Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Bhaskar Mukhoty
Instructor: Purushottam Kar
Date: October 6, 2016

LECTURE
17

Stochastic Optimization

1 Introduction

In this lecture we would look at stochastic optimization techniques. But before delving into the topic, we would review the measure theoretic definition of probalitiy and few useful notions associated with it.

σ-Algebra

Let, Ω be a any set (countable or uncountable) and its power set be 2^{Ω}. A collection of subsets of Ω, say $\mathcal{F} \subset 2^{\Omega}$, is called a σ-algebra, if the following conditions hold.

i. $\Omega \in \mathcal{F}$
ii. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, where $A^c = \Omega - A$
iii. $A_1, A_2, ... \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Observe that the null set $\phi \in \mathcal{F}$, by rule (i) and (ii).
The smallest possible σ-algebra over Ω, is $\{\phi, \Omega\}$, called the trivial σ-algebra.

The tuple $(\Omega, \mathcal{F})$ is called a measureable space, and each element of $A \in \mathcal{F}$, is a measurable set.

Measure

Let, $[0, \infty]$ is the set of non-negetive real numbers with infinity.

A positive measure, associated with a measurable space $(\Omega, \mathcal{F})$, is a function $\mu : \mathcal{F} \to [0, \infty]$, which satisfies the following conditions.

i. $\mu(\phi) = 0$
ii. $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$, where $A_i \cap A_j = \phi$, whenever $i \neq j$.

A measurable space $(\Omega, \mathcal{F})$, with a measure μ defined on it, is called a measure space, denoted by $(\Omega, \mathcal{F}, \mu)$.

Probability Measure

A probablity measure $(\Omega, \mathcal{F}, P)$, is a positive measure with the property, $P(\Omega) = 1$.

Here, Ω is the set of all possible outcomes of a random experiment, called the sample space. An event space is a σ-algebra $\mathcal{F}$ defined on Ω. Any element, $A \in \mathcal{F}$ is an event.

Random Variable

Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a function $X : \Omega \to \mathbb{R}$, from the sample space to the set of real numbers, such that,

$$\forall x \in \mathbb{R}, \quad \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

A function between two measureable spaces is called a measurable function, if every pre-image of a measurable set is measurable. The above condition ensures that the random variable X, is a measureble functionRosenthal (2006).

Cumulative Distribution of a random variable

The cumulative distribution function(CDF) of a random varible X, gives the probability that the random variable takes a value less than equal to some value of the variable. It is given by,

$$F_X(x) = P(X \leq x)$$

Equivalently, for a continuous random variable X, it is given as:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx$$

Here, $f_X(x)$ is the probability density function(PDF) of the continuous random variable X. The area defined by an interval $[a, b]$ under $f_X(x)$, gives the probablity of X taking the value in the interval.

$$P(a \leq X \leq b) = \int_{a}^{b} f_X(x)\, dx$$

Expectation of a random variable

Case 1: X is a discrete random variable, i.e. the values that X takes are countable.
$X \in \{x_1, x_2, ...\}$, then the expectation of X is given by,

$$E[X] = \sum_{i=0}^{\infty} x_i\, P(X = x_i)$$

Ex. Jacod and Protter (2003) Suppose X takes all its values in $N = \{0, 1, 2, ..\}$

Then,

$$E[X] = \sum_{i=0}^{\infty} i\, P(X = i) = \sum_{i=1}^{\infty} \sum_{j=1}^{i} P(X = i) = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} P(X = i) = \sum_{j=1}^{\infty} P(X \geq j)$$

Case 2: X is continuous random variable, with a probability density function $f_X(x)$

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

If $X > 0$,

$$E[X] = \int_{0}^{\infty} x\, f_X(x)\, dx = \int_{0}^{\infty} \int_{0}^{x} f_X(x)\, dt\, dx = \int_{0}^{\infty} \int_{t}^{\infty} f_X(x)\, dx\, dt = \int_{0}^{\infty} P(X \geq t)\, dt$$

Also, in general,

$$E[X] = \int_0^\infty P(X \geq t)\, dt - \int_{-\infty}^0 P(X \leq -t)\, dt$$

We can prove that expection is a linear operator. That is, the following holds with a, b, c being constants.

$$E[c] = c$$
$$E[aX + bY] = aE[X] + bE[Y]$$

The conditional expectation of a discrete random variable X, given another discrete random Y taking a value y, is defined as,

$$E[X|Y] = E[X|Y = y] = \sum_x x\, P[X = x|Y = y]$$

We can observe that it is a random variable on Y.

