First-order methods
Convexity

10-725 Optimization
Geoff Gordon
Ryan Tibshirani
• 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

\[
\begin{align*}
\mathbf{x}_i & \quad \text{pixels of char } \mathbf{i} \\
\mathbf{y}_i & \quad a \ldots z \\
\varphi_j(\mathbf{x}_i) & \quad \text{feature of a char} \\
\psi_{ijk}(\mathbf{x}_i, \mathbf{y}_i) & \quad \phi_j(\mathbf{x}_i) \mathbf{S}(\mathbf{y}_i = \mathbf{k}) \\
\chi_{ikl}(\mathbf{y}_i, \mathbf{y}_{i+1}) & \quad \mathbf{S}(\mathbf{y}_i = \mathbf{k}) \mathbf{H}(\mathbf{y}_{i+1} = \mathbf{l})
\end{align*}
\]

\[
L(x,y;v,w) = \sum \varphi_{ik}(\mathbf{x}_i, \mathbf{y}_i) + \sum \psi_{ikl}(\mathbf{y}_i, \mathbf{y}_{i+1}) + \sum \chi_{ikl}(\mathbf{y}_i, \mathbf{y}_{i+1})
\]

Classifier:
\[
y = \arg \max_y L(x, y; v, w)
\]
Learning structured classifier

- Get it right if: \( L(x, y; v, w) > L(x, y'; v, w) \)
- So, want: \( L(x, y; v, w) \geq \max_y (L(x, y'; v, w) + \text{proj}(y, y')) \)
- Where \( \pi(y, y') = \{
  \begin{array}{ll}
    0 & y = y' \\
    > 0 & y \neq y'
  \end{array}
\)
- 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

- Strictly convex:
  \[ f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \]

- k-strongly convex:
  \[ f(y) \geq f(x) + (y-x)^T f'(x) + \frac{k}{2} |y-x|^2 \]

\[ x, y \in \text{dom} f \quad x \neq y \quad k \in \mathbb{R}^+ \]
Extended reals

• Suppose dom $f \subset \mathbb{R}^n$

• Define $g(x) = \begin{cases} 
    f(x) & x \in \text{dom } f \\
    \infty & \text{otherwise} 
\end{cases}$

• $f$ convex $\iff$ $g$ convex
  \begin{itemize}
  \item $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$
  \item $g(tx + (1-t)y) \leq tg(x) + (1-t)g(y)$
  \item cases:
  \begin{align*}
    x, y & \in \text{dom } f \\
    x, y & \notin \text{dom } f \\
    x \in \text{dom } f & y \notin \text{dom } f \\
    x \notin \text{dom } f & y \in \text{dom } f \\
    x \notin \text{dom } f & y \notin \text{dom } f 
  \end{align*}
  \end{itemize}
Function $f(x)$ is Lipschitz (in norm $||x||$) if:

- $|f(x) - f(y)| \leq L \cdot ||x - y||$

Lipschitz functions:
- $x^2/3$ (not Lipschitz)
- $|x|$ (Lipschitz with $L = 1$)
- $\text{sgn}(x)$ (not Lipschitz)
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: $O(1/\epsilon^2)$  
  ‣ $\nabla f$ Lipschitz: $O(1/\epsilon)$
  ‣ $f$ strongly convex: $O(\ln(1/\epsilon))$

• Constant in $O(\ldots)$: \textit{conditioning} of $f$
Conditioning

Figure 9.14 The iterates of steepest descent with norm $\| \cdot \|_{P_1}$, after the change of coordinates. This change of coordinates reduces the condition number of the sublevel sets, and so speeds up convergence.

Figure 9.15 The iterates of steepest descent with norm $\| \cdot \|_{P_1}$, after the change of coordinates. This change of coordinates increases the condition number of the sublevel sets, and so slows down convergence.
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 \( f(x) = \mathbb{E}(f_i(x) \mid i \sim P) \))
  - 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