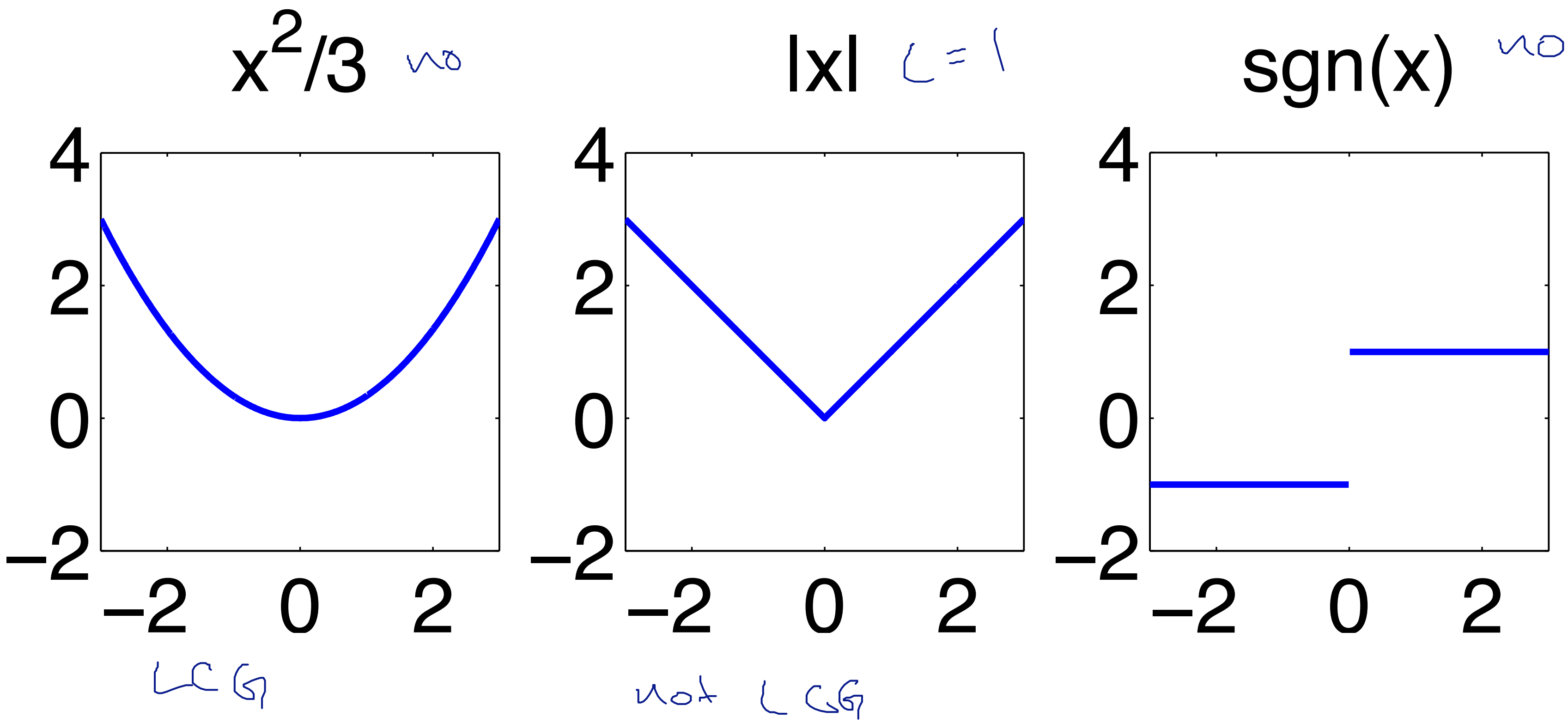
$|x| + x^2 \Rightarrow$ 2-strongly

Extended reals

- Suppose dom f $\subset R^n$
- Define $g(x) = \{$ $f(x)$ $x \in \mathrm{dom}\, f$
  ∞ o/w
- f convex $\Leftrightarrow$ g convex
  - $f(tx+(1-t)y) \le tf(x) + (1-t)f(y)$
  - $g(tx+(1-t)y) \le tg(x) + (1-t)g(y)$
  - cases: $x, y \in \mathrm{dom}\, f$
    $x, y \notin \mathrm{dom}\, f$
    $x \in \mathrm{dom}\, f \quad y \notin \mathrm{dom}\, f$

Lipschitz

- Function f(x) is Lipschitz (in norm ||x||) if:
  - $|f(x) - f(y)| \le L \cdot \|x - y\|$