The following identity is called the law of total expectation:

$E[E[X|Y]]$
$= E[\sum_x x\, P[X = x|Y = y]]$
$= \sum_y \sum_x x\, P[X = x|Y = y]P[Y = y]$
$= \sum_x \sum_y x\, \frac{P[X = x \cap Y = y]}{P[y = y]} P[y = y]$
$= \sum_x x \sum_y P[X = x \cap Y = y]$
$= \sum_x x\, P[X = x] = E[X]$

The random variables X and Y are said to be independent, if

$$E[X|Y] = E[X]$$

Variance

The variance of a random variable is a measure of dispersion from its mean. Defined as follows,
$Var[X] = E[(X - E[X])^2]$
$= E[X^2 - 2X\, E[X] + E[X]^2]$
$= E[X^2] - 2E[X]^2 + E[X]^2]$ (by linearity of expectation)
$= E[X^2] - E[X]^2$

Notion of Independence

Let, $\{A_1, A_2, ..., A_n\}$ be a finite set of events. The events are pairwise independent, if,

$$P(A_i \cap A_j) = P(A_i)P(A_j) \qquad \forall i, j$$

And the events are mutually independent, if

$$P(\bigcap_{i=1}^k A_i) = \prod_{i=1}^k P(A_i) \qquad \forall\, k \leq n$$

Equivalent notions of pairwise and mutual independence hold for random variables. Two random variables X and Y are said to be independent, if the events A_x and B_y are independent for all possible x and y.

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$
$$B_y = \{\omega \in \Omega : Y(\omega) \leq y\}$$

The independence of a finite set of random variables, are defined in terms of events constructed as above.

2 Stochastic Approximation

Stochastic approximation is a family of iterative algorithms, which tries find some property of the function

$$f(x) = \mathop{\mathbb{E}}_{\theta}[F(x, \theta)]$$

The expectation is over the parameter θ and it is not directly computable. Instead, samples or gradient of the function f at some choosen point x is available to the method. We assume that there is an oracle, that knows the function f and provides the method information about f at the specific points x that are quaried to it.

We have different models of the oracle depending upon the different kind of quires about the function. Following are few of them.

Definition 17.1. Deterministic First Order oracle: When queried with a point x, it returns $g(x) \leftarrow DFO(f, x)$, where $g(x) \in \partial f(x)$, is a subgradient of the function f at the point x. Observe, that the parameter f, passed in the DFO query is implicit. The method has no knowledge about it, but the oracle does.

Definition 17.2. Stochastic First Order oracle: When queried with a point x, it returns $g(x) \leftarrow SFO(f, x)$, where $E[g(x)] \in \partial f(x)$. If f is differentiable at the point x, then $E[g(x)] = \nabla f(x)$

Definition 17.3. Stochastic Zeroth Order oracle: Similarly, zero'th order oracle $\hat{f}(x) \leftarrow SO(f, x)$, where $E[\hat{f}(x)] = f(x)$. If f is differentiable at the point x, then $E[g(x)] = \nabla f(x)$

2.1 Robbins-Monro algorithm

This algorithm introduced in 1951Robbins and Monro (1951) solves the problem,

$$f(x) = \alpha, \text{ where, } f(x) = \mathop{\mathbb{E}}_{\theta}[F(x, \theta)]$$

They proved that the following iterative rule converges to x^* in L_2, where $f(x^*) = \alpha$

$$x^{t+1} \leftarrow x^t - \eta_t(SO(f, x^t) - \alpha)$$

2.2 Widrow-Hoff algorithm

Now, consider the linear regression problem where we want to minimize the loss function L,Schapire (2013)

$$\min_w \ L(w, x, y) = \min_w \ (y - w^T x)^2$$
$$\nabla_w L(w, x, y) = -2(y - w^T x)x$$

The irerative algorithm also known as Adaptive Linear Neurone or ADALINE was proposed by Widrow-Hoff, uses the following update function.

$$w^{t+1} \leftarrow w^t - \frac{\eta_t}{2} \nabla_w L(w, x, y)$$
$$w^{t+1} \leftarrow w^t + \eta_t (y^t - (w^t)^T x^t) x^t$$

We will later see, that this a stochastic gradient descent update rule for linear regression.

