Convex Optimization

Instructor: Pranjal Awasthi
So Far

• Learning Algorithms such as
  – SVMs
  – Regression
  – Ridge
  – Maximum likelihood estimation in graphical models

• All rely on solving convex problems

• Convex Optimization
  – Concerns the study of algorithm for optimizing convex functions
  – The engine underneath most successful ML algorithms
Convexity

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if
  - $\forall x, y \in \mathbb{R}^n, 0 < t < 1,$
  - $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$

- The function value at the average is less than the average of the function values.
Convexity

• A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if and only if
  – $\forall x, y \in \mathbb{R}^n$, $0 < t < 1$,
  – $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$

• If $f$ is twice differentiable then convexity equivalent to
  – $H(x) \geq 0$, $\forall x$
  – $H(x)$ is an $n \times n$ matrix, $H_{i,j}(x) = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$
  – Equivalent to $\lambda_{min}(H(x)) \geq 0$, $\forall x$
Basic Convex Functions

- Affine: $ax + b$
- Exponential: $e^{ax}$
- Power: $x^p$, for $p \geq 1$ or $p \leq 0$
- Norms: $\|x\|_2, \|x\|_1, ...$

- Convexity preserving operations
  - Positive scaling: $cf(x), c > 0$
  - Sum: $f_1(x) + f_2(x)$
  - Pointwise Maximum: $\max (f_1(x), f_2(x))$
Examples

\[ \min \frac{1}{2} \| \mathbf{w} \|^2 \]

Subject to: \( y_i (\mathbf{w} \cdot \mathbf{x}_i) \geq 1, \forall i \)

\[ L = \frac{1}{2} \| \mathbf{w} \|^2 \]

\[ \frac{\partial L}{\partial w_j} = w_j \]

\[ \frac{\partial^2 L}{\partial w_j^2} = 1, \quad \frac{\partial^2 L}{\partial w_j \partial w_k} = 0 \]

\[ H = \begin{bmatrix} 1 & \cdots & 0 \\ 0 & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = I \]

always positive semidefinite
Examples

\[ \hat{w} = \arg\min_w \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{X}_i \cdot w)^2 \]

\[ \frac{\partial L}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} 2(y_i - \hat{X}_i \cdot w)(-X_{ij}) \]

\[ \frac{\partial^2 L}{\partial w_j \partial w_k} = \frac{1}{m} \sum_{i=1}^{m} 2 X_{ij} X_{ik} \]

\[ H = \frac{1}{m} X^T X \quad \text{(covariance matrix)} \]

always positive semidefinite
Examples

\[ \hat{w} = \arg\min_w \frac{1}{m} \sum_{i=1}^{m} \left( y_i - \hat{X}_i \cdot w \right)^2 + ||w||^2 \]

Sum of two convex functions
Examples

\[ \hat{w} = \arg\min_w \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{X}_i \cdot w)^2 + \|w\|_1 \]
Examples

\[-\frac{1}{m} \sum_{s=1}^{m} \sum_{i=1}^{s} \theta_i X_{si} + \sum_{i \sim j} \theta_{i,j} X_{si} X_{sj} + A(\theta)\]

\[\frac{\partial^2 A(\theta)}{\partial \theta_i \partial \theta_j} = \text{cov}(X_i, X_j)\]

Hessian = Covariance matrix

Always positive semidefinite
Checking for Convexity

1. Use definition
   - $\forall x, y \in \mathbb{R}^n, \ 0 < t < 1,$
   - $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$

2. If $f$ is twice differentiable then check Hessian
   - Equivalent to $\lambda_{min}(H(x)) \geq 0, \forall x$ — very useful

3. Show that $f$ is obtained from a convex functions using convexity preserving operations

4. In general, checking convexity is NP-hard
Advantage of Convexity

• In general, checking convexity is NP-hard
• But minimizing a convex function is easy!
• General method
  – Ellipsoid Algorithm
  – Can output a point $x$ such that $f(x) \leq f(x^*) + \epsilon$
    • $f(x^*) = \min_x f(x)$
    • Runtime $\sim O(n^4 \log(1/\epsilon))$

• Can do better if functions are nicer
Practical Algorithms for Convex Minimization

- First Order methods
  - Gradient Descent
  - Subgradient Descent
  - Stochastic Gradient Descent
  - Accelerated Gradient Descent

