Gradient Descent

Instructor: Pranjal Awasthi
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 \( \lambda_{min}(H(x)) \geq 0, \forall x \)
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
Gradient Descent

• Assume $f$ is continuous and differentiable
• Start at arbitrary $x_0$.
• For $t=1,2,\ldots$:
  
  $x_{t+1} = x_t - \eta \nabla f(x_t)$
Gradient Descent

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

• 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

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

*Analysis:*

1. $f(x_t)$ decreases monotonically
2. Analyze the rate of convergence.
Gradient Descent

1. \( f(y) \geq f(x) + \nabla f(x) \cdot (y-x) + \frac{1}{2} \| y-x \|^2 \)

2. \( \| \nabla f(x) - \nabla f(y) \| \leq L \| x-y \| \)

Want to show, \( f(x_{t+1}) \leq f(x_t) \)

Use 1 with \( y=x_t, x=x_{t+1} \)

\( = f(x_t) \geq f(x_{t+1}) + (\nabla f(x_{t+1})) \cdot (x_t - x_{t+1}) \)

\( = f(x_{t+1}) + \nabla f(x_{t+1}) \cdot (\nabla f(x_t)) \)

\( = f(x_{t+1}) \leq f(x_t) - h \nabla f(x_t) \cdot \nabla f(x_{t+1}) \)
Gradient Descent

\[
\begin{align*}
  f(x_{t+1}) & \leq f(x_t) - h \nabla f(x_t) \cdot \nabla f(x_t) \\
  & = f(x_t) - h \nabla f(x_t) [\nabla f(x_t) + \nabla f(x_{t+1}) - \nabla f(x_t)] \\
  & = f(x_t) - h \| \nabla f(x_t) \|^2 + h \nabla f(x_t) \cdot (\nabla f(x_t) - \nabla f(x_{t+1})) \\
  \min_x h \nabla f(x_t) \cdot (\nabla f(x_t) - \nabla f(x_{t+1})) & \leq h \| \nabla f(x_t) \| \| \nabla f(x_t) - \nabla f(x_{t+1}) \| \\
  & \leq (h \| \nabla f(x_t) \|)^2 \| x_t - x_{t+1} \| \\
  & \leq h^2 L \| \nabla f(x_t) \|^2 \\
  \Rightarrow \quad f(x_{t+1}) & \leq f(x_t) - h \| \nabla f(x_t) \|^2 + h^2 L \| \nabla f(x_t) \|^2 \\
  & = f(x_t) - h \| \nabla f(x_t) \|^2 (1 - h L) \\
  \text{If } h < \frac{1}{2L}, \quad \frac{f(x_{t+1})}{2} & \leq f(x_t) - h \| \nabla f(x_t) \|^2
\end{align*}
\]
Gradient Descent

Step 2:

\[ f(x_{t+1}) \leq f(x_t) - \frac{n}{2} \| \nabla f(x_t) \|^2 \]

Want to compare \( f(x_{t+1}) \) with \( f(x^*) \)

1. \( f(y) \geq f(x) + \nabla f(x) \cdot (y - x) \)

Set \( y = x^* \), \( x = x_t \)

\[ \Rightarrow f(x^*) \geq f(x_t) + \nabla f(x_t) \cdot (x^* - x_t) \]

\[ \Rightarrow f(x_t) \leq f(x^*) - \nabla f(x_t) \cdot (x^* - x_t) \]

\[ f(x_{t+1}) \leq f(x^*) - \nabla f(x_t) \cdot (x^* - x_t) - \frac{n}{2} \| \nabla f(x_t) \|^2 \]

\[ = f(x_t) - \frac{1}{n} (x_t - x_{t+1}) \cdot (x^* - x_t) - \frac{1}{2n} \| x_t - x_{t+1} \|^2 \]
Gradient Descent

\[
f(x_{t+1}) \leq f(x^*) - \frac{1}{n} (x_t - x_{t+1})(x^* - x_t) \leq \frac{1}{2n} ||x_t - x_{t+1}||^2
\]

\[
= f(x^*) + \frac{1}{2n} \left[ 2(x_{t+1} - x_t)(x_t - x^*) - ||x_t - x_{t+1}||^2 \right]
\]

\[
= f(x^*) + \frac{1}{2n} \left[ 2x_{t+1} \cdot x^* - 2x_{t+1} \cdot x_t - 2x_t \cdot x^* + 2x_t \cdot x_e - x_t \cdot x_t - x_{t+1} \cdot x_{t+1} + 2x_e \cdot x_{t+1} \right]
\]

\[
= f(x^*) + \frac{1}{2n} \left[ (x_t \cdot x_t - 2x_t \cdot x^*) - (x_{t+1} \cdot x_{t+1} - 2x_{t+1} \cdot x^*) \right]
\]

\[
f(x_{t+1}) \leq f(x^*) + \frac{1}{n} \left[ 2 ||x_t - x^*||^2 - 2 ||x_{t+1} - x^*||^2 \right]
\]

\[
\sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*) \leq \sum_{t=0}^{T-1} \frac{1}{2n} \left[ 2 ||x_t - x^*||^2 - 2 ||x_{t+1} - x^*||^2 \right]
\]
Gradient Descent

