Lecture 6: ML & Opt

- Review of Taylor Series
- Oracle Models of Optimization
- Optimization as Minimization
- Convex Sets, Functions; Convex Optim Probs
- First & 2nd order characterizations of Convexity
- Examples
Taylor Series

Taylor Series is a local polynomial approximation about a point, \( x \), of an arbitrary smooth function.

\[
f(x) + f'(x) \cdot h \approx f(x+h)
\]

\[
f(y+h) \approx f(y) + f'(y) \cdot h + \frac{1}{2} f''(y) h^2
\]

Second-order TS expansion of \( f \) at \( y \)
Multivariate TS

First-order TS

\[ f(y) = f(x) + \langle \nabla f(x), y - x \rangle \]

\[ \nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} \]

\[ f(y) = f(x) + \frac{\partial f(x)}{\partial x_1} \cdot (y_1 - x_1) + \frac{\partial f(x)}{\partial x_2} \cdot (y_2 - x_2) \]
Define the Hessian matrix of 2nd derivatives of \( f \) at \( x \):

\[
H = \nabla^2 f(x) = \begin{bmatrix}
\frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\
\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2}
\end{bmatrix}
\]

It is symmetric.

2nd-order Taylor Series:

\[
f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T H (y - x)
\]

\[
f(y) \approx f(x) + \frac{\partial f(x)}{\partial x_1} (y_1 - x_1) + \frac{\partial f(x)}{\partial x_2} (y_2 - x_2)
\]

\[
+ \frac{1}{2} \left[ \frac{\partial^2 f(x)}{\partial x_1^2} (y_1 - x_1)^2 + 2 \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} (y_1 - x_1)(y_2 - x_2) + \frac{\partial^2 f(x)}{\partial x_2^2} (y_2 - x_2)^2 \right]
\]
In general, \( f: \mathbb{R}^d \rightarrow \mathbb{R} \)

\[
f(y) \approx f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle y - x, H(y - x) \rangle
\]

where \( H \in \mathbb{R}^{d \times d} \) and \( H_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \)

(symmetric)
Oracle Models of Optimization

To solve a general optim prob numerically, we use
iterative methods

\[ x_0 \leftarrow \text{initial guess of the solution} \]

- compute \( f(x_0) \) and/or \( \nabla f(x_0) \) and/or \( \nabla^2 f(x_0) \)
- form a local model
- move to a new point based on the local model
- repeat until "convergence"
Convex Optim

- We've seen we reduce ML to optim probs (ERM and MLE or MAP)

We also know that we only want our numerical soln $f_T$ to satisfy

$$R_n(f_T) \leq R_n(\hat{f}_T) + O(\varepsilon)$$

so we only care about solving optim probs to $\varepsilon$-suboptimality
In general, optim is NP-Hard so we restrict ourselves to convex optim problems (for now).

\[ x^* = \arg\min_{x \in C} f(x) \]

- **Convex**: Local minima are equal to global minima.
- **Nonconvex**: Global minima are not unique. Local minima are not always global minima.
- Why restrict ourselves (for now) to convex optim probs
  - very expressive model class
  - they are generally tractable
  - local minima are global minima, so iterative algos are guaranteed to find global minima

- Many ML probs are naturally convex:
  - GLMs fit using MLE give convex optim probs
  - many efficient alg exist, tailored for different settings
Convexity

We say a set $C$ is convex if whenever $x, y \in C$, the line segment $[x, y] = \{ \alpha \in [0, 1] : \alpha x + (1-\alpha)y \}$ is contained in $C$.

Fact: A set $C$ is convex iff for all $x_1, \ldots, x_k \in C$ and any numbers $\Theta_1, \ldots, \Theta_k$ s.t. $\Theta_k \geq 0$ for all $i$ and $\sum \Theta_i = 1$, the convex combination $\Theta_1 x_1 + \cdots + \Theta_k x_k \in C$. 
Given $x_1, \ldots, x_k$ the set of all convex combinations of these pts is called their **convex hull**

**Fact** If $C$ is a convex set and $X$ is a r.v. taking values in $C$, then $E(X) \in C$
Ex of convex sets

- empty set
- $\mathbb{R}, \mathbb{R}_+, \mathbb{R}^d$, rays, lines, line segments
- hyperplanes: $\{ x \mid x^T a = b \}$

- positive orthant
  $x \geq 0$

- convex hull of an arb. set $S$, i.e. the smallest convex set containing $S$

NNLS: $\min \|AX - b\|_2^2$
$x \geq 0$
- The Euclidean unit ball \( \mathbb{S}^n \) \( \| x \|_2 = 1 \)

- Norm balls \( B_{\| \cdot \|_p} (x, r) = \{ y \mid \| x - y \|_p \leq r \} \)

- \( B_{\| \cdot \|_1} (0, 1) \)

- \( B_{\| \cdot \|_\infty} (0, 1) \)

- \( \| x \|_p = \max_{i} \| x_i \| \) and \( \lim_{p \to \infty} \| x \|_p \)
- **Polyhedrons**: set of solutions to a finite number of linear equalities & inequalities

\[ \{ x : a_i^T x \leq b_i \} = \{ x : Ax \leq b \} \]

\[ A = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \]

- Intersection of arbitrary \# of convex sets is convex

\[ \bigcap_{i \in I} C_i \text{ is convex if } C_i \text{ is convex for all } i \in I \]
Convex function

A function \( f \) defined on a convex set \( C \) is called convex if it satisfies

\[
f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)
\]

for all \( \alpha \in [0,1] \) and \( x, y \in C \)

\( f(x) = \frac{1}{x} \) is convex on \( \mathbb{R}_{++} \)

Graphically: a function is convex if it lies under its chords
**Strict convexity**

\[ f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y) \] is satisfied for all \( x,y \in \text{dom}(f) \) and \( \alpha \in (0,1) \)

---

\[ f = 0 \]

not strictly convex

\[ x = 1 \quad (1-x)^+ \]

---

Most important consequence of convexity

If a function \( f \) is convex on a convex set \( C \), then any local minimum of \( f \) on \( C \) is a global minimum.
Fact

If $f$ is strictly convex on $C$ then the minimizer of $f$ on $C$ (if any exists) is unique.