CIS 700: “algorithms for Big Data”

Lecture 8: Gradient Descent

Slides at http://grigory.us/big-data-class.html

Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

Smooth Convex Optimization

- Minimize f over $\mathbb{R}^n$:
  - f admits a minimizer x^* ($\nabla f(x^*) = 0$)
  - f is continuously differentiable and convex on $\mathbb{R}^n$:
  $\forall x, y \in \mathbb{R}^n$: $f(x) - f(y) \geq (x - y)\nabla \mathrm{f}(\mathrm{y})$
  - f is β-smooth (∇f is β-Lipschitz)
  $\forall x, y \in \mathbb{R}^n$: $\left|\left|\nabla f(x) - \nabla f(y)\right|\right| \leq \beta||x - y||$
- Example:
  - $f = \frac{1}{2}x^T Ax - b^T x$
  - $\nabla f = Ax - b \Rightarrow x^* = A^{-1}b$

Gradient Descent Method

- Gradient descent method:
  - Start with an arbitrary x_1
  - Iterate $x_{s+1} = x_s - \eta \cdot \nabla f(x_s)$
- **Thm.** If $\eta = 1/\beta$ then:

$$f(x_t) - f(x^*) \leq \frac{2\beta \left|\left|x_1 - x^*\right|\right|_2^2}{t+3}$$

- "Linear convergence", can be improved to quadratic using Nesterov's accelerated descent

Gradient Descent: Analysis

- **Lemma 1:** If f is β-smooth then $\forall x, y: \in \mathbb{R}^n$:
$$f(x) \le f(y) + \nabla f(y)^T (x-y) + \frac{\beta}{2} ||x-y||^2$$
- $f(x) - f(y) \ - \nabla f(y)^T (x-y) =$
$$\int_0^1 \nabla f\big(y + t(x-y)\big)^T (x-y) dt \ - \nabla f(y)^T (x-y)$$
$$\le \int_0^1 \beta t ||x-y||^2 dt \ = \frac{\beta}{2} ||x-y||^2$$
- Convex and β-smooth is equivalent to:
$$f(y) + \nabla f(y)^T (x-y) \ \le f(x)$$
$$\le f(y) + \nabla f(y)^T (x-y) + \frac{\beta}{2} ||x-y||^2$$

Gradient Descent: Analysis

- **Lemma 2:** If f convex and β-smooth then $\forall x, y: \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \nabla \mathrm{f}(x)^T (y - x) + \frac{1}{2\beta} \left|\left|\nabla f(x) - \nabla f(y)\right|\right|_2^2$$

- **Cor:** $\left(\nabla \mathrm{f}(\mathrm{x}) - \nabla f(y)\right)^T (x - y) \geq \frac{1}{\beta} \left|\left|\nabla f(x) - \nabla f(y)\right|\right|^2$
- $\phi^x(y) = f(y) - \nabla f(x)^T y$
- $\nabla \phi^x(y) = \nabla f(y) - \nabla f(x)$
- ϕ^x is convex, β-smooth and minimized at x:

$$\phi^x(x) - \phi^x(y) = f(x) - \nabla f(x)^T x - f(y) + \nabla f(x)^T y$$
$$\geq (x - y) \nabla \phi^x(\mathrm{y})$$

$$\left|\left|\nabla \phi^x(y_1) - \nabla \phi^x(y_2)\right|\right| = \left|\left|\nabla f(y_1) - \nabla f(y_2)\right|\right| \leq \beta ||y_1 - y_2||$$

Gradient Descent: Analysis

- **Lemma 2:** If f convex and β-smooth then $\forall x, y :\in \mathbb{R}^n$:

$$f(y) \geq f(x) + \nabla \mathrm{f}(x)^T (y - x) + \frac{1}{2\beta} \left|\left|\nabla f(x) - \nabla f(y)\right|\right|_2^2$$

- $\phi^x(y) = f(y) - \nabla f(x)^T y$
- $\nabla \phi^x(y) = \nabla f(y) - \nabla f(x)$
- $f(x) - f(y) - \nabla \mathrm{f}(x)^T (y - x) = \phi^x(x) - \phi^x(y)$

$$\leq \phi^x\left(y - \frac{1}{\beta} \nabla \phi^x(y)\right) - \phi^x(y)$$

$$\leq \nabla \phi^x(y)^T \left(-\frac{1}{\beta} \nabla \phi^x(y)\right) + \frac{\beta}{2} \left\|\frac{1}{\beta} \nabla \phi^x(y)\right\|^2 \; (by\ Lemma\ 1)$$

$$= -\frac{1}{2\beta} \left|\left|\nabla \phi^x(y)\right|\right|^2 = -\frac{1}{2\beta} \left|\left|\nabla f(x) - \nabla f(y)\right|\right|^2$$

