Lecture 6: stochastic gradient descent

Sanjeev Arora
Elad Hazan
Admin

- Exercise 2 (implementation) this Thu, in class
- Exercise 3 (written), this Thu, in class
- Movie – “Ex Machina” + discussion panel w. Prof. Hasson (PNI)
  Wed Oct. 5\textsuperscript{th} 19:30
  tickets from Bella; room 204 COS

- Today: special guest - Dr. Yoram Singer @ Google
Recap

• Definition + fundamental theorem of statistical learning, motivated efficient algorithms/optimization
• Convexity and it’s computational importance
• Local greedy optimization – gradient descent
Agenda

• Stochastic gradient descent
• Dr. Singer on opt @ google & beyond
Mathematical optimization

Input: function $f: K \mapsto R$, for $K \subseteq R^d$
Output: point $x \in K$, such that $f(x) \leq f(y) \ \forall \ y \in K$
Prefer Convex Problems
Convex functions: local $\rightarrow$ global

Sum of convex functions $\rightarrow$ also convex
Convex Functions and Sets

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \text{dom} f$ and any $a \in [0, 1]$,

$$f(ax + (1 - a)y) \leq af(x) + (1 - a)f(y)$$

A set $C \subseteq \mathbb{R}^n$ is convex if for $x, y \in C$ and any $a \in [0, 1]$,

$$ax + (1 - a)y \in C$$
Special case: optimization for linear classification

Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, find hyperplane (through the origin w.l.o.g) such that:

$$\min_w \# \text{ of mistakes}$$

$$\min \ell(w^T x_i, y_i) \text{ for a convex loss function}$$
Convex relaxation for 0-1 loss

1. Ridge / linear regression
   \[ \ell(w, x_i, y_i) = (w^T x_i - y_i)^2 \]

2. SVM
   \[ \ell(w, x_i, y_i) = \max\{0, 1 - y_i w^T x_i\} \]

i.e. for \(|w| = |x_i| = 1, \) We have:

\[ 1 - y_i w^T x_i = \begin{cases} 
  0 & y_i = w^T x_i \\
  \leq 2 & y_i \neq w^T x_i 
\end{cases} \]
Greedy optimization: gradient descent

• Move in the direction of steepest descent:

\[ x_{t+1} \leftarrow x_t - \eta \nabla f(x_t) \]

We saw: for certain step size choice,

\[ f\left(\frac{1}{T} \sum_t x_t\right) \leq \min_{x^* \in K} f(x^*) + \frac{1}{\sqrt{T}} \]
GD for linear classification

\[
\min_{|w| \leq 1} \frac{1}{m} \sum_{i} \ell(w^T x_i, y_i)
\]

\[
w_{t+1} = w_t - \eta \frac{1}{m} \sum_{i} \ell' (w_t^T x_i, y_i)x_i
\]

- Complexity? \( \frac{1}{\epsilon^2} \) iterations, each taking \( \sim \) linear time in data set
- Overall \( O \left( \frac{md}{\epsilon^2} \right) \) running time, \( m \)=# of examples in \( \mathbb{R}^d \)
- Can we speed it up??
GD for linear classification

- What if we take a single example, and compute gradient only w.r.t its loss??

- Which example?
  - --> uniformly at random...

- Why would this work?
SGD for linear classification

\[
\min_{|w| \leq 1} \frac{1}{m} \sum_i \ell(w^T x_i, y_i)
\]

\[
w_{t+1} = w_t - \eta \ell'(w_t^T x_{i_t}, y_{i_t}) x_{i_t}
\]

• Uniformly at random?! \(i_t \sim U[1, ..., m]\)

• Each iteration is much faster \(O(md) \rightarrow O(d)\), convergence??

Has expectation = full gradient
Crucial for SGD: linearity of expectation and derivatives

Let \( f(w) = \frac{1}{m} \sum_i \ell_i(w) \), then for \( i_t \sim U[1, \ldots, m] \) chosen uniformly at random, we have

\[
E[\nabla \ell_{i_t}(w)] = \sum_{i=1}^{m} \frac{1}{m} \nabla \ell_i(w) = \nabla \left( \frac{1}{m} \sum_i \ell_i(w) \right) = \nabla \ell(w)
\]
Greedy optimization: gradient descent

• Move in a random direction, whose expectation is the steepest descent:

• Denote by $\nabla f (w)$ a vector random variable whose expectation is the gradient,

$$E[\nabla f (w)] = \nabla f (w)$$

$$x_{t+1} \leftarrow x_t - \eta \nabla f (x_t)$$
Stochastic gradient descent – constrained case

\[
y_{t+1} \leftarrow x_t - \eta \nabla f(x_t), \quad E[\nabla f(x_t)] = \nabla f(x_t)
\]

\[
x_{t+1} = \arg \min_{x \in K} |y_{t+1} - x|
\]
Stochastic gradient descent – constrained set

Let:
• $G = \text{upper bound on norm of gradient estimators}$
  \[ |\nabla f(x_t)| \leq G \]

• $D = \text{diameter of constraint set}$
  \[ \forall x, y \in K . \ |x - y| \leq D \]

Theorem: for step size $\eta = \frac{D}{G \sqrt{T}}$

\[ E[f\left(\frac{1}{T} \sum_t x_t\right)] \leq \min_{x^* \in K} f(x^*) + \frac{DG}{\sqrt{T}} \]
Proof:

1. We have proved: (for any sequence of $\nabla_t$)

$$\left(\frac{1}{T} \sum_t \nabla_t^T x_t\right) \leq \min_{x^* \in K} \frac{1}{T} \sum_t \nabla_t^T x^* + \frac{DG}{\sqrt{T}}$$

2. By property of expectation:

$$E[f\left(\frac{1}{T} \sum_t x_t\right) - \min_{x^* \in K} f(x^*)] \leq \left(\frac{1}{T} \sum_t \nabla f(x_t)^T (x_t - x^*)\right) \leq \frac{DG}{\sqrt{T}}$$
Summary

• Mathematical & convex optimization
• Gradient descent algorithm, linear classification
• Stochastic gradient descent