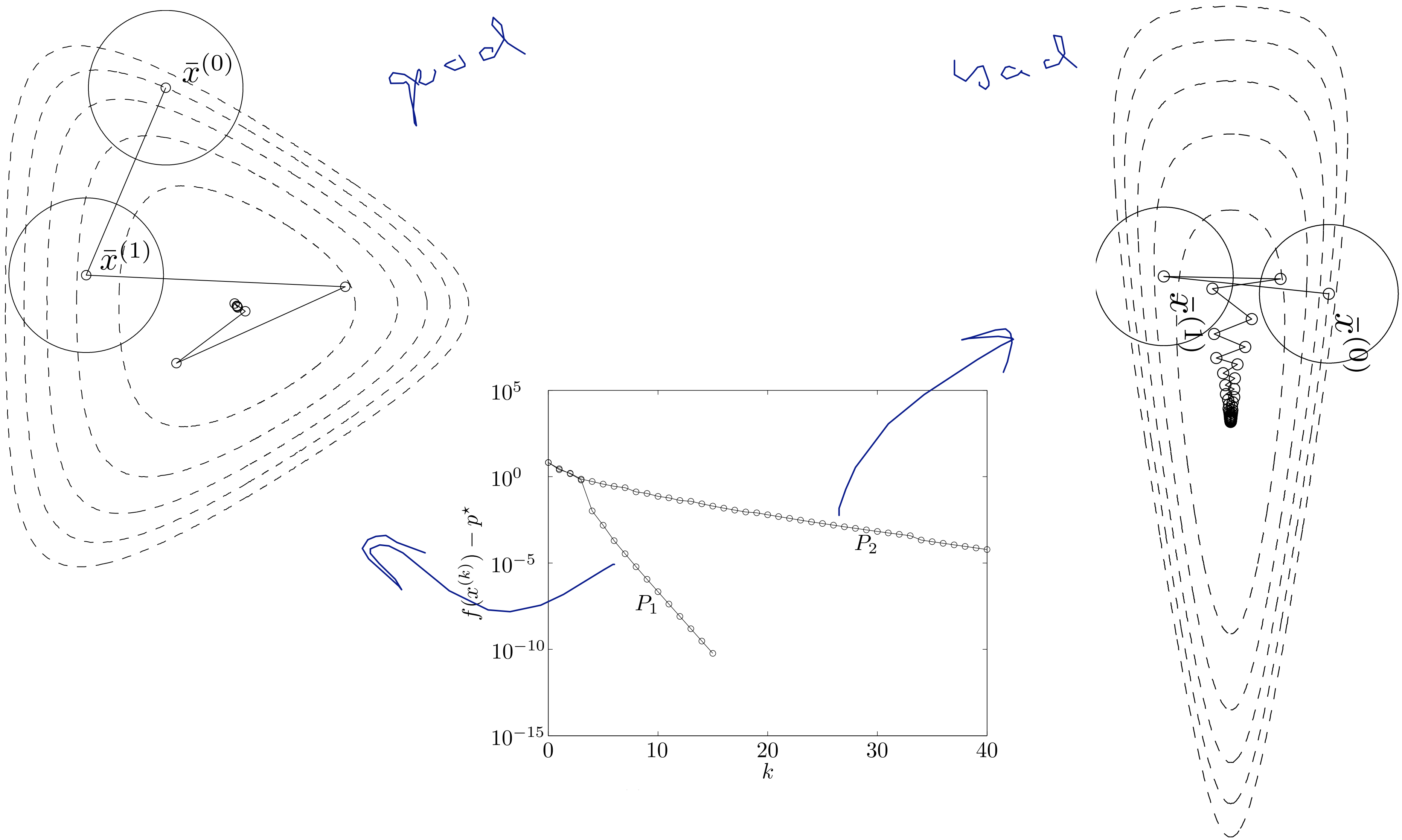


Back to gradient descent

- Suppose f(x) is convex, $\nabla f(x)$ exists
- Iterations to get to accuracy $\epsilon(f(x_0)-f(x^*))$: if
  - f Lipschitz: $O(1/\epsilon^2)$ — $t_k \approx 1/\sqrt{k}$
  - ∇f Lipschitz: $O(1/\epsilon)$
  - f strongly convex: $O(\ln(1/\epsilon))$

  t_k const
- Constant in O(…): ***conditioning*** of f

Conditioning



Extensions

- Subgradient descent
- Prox operator (e.g., $g_t = \arg\min_g \|g\|_p^2 + \nabla f \cdot g$)
- FISTA, mirror descent, conjugate gradient
- Nesterov's smoothing
- Line search (BV sec 9.2)
- Stochastic GD (when $f(x) = E(f_i(x) \mid i \sim P)$)
  - sample one i on each iter, use $\nabla f_i(x)$
  - or, minibatches: sample a few i's, use mean $\nabla f_i(x)$

Comparison: stochastic GD

- Iteration bounds for stochastic GD
  - f Lipschitz: $O(1/\epsilon^2)$ *(worse const, same O() as GD)*
  - f strongly convex: $O(\ln(1/\epsilon)/\epsilon)$ *(much worse)*
- f Lipschitz: stochastic GD often wins big
- Even if f strongly convex:
  - plain GD: each iter $O(N)$, #iters $O(\ln(1/\epsilon))$
  - stochastic: each iter $O(1)$, #iters $O(\ln(1/\epsilon)/\epsilon)$
  - stochastic can win if lots of data, loose tolerance
  - could make sense to throw data away, use full GD