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

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

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

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

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

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

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

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

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

$$\left||x_s - x^*|\right|^2$$

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:

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

Accelerated Gradient Descent: Analysis

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

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

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

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

Accelerated Gradient Descent: Analysis

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

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

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

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

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

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

Accelerated Gradient Descent: Analysis

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

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

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

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

- It holds that:

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

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.

Constrained Convex Optimization

- Non-convex optimization is NP-hard:

$$\sum_i x_i^2(1 - x_i)^2 = 0 \Leftrightarrow \forall i\colon\ x_i \in \{0,1\}$$

- Knapsack:
  - Minimize $\sum_i c_i x_i$
  - Subject to: $\sum_i w_i x_i \leq W$
- Convex optimization can often be solved by ellipsoid algorithm in $poly(n)$ time, but too slow

Convex multivariate functions

- Convexity:
  - $\forall x, y \in \mathbb{R}^n$: $f(x) \geq f(y) + (x - y)\nabla \mathrm{f(y)}$
  - $\forall x, y \in \mathbb{R}^n, 0 \leq \lambda \leq 1$:
    $$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y)$$
- If higher derivatives exist:
  $$f(x) = f(y) + \nabla f(y) \cdot (x - y) + (x - y)^T \nabla^2 f(x)(x - y) + \cdots$$
- $\nabla^2 f(x)_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$ is the Hessian matrix
- f is convex iff it's Hessian is positive semidefinite, $y^T \nabla^2 f y \geq 0$ for all y.

Examples of convex functions

- ℓ_p-norm is convex for $1 \leq p \leq \infty$:

$$\left|\left|\lambda x + (1-\lambda)y\right|\right|_p \leq \left|\left|\lambda x\right|\right|_p + \left|\left|(1\ -\lambda)y\right|\right|_p$$
$$= \lambda\left|\left|x\right|\right|_p + (1\ -\lambda)\left|\left|y\right|\right|_p$$

- $f(x) = \log(e^{x_1} + e^{x_2} + \cdots e^{x_n})$

$$\max(x_1, \ldots, x_n) \leq f(x) \leq \max(x_1, \ldots, x_n) + \log n$$

- $f(x) = x^T A x$ where A is a p.s.d. matrix, $\nabla^2 f = A$
- Examples of constrained convex optimization:
  - (Linear equations with p.s.d. constraints):

  minimize: $\frac{1}{2} x^T A x - b^T x$ (solution satisfies $Ax = b$)
  - (Least squares regression):

  Minimize: $\left|\left|Ax - b\right|\right|_2^2 = \mathrm{x}^\mathrm{T}\mathrm{A}^\mathrm{T}\mathrm{A}\,\mathrm{x}\ - 2\ (\mathrm{Ax})^\mathrm{T}\mathrm{b} + \mathrm{b}^\mathrm{T}\mathrm{b}$

Constrained Convex Optimization

- General formulation for convex f and a convex set K:

$$minimize: f(x) \text{ subject to: } x \in K$$

- Example (SVMs):
  - Data: $X_1, \dots, X_N \in \mathbb{R}^n$ labeled by $y_1, \dots, y_N \in \{-1,1\}$ (spam / non-spam)
  - Find a linear model:

$$W \cdot X_i \geq 1 \Rightarrow X_i \text{ is spam}$$

$$W \cdot X_i \leq -1 \Rightarrow X_i \text{ is non-spam}$$

$$\forall i: 1 - y_i W X_i \leq 0$$

- More robust version:

$$minimize: \sum_i Loss\left(1 - W(y_i X_i)\right) + \lambda \left|\left|W\right|\right|_2$$

  - E.g. hinge loss Loss(0,t)=max(0,t)
  - Another regularizer: $\lambda \left|\left|W\right|\right|_1$ (favors sparse solutions)

Gradient Descent for Constrained Convex Optimization

- (Projection): $x \notin K \rightarrow y \in K$

$$\mathrm{y} = \mathrm{argmin}_{z \in K} \left|\left|z - x\right|\right|_2$$

- Easy to compute for $\left|\left|\cdot\right|\right|_2^2$: $y = x / \left|\left|x\right|\right|_2^2$
- Let $\left|\left|\nabla f(x)\right|\right|_2 \leq G, \max_{x,y \in K} \left(\left|\left|x - y\right|\right|_2\right) \leq D$.
- Let $T = \frac{4D^2G^2}{\epsilon^2}$
- Gradient descent (gradient + projection oracles):
  - Let $\eta = D / G\sqrt{T}$
  - Repeat for $i = 0, \dots, T$:
    - $y^{(i+1)} = x^{(i)} + \eta \nabla f\left(x^{(i)}\right)$
    - $x^{(i+1)} =$ projection of $y^{(i+1)}$ on K
  - Output $z = \frac{1}{T} \sum_i x^{(i)}$