3 Stochastic Optimization

Stochastic optimization is a class of stochastic approximation problems where we want to minimize the funtion f, as defined in the previous section.

$$\min_{x \in \mathcal{C}} \ f(x) \text{ where,} \qquad f(x) = \mathop{\mathbb{E}}_{\theta}[F(x, \theta)]$$

We can pose, our well known problems in this format.

Example 17.1. Support Vector Machine :

$$\min_x \quad \frac{1}{2}\|x\|_2^2 + \frac{c}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \theta_i^T x)$$

Stochastic form:

$$\min_x \quad \mathop{\mathbb{E}}_{\theta} \big[\frac{1}{2}\|x\|_2^2 + c \ \max(0, 1 - y\theta^T x) \big]$$

Observe that f is a non-differentiable function over here. Hence, SFO will return a subgradient, instead of gradient. And,

$$g^t = \begin{cases} x, & \text{if } 1 - y\theta^T x \leq 0 \\ x - cy\theta, & 1 - y\theta^T x > 0 \end{cases}$$

Example 17.2. Linear Regression :

$$\min_x \quad \frac{\lambda}{2}\|x\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} (y_i - x^T \theta_i)^2$$

Stochastic form:

$$\min_x \quad \mathop{\mathbb{E}}_{\theta} \big[\frac{\lambda}{2}\|x\|_2^2 + (y - x^T \theta)^2 \big]$$

Algorithm 1: Stochastic Gradient Descent

Input: $SFO(f,.)$, step lengths η_t, constraint set $\mathcal{C}$
Output: $\mathbf{x}* \in \mathcal{C}$

1: $x^0 \in \mathcal{C}$
2: **for** $t = 0, 1, \ldots, T$ **do**
3: $\quad \mathbf{g}^t \leftarrow SFO(f, \mathbf{x}^t)$ `//gradient obtained from Stochastic First Order oracle at point` $\mathbf{x}^t$
4: $\quad \mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{x}^t - \eta_t \mathbf{g}^t)$
5: **end for**
6: **return** $\mathbf{x}^* = \frac{1}{T} \sum_{t=0}^{T} \mathbf{x}^t$

3.1 Finite Sample Approximation

One of simplest way to estimate $\underset{\theta}{\mathbb{E}}[F(x,\theta)]$ is to take a finite number of instances of the function $F(x,\theta)$. To optimize f on x, we could take $\theta^1, \theta^2, ..., \theta^m$ iid's over θ. Then,

$$\hat{x}_m = \arg\min_{x \in \mathcal{C}} \frac{1}{m} \sum_{i=1}^{m} F(x, \theta^i)$$

In order to understand effectiveness of such a method, it can be proved that the following result holds:

$$\mathbb{P}\left[f(\hat{x}_m) \leq f^* + c\sqrt{\frac{lg(\frac{1}{\delta})}{m}} \right] = 1 - \delta \quad \text{where, } f^* = \min_{x \in \mathcal{C}} f(x)$$

3.2 Stochastic Programming

In this method we repeadly request θ^t. Then, use $F(x, \theta^t)$ in gradient descent or Newton's Method.

3.3 Stochastic Gradient Descent

The algorithm 1 lists the stochastic gradient descent algorithm. The algorithm assumes that it has access to a stochastic first order oracle. The main difference of the method from the original gradient descent is that it performs each iteration step with respect to single data point, while gradient descent does each parameter updation by taking the whole training set. As a result SGD oscilates before convering to a local minima, but is possible to implement for large data sets.

References

Jean Jacod and Philip E Protter. *Probability essentials*. Springer Science & Business Media, 2003.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Jeffrey Seth Rosenthal. *A first look at rigorous probability theory.* World Scientific, 2006.

Rob Schapire. COS 511: Theoretical Machine Learning. (Lecture 17), 2013.