- Second Order Methods
  - Newton’s Method
  - Gauss-Newton
  - BFGS
Gradient Descent

• Assume $f$ is continuous and differentiable
• What makes minimizing $f$ easy?
  – Any local minimum is a global minimum

  Suppose $x_0$ is a local min
  
  $\Rightarrow \forall z$ such that $\|z - x_0\| \leq R$, $f(z) \geq f(x_0)$

  Suppose $y \neq x_0$ is the global min
  
  $\Rightarrow$
  ① $f(y) < f(x_0)$
  ② $\|y - x_0\| > R$

  Let $z = \theta y + (1 - \theta) x_0$ for $\theta = \frac{R}{2\|y - x_0\|}$

  Then
  ③ $\|z - x_0\| = \theta \|y - x_0\| = \frac{R}{2}$
  ④ $f(z) \leq \theta f(y) + (1 - \theta) f(x_0) < f(x_0)$

  ③ + ④ contradicts local optimality of $x_0$. 
Gradient Descent

- Assume $f$ is continuous and differentiable
- What makes minimizing $f$ easy?
  - Local information implies global information
  - $f(y) \geq f(x) + \Delta f(x)^T(y - x), \forall x, y$

Proof in 1-dimensions:

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x)$$

$$= t(f(y) - f(x)) + f(x)$$

$$\Rightarrow tf(y) \geq f(x) + \left[\frac{f(ty + (1-t)x) - f(x)}{t(y-x)}\right] (y - x)$$

$$\Rightarrow f(y) \geq f(x) + \left[\lim_{t \to 0} \frac{f(ty + (1-t)x) - f(x)}{t(y-x)}\right] (y - x)$$

$$= f(x) + \nabla f(x)(y - x) \quad \Box$$
Gradent Descent

- Assume $f$ is continuous and differentiable
- What makes minimizing $f$ easy?
  - Local information implies global information
  - $f(y) \geq f(x) + \Delta f(x)^T(y - x), \forall x, y$
Gradient Descent

• Assume $f$ is continuous and differentiable
• What makes minimizing $f$ easy?
  – Local information implies global information
  – $f(y) \geq f(x) + \Delta f(x)^T(y - x), \forall x, y$
  – Given a point $x_t$ and $\Delta f(x_t)$,
    • Adding gradient can only hurt
    • Must go in the other direction

  $x_{t+1} = x_t - \eta \Delta f(x_t)$
  $\eta = \text{step size}$
Gradient Descent

- Assume $f$ is continuous and differentiable
- Start at arbitrary $x_0$. If $\Delta f(x_0) = 0$, return $x_0$
- For $t=1,2,\ldots$
  - $x_{t+1} = x_t - \eta \Delta f(x_t)$

- What guarantee do we have on the optimal solution?
- When to stop?
Gradient Descent

• Assume $f$ is continuous and differentiable
• Start at arbitrary $x_0$. If $\Delta f(x_0) = 0$, return $x_0$
• For $t=1,2,\ldots$:
  \[ x_{t+1} = x_t - \eta \Delta f(x_t) \]

• Assumption: $f$ is $L$-Lipschitz
  – Gradients at nearby points cannot change drastically
  – A notion of smoothness of $f$
  – $\|\Delta f(x) - \Delta f(y)\|_2 \leq L \|x - y\|_2$
**Gradient Descent**

**Theorem:** Let $x^*$ be the optimal solution and $x_0$ be the initial point. Then in $t$ steps, we have

$$f(x_t) - f(x^*) \leq \frac{||x_0 - x^*||^2}{2\eta t},$$

provided $\eta \leq \frac{1}{2L}$

- $\eta$ too large: Won’t converge
- $\eta$ too small: Convergence is slow
- In practice: $\eta$ is chosen to decay with $t$
- #iterations independent of $n$, the dimensionality!!
Gradient Descent

**Theorem:** Let $x^*$ be the optimal solution and $x_0$ be the initial point. Then in $t$ steps, we have

$$f(x_t) - f(x^*) \leq \frac{||x_0-x^*||^2}{2\eta t}, \text{ provided } \eta \leq \frac{1}{2L}$$

- When to stop
  - When gradient becomes small
  - Value of $f()$ does not change much