Gradient Descent: Analysis

- Gradient descent: $x_{s+1} = x_s - 1/\beta \cdot \nabla f(x_s)$
- **Thm:** $f(x_t) - f(x^*) \leq \frac{2\beta\|x_1 - x^*\|_2^2}{t+3}$

$$f(x_{s+1}) - f(x_s) \leq \nabla f(x_s)^T (x_{s+1} - x_s) + \frac{\beta}{2}\|x_{s+1} - x_s\|^2$$
$$= -\frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

- Let $\delta_s = f(x_s) - f^*$. Then $\delta_{s+1} \leq \delta_s - \frac{1}{2\beta}\|\nabla f(x_s)\|^2$.
- $\delta_s \leq \nabla f(x_s)^T (x_s - x^*) \leq \|x_s - x^*\| \|\nabla f(x_s)\|$
- **Lem:** $\|x_s - x^*\|$ is decreasing with s.
- $\delta_{s+1} \leq \delta_s - \frac{\delta_s^2}{2\beta\|x_1 - x^*\|^2}$

Gradient Descent: Analysis

- $\delta_{s+1} \leq \delta_s - \frac{\delta_s^2}{2\beta\left|\left|x_1 - x^*\right|\right|^2}$; $\omega = \frac{1}{2\beta\left|\left|x_1 - x^*\right|\right|^2}$
- $\omega\delta_s^2 + \delta_{s+1} \leq \delta_s \Leftrightarrow \frac{\omega\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leq \frac{1}{\delta_{s+1}}$
- $\frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geq \omega \Rightarrow \frac{1}{\delta_t} \geq \omega(t-1) + \frac{1}{f(x_1) - f(x^*)}$
- $f(x_1) - f(x^*) \leq$

$$\nabla f(x^*)(x_1 - x^*) + \frac{\beta}{2}\left|\left|x_1 - x^*\right|\right|^2 = \frac{1}{4\omega}$$

- $\delta_t \leq \frac{1}{\omega(t+3)}$

Gradient Descent: Analysis

- **Lem:** $||x_s - x^*||$ is decreasing with s.
- $\left(\nabla \mathrm{f}(\mathrm{x}) - \nabla f(y)\right)^T (x - y) \geq \frac{1}{\beta} ||\nabla f(x) - \nabla f(y)||^2$

$$\Rightarrow \nabla f(y)(y - x^*) \geq \frac{1}{\beta} ||\nabla f(y)||^2$$

- $||x_{s+1} - x^*||^2 = \left|\left|x_s - \frac{1}{\beta} \nabla f(x_s) - x^*\right|\right|^2$

$$= ||x_s - x^*||^2 - \frac{2}{\beta} \nabla f(x_s)^T (x_s - x^*) + \frac{1}{\beta^2} ||\nabla f(x_s)||^2$$

$$\leq ||x_s - x^*||^2 - \frac{1}{\beta^2} ||\nabla f(x_s)||^2$$

$$||x_s - x^*||^2$$

Nesterov's Accelerated Gradient Descent

- Params: $\lambda_0 = 0, \lambda_s = \frac{1+\sqrt{1+4\lambda_{s-1}^2}}{2}, \gamma_s = \frac{1-\lambda_s}{\lambda_{s+1}}$
- Accelerated Gradient Descent ($x_1 = y_1$):
  - $y_{s+1} = x_s - \frac{1}{\beta}\nabla f(x_s)$
  - $x_{s+1} = (1-\gamma_s)y_{s+1} + \gamma_s y_s$
- Optimal convergence rate $O(1/t^2)$
- **Thm.** If f is convex and β-smooth then:

$$f(y_t) - f(x^*) \le \frac{2\beta\left|\left|x_1 - x^*\right|\right|^2}{t^2}$$

Accelerated Gradient Descent: Analysis

- $f\left(x - \frac{1}{\beta}\nabla \mathrm{f(x)}\right) - f(y) \leq$

$$\leq f\left(x - \frac{1}{\beta}\nabla \mathrm{f(x)}\right) - f(x) + \nabla f(x)^T(x-y)$$

$$\leq \nabla f(x)^T\left(x - \frac{1}{\beta}\nabla \mathrm{f(x)} - x\right) + \frac{\beta}{2}\left\|x - \frac{1}{\beta}\nabla \mathrm{f(x)} - x\right\|_2^2 + \nabla f(x)^T(x-y) \qquad \text{(by Lemma 1)}$$

$$= -\frac{1}{2\beta}\left||\nabla f(x)|\right|^2 + \nabla \mathrm{f(x)^T(x - y)}$$