Gradient Descent for Constrained Convex Optimization

- $\left\|x^{(i+1)} - x^*\right\|_2^2 \le \left\|y^{(i+1)} - x^*\right\|_2^2$

$$= \left\|x^{(i)} - x^* - \eta \nabla f\left(x^{(i)}\right)\right\|_2^2$$

$$= \left\|x^{(i)} - x^*\right\|_2^2 + \eta^2 \left\|\nabla f\left(x^{(i)}\right)\right\|_2^2 - 2\eta \nabla f\left(x^{(i)}\right) \cdot \left(x^{(i)} - x^*\right)$$

- Using definition of G:

$$\nabla f\left(x^{(i)}\right) \cdot \left(x^{(i)} - x^*\right) \le \frac{1}{2\eta}\left(\left\|x^{(i)} - x^*\right\|_2^2 - \left\|x^{(i+1)} - x^*\right\|_2^2\right) + \frac{\eta}{2} G^2$$

- $f\left(x^{(i)}\right) - f(x^*) \le \frac{1}{2\eta}\left(\left\|x^{(i)} - x^*\right\|_2^2 - \left\|x^{(i+1)} - x^*\right\|_2^2\right) + \frac{\eta}{2} G^2$
- Sum over $i = 1, \dots, T$:

$$\sum_{i=1}^{T} f\left(x^{(i)}\right) - f(x^*) \le \frac{1}{2\eta}\left(\left\|x^{(0)} - x^*\right\|_2^2 - \left\|x^{(T)} - x^*\right\|_2^2\right) + \frac{T\eta}{2} G^2$$

Gradient Descent for Constrained Convex Optimization

- $\sum_{i=1}^{T} f\left(x^{(i)}\right) - f(x^*) \leq \frac{1}{2\eta}\left(\left\|x^{(0)} - x^*\right\|_2^2 - \left\|x^{(T)} - x^*\right\|_2^2\right) + \frac{T\eta}{2} G^2$
- $f\left(\frac{1}{T}\sum_i x^{(i)}\right) \leq \frac{1}{T}\sum_i f\left(x^{(i)}\right)$:

$$f\left(\frac{1}{T}\sum_i x^{(i)}\right) - f(x^*) \leq \frac{D^2}{2\eta T} + \frac{\eta}{2} G^2$$

- Set $\eta = \frac{D}{G\sqrt{T}} \Rightarrow$ RHS $\leq \frac{DG}{\sqrt{T}} \leq \epsilon$

Online Gradient Descent

- Gradient descent works in a more general case:
- $f \rightarrow$ sequence of convex functions $f_1, f_2 \dots, f_T$
- At step i need to output $x^{(i)} \in K$
- Let x^* be the minimizer of $\sum_i f_i(w)$
- Minimize regret:

$$\sum_i f_i\left(x^{(i)}\right) - f_i(x^*)$$

- Same analysis as before works in online case.

Stochastic Gradient Descent

- (Expected gradient oracle): returns g such that $\mathbb{E}_g[g] = \nabla f(x)$.
- Example: for SVM pick randomly one term from the loss function.
- Let g_i be the gradient returned at step i
- Let $f_i = g_i x$ be the function used in the i-th step of OGD
- Let $z = \frac{1}{T}\sum_i x^{(i)}$ and x^* be the minimizer of f.

Stochastic Gradient Descent

- **Thm.** $\mathbb{E}[f(z)] \leq f(x^*) + \frac{DG}{\sqrt{T}}$ where G is an upper bound of any gradient output by oracle.
- $f(z) - f(x^*) \leq \frac{1}{T}\sum_i (f\left(x^{(i)}\right) - f(x^*))$ (convexity)

$$\leq \frac{1}{T}\sum_i \nabla f\left(x^{(i)}\right)\left(x^{(i)} - x^*\right)$$

$= \frac{1}{T}\sum_i \mathbb{E}\left[g_i(x^{(i)} - x^*)\right]$ (grad. oracle)

$$= \frac{1}{T}\sum_i \mathbb{E}[f_i(x^{(i)}) - f_i(x^*)]$$

$$= \frac{1}{T}\mathbb{E}[\sum_i f_i(x^{(i)}) - f_i(x^*)]$$

- $\mathbb{E}[]$ = regret of OGD , always $\leq \epsilon$