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

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

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

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

- What if $f$ is not differentiable?

$$\hat{w} = \text{argmin}_w \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{X}_i \cdot w)^2 + \lambda ||w||_1$$
Gradient Descent

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

\[ \frac{\partial L}{\partial w} = \frac{1}{m} \sum_{i=1}^{m} (y_i - X_i \cdot w)(X_i \cdot w) + 1 \]
Sub-Gradient Descent

- What makes minimizing $f$ easy?
  - Any local minimum is a global minimum
  - $f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y$

- Subgradient at $x$: Any vector $g \in \mathbb{R}^n$ such that
  - $f(y) \geq f(x) + g^T (y - x), \forall x, y$
  - Generalizes the notion of gradient
  - Always exists for convex functions
Sub-Gradient Descent

• Assume $f$ is continuous and differentiable
• Start at arbitrary $x_0$.
• For $t=1,2,\ldots$
  - $x_{t+1} = x_t - \eta \nabla f(x_t)$
  - $x_{t+1} = x_t - \eta g(x_t)$
Sub-Gradient Descent

\[ f(x) = |x| \]

Subgradient at \( x \): \( \forall y \)

\[ f(y) \geq f(x) + g(y-x) \]

1. \( x > 0 \), \( 1y \geq x + g(y-x) , \forall y \)

\[ g(x) = 1 \]

2. \( x < 0 \), \( g(x) = -1 \)

3. \( x = 0 \), \( f(y) \geq gy , \forall y \)

\[ 1y \geq gy , \forall y \]

\[ g \in [-1, 1] \]
Sub-Gradient Descent

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

*If \( f(X_t) \) is differentiable at \( X_t \):
  \[ g(X_t) = \nabla f(X_t) \]*

*If \( f(X_t) \) is not differentiable:
  There can be many sub-gradients.
  \[ X_{t+1} = X_t - \eta g(X_t) \]
Sub-Gradient Descent

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

• Assumption: $f$ is $L$-Lipschitz
  – Gradients at nearby points cannot change drastically
  – A notion of smoothness of $f$
  \[ ||\Delta f(x) - \Delta f(y)||_2 \leq L \ ||x - y||_2 \]
• Instead: Sub-gradients are not too large
  \[ ||g(x)||^2 \leq G, \forall x \]
Sub-Gradient vs Gradient Descent
Sub-Gradient Descent

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

$$f(x_{best}) - f(x^*) \leq \frac{||x_0 - x^*||G}{\sqrt{t}}, \quad \eta = \frac{||x_0 - x^*||}{G\sqrt{t}}$$

- **Key differences from gradient descent**
  - Only guarantee error of $x_{best}$
  - Convergence is slower by $\sqrt{factor}$
  - Can’t say much about fixed step size
Gradient Descent

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

• For ML problems, gradient computation is still expensive

$$\sum_{i=1}^{m} (y_i - x_i^T \omega)^2 + \|\omega\|_1$$

$O(m)$ computations per iteration
Stochastic Gradient Descent

- Want to minimize: \( L(w) = \sum_{i=1}^{m} f(w, X_i, y_i) + h(w) \)
  - \( f, h \) is convex

- For ML problems, gradient computation is still expensive

- Stochastic Gradient Descent
  - Just compute gradient on a randomly chosen point.
  - In each iteration, choose \( p \in \{1, 2, \ldots, m\} \) at random
    - \( w_{t+1} = w_t - \eta \nabla f(w_t, X_p, y_p) \)
Stochastic Gradient Descent

• Philosophy/Advantages
  – We are already approximating true loss by minimizing loss over training set.
    • Adding another approximation will not hurt much
  – In many cases, can’t really compute full gradient
    • Online/Streaming data
    • Distributed data
  – Training time is a huge bottleneck, need fast methods
Stochastic-Gradient Descent

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

$$E[f(x_{best})] - f(x^*) \leq \frac{||x_0-x^*||}{\sqrt{t}} \frac{G}{\eta}, \quad \eta = \frac{||x_0-x^*||}{G\sqrt{t}}$$

- Key differences from (sub)gradient descent
  - Only guarantee error of $E[f(x_{best})]$
  - Convergence rate is the same as sub-gradient descent!
  - Has high variance if sub-gradients are too noisy
    - Lot of recent work on variance reduction techniques!
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