CSCI B609:
“Foundations of Data Science”

Lecture 13/14: Gradient Descent and Applications

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

Grigory Yaroslavtsev

http://grigory.us

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:
  $$f(x) = f(y) + \nabla f(y) \cdot (x - y) \\ + (x-y)^T \nabla^2 f(x)(x-y) + \cdots$$
- $\nabla^2 f(x)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ 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: $\frac{1}{2} \mathrm{x}^T A x - b^T x$ (solution satisfies $Ax = b$)
  - (Least squares regression):

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

Constrained Convex Optimization

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

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

- Example (SVMs):
  - Data: $X_1, \ldots, X_N \in \mathbb{R}^n$ labeled by $y_1, \ldots, y_N \in \{-1,1\}$ (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(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 $i = 0, \ldots, T$:
    - $y^{(i+1)} = x^{(i)} - \eta \nabla f\left(x^{(i)}\right)$
    - $x^{(i+1)} =$ projection of $y^{(i+1)}$ on K
  - Output $z = \frac{1}{T} \sum_i x^{(i)}$

Gradient Descent for Constrained Convex Optimization

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

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

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

- Using definition of G:

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

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

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

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

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

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

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

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

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

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