Accelerated Gradient Descent: Analysis

- $f\left(x - \frac{1}{\beta}\nabla \mathrm{f}(\mathrm{x})\right) - f(y) \leq -\frac{1}{2\beta}||\nabla f(x)||^2 + \nabla \mathrm{f}(\mathrm{x})^{\mathrm{T}}(\mathrm{x} - \mathrm{y})$
- Apply to $x = x_s, y = y_s$:

$$f(y_{s+1}) - f(y_s) = f\left(x_s - \frac{1}{\beta}\nabla \mathrm{f}(\mathrm{x_s})\right) - f(y_s)$$

$$\leq -\frac{1}{2\beta}||\nabla f(x_s)||^2 + \nabla f(x_s)(x_s - y_s)$$

$$= -\frac{\beta}{2}||y_{s+1} - x_s||^2 - \beta(y_{s+1} - x_s)^T(x_s - y_s) \quad (1)$$

- Apply to $x = x_s, y = x^*$:

$$f(y_{s+1}) - f(x^*) \leq -\frac{\beta}{2}||y_{s+1} - x_s||^2 - \frac{\beta}{2}(y_{s+1} - x_s)^T(x_s - x^*)$$

(2)

Accelerated Gradient Descent: Analysis

- (1) x $(\lambda_s - 1)$ + (2), for $\delta_s = f(y_s) - f(x^*)$:

$$\lambda_s \delta_{s+1} - (\lambda_s - 1)\delta_s \leq$$
$$-\frac{\beta}{2}\lambda_s ||y_{s+1} - x_s||^2 - \beta(y_{s+1} - x_s)^T(\lambda_s x_s - (\lambda_s - 1)y_s - x^*)$$

- (x) λ_s and use $\lambda_{s-1}^2 = \lambda_s^2 - \lambda_s$:

$$\lambda_s^2 \delta_{s+1} - \lambda_{s-1}^2 \delta_s$$
$$\leq -\frac{\beta}{2}(||\lambda_s(y_{s+1} - x_s)||^2 + 2\lambda_s(y_{s+1} - x_s)^T(\lambda_s x_s - (\lambda_s - 1)y_s - x^*))$$

- It holds that:

$$||\lambda_s(y_{s+1} - x_s)||^2 + 2\lambda_s(y_{s+1} - x_s)^T(\lambda_s x_s - (\lambda_s - 1)y_s - x^*)) =$$
$$||\lambda_s y_{s+1} - (\lambda_s - 1)y_s - x^*||^2 - ||\lambda_s x_s - (\lambda_s - 1)y_s - x^*||^2$$

Accelerated Gradient Descent: Analysis

- By definition of AGD:

$$x_{s+1} = y_{s+1} + \gamma_s(y_s - y_{s+1}) \Leftrightarrow$$
$$\lambda_{s+1}x_{s+1} = \lambda_{s+1}y_{s+1} + (1 - \lambda_s)(y_s - y_{s+1}) \Leftrightarrow$$
$$\lambda_{s+1}x_{s+1} - (\lambda_{s+1} - 1)y_{s+1} = \lambda_s y_{s+1} - (\lambda_s - 1)y_s$$

- Putting last three facts together for $u_s = \lambda_s x_s - (\lambda_s - 1)y_s - x^*$ we have:

$$\lambda_s^2 \delta_{s+1} - \lambda_{s-1}^2 \delta_s \leq \frac{\beta}{2}\left(||u_s||^2 - ||u_{s+1}||^2\right)$$

- Adding up over $s = 1$ to $s = t - 1$:

$$\delta_t \leq \frac{\beta}{2\lambda_{t-1}^2}||u_1||^2$$

- By induction $\lambda_{t-1} \geq \frac{t}{2}$. Q.E.D.