Foundations for Machine Learning

L. Y. Stefanus

TU Dresden, June-July 2018
Reference

- Shai Shalev-Shwartz and Shai Ben-David. UNDERSTANDING MACHINE LEARNING: From Theory to Algorithms. Cambridge University Press, 2014.
Stochastic Gradient Descent (SGD)
Gradient Descent

• Before we study the **stochastic gradient descent** method, we first study the **standard gradient descent** approach for minimizing a differentiable convex function $f(w)$.

• Gradient descent is an iterative optimization procedure in which at each step we improve the solution by taking a step along the negative of the gradient of the function to be minimized at the current point.
• The gradient of a differentiable function \( f: \mathbb{R}^d \to \mathbb{R} \) at \( w \), denoted as \( \nabla f(w) \), is the vector of partial derivatives of \( f \), namely,

\[
\nabla f(w) = \left( \frac{\partial f(w)}{\partial w[1]}, \ldots, \frac{\partial f(w)}{\partial w[d]} \right).
\]

• Gradient descent is an iterative algorithm. We start with an initial value of \( w \) (say, \( w^{(1)} = 0 \)). Then, at each iteration, we take a step in the direction of the negative of the gradient at the current point. That is, the update step is

\[
\tag{14.1} w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)}),
\]

where \( \eta > 0 \) is a parameter.

• Intuitively, …
• Intuitively, since the gradient points in the direction of the greatest rate of increase of $f$ around $w^{(t)}$, the algorithm makes a small step in the opposite direction, thus decreasing the value of the function.

• Eventually, after $T$ iterations, the algorithm outputs the averaged vector, $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$, the last vector, $w^{(T)}$, or the best performing vector, $\arg\min_{t \in [T]} f(w^{(t)})$. But taking the average turns out to be rather useful, especially when we generalize gradient descent to non-differentiable functions and to the stochastic case.
Complexity of GD

Corollary 14.2

Let $f$ be a convex, $\rho$-Lipschitz function, and let $w^* \in \text{argmin}_{\{w: \|w\| \leq B\}} f(w)$. If we run the GD algorithm on $f$ for $T$ steps with $\eta = \frac{B^2}{\rho^2 T}$, then the output vector $\bar{w}$ satisfies

$$f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Furthermore, for every $\epsilon > 0$, to achieve $f(\bar{w}) - f(w^*) \leq \epsilon$, it suffices to run the GD algorithm for a number of iterations that satisfies

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}.$$
Subgradients

• The GD algorithm requires that the function \( f \) be differentiable.

• It turns out that the GD algorithm can be applied to non-differentiable functions by using subgradient of \( f(w) \) at \( w^{(t)} \), instead of the gradient.

• Notice that for a convex function \( f \), the gradient at \( w \) defines the slope of a tangent that lies below \( f \), that is,

\[
\forall u, f(u) \geq f(w) + < u - w, \nabla f(w) >.
\]
The existence of a tangent that lies below convex $f$ is an important property of convex functions, which is in fact an alternative characterization of convexity.

**Lemma 14.3** Let $S$ be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff for every $w \in S$ there exists $v$ such that

$$\forall u \in S, \quad f(u) \geq f(w) + \langle u - w, v \rangle.$$  

(14.8)

**Definition 14.4**
A vector $v$ that satisfies Equation (14.8) is called a subgradient of $f$ at $w$. The set of subgradients of $f$ at $w$ is called the differential set and denoted $\partial f(w)$. 
• Illustration of subgradients of a non-differentiable convex function:

For scalar functions, a subgradient of a convex function $f$ at $w$ is a slope of a line that touches $f$ at $w$ and is not above $f$ elsewhere.
Computing Subgradients

- How do we construct subgradients of a given convex function?
- If a convex function $f$ is differentiable at a point $w$, then the differential set is trivial, because $\partial f(w)$ contains a single element, namely, the gradient of $f$ at $w$, $\nabla f(w)$.
- For many practical uses, we do not need to calculate the whole set of subgradients at a given point, as one member of this set would suffice.
Example

- Consider the absolute value function $f(x) = |x|$.
- We can easily construct the **differential set** for the differentiable parts of $f$, and the only point that requires special attention is $x = 0$.
- At that point, the differential set is the set of all numbers between $-1$ and $1$.
- Hence:

$$\partial f(x) = \begin{cases} 
\{1\} & \text{if } x > 0 \\
\{-1\} & \text{if } x < 0 \\
[-1,1] & \text{if } x = 0 
\end{cases}$$
Subgradients of Lipschitz Functions

• Recall that a function $f: A \rightarrow \mathbb{R}$ is $\rho$-Lipschitz if for all $u, v \in A$:

$$|f(u) - f(v)| \leq \rho \|u - v\|.$$

• The following lemma gives an equivalent definition using norms of subgradients.

**Lemma 14.7** Let $A$ be a convex open set and let $f: A \rightarrow \mathbb{R}$ be a convex function. Then, $f$ is $\rho$-Lipschitz over $A$ iff for all $w \in A$ and $v \in \partial f(w)$ we have that $\|v\| \leq \rho$. 
Proof. Assume that for all \( v \in \partial f(w) \) we have that \( ||v|| \leq \rho \). Since \( v \in \partial f(w) \) we have

\[
  f(w) - f(u) \leq \langle v, w - u \rangle.
\]

Bounding the right-hand side using Cauchy-Schwartz inequality we obtain

\[
  f(w) - f(u) \leq \langle v, w - u \rangle \leq ||v|| ||w - u|| \leq \rho ||w - u||.
\]

An analogous argument can show that \( f(u) - f(w) \leq \rho ||w - u|| \). Hence \( f \) is \( \rho \)-Lipschitz.

Now assume that \( f \) is \( \rho \)-Lipschitz. Choose some \( w \in A, v \in \partial f(w) \). Since \( A \) is open, there exists \( \varepsilon > 0 \) such that \( u = w + \varepsilon v/||v|| \) belongs to \( A \). Therefore,

\[
  \langle u - w, v \rangle = \varepsilon ||v|| \quad \text{and} \quad ||u - w|| = \varepsilon.
\]

From the definition of the subgradient,

\[
  f(u) - f(w) \geq \langle v, u - w \rangle = \varepsilon ||v||.
\]

On the other hand, from the Lipschitzness of \( f \) we have

\[
  \rho \varepsilon = \rho ||u - w|| \geq f(u) - f(w).
\]

Combining the two inequalities we conclude that \( ||v|| \leq \rho \). □
Subgradient Descent

• The gradient descent algorithm can be generalized to non-differentiable functions by using a subgradient of $f(w)$ at $w^{(t)}$, instead of the gradient.
Stochastic Gradient Descent (SGD)

• In stochastic gradient descent we do not require the update direction to be based exactly on the gradient. Instead, we allow the direction to be a random vector and only require that its expected value at each iteration will equal the gradient direction.

• Or, more generally, we require that the expected value of the random vector will be a subgradient of the function at the current vector.
Stochastic Gradient Descent (SGD) for minimizing $f(w)$

parameters: Scalar $\eta > 0$, integer $T > 0$
initialize: $w^{(1)} = 0$
for $t = 1, 2, \ldots, T$
  choose $v_t$ at random from a distribution such that $E[v_t \mid w^{(t)}] \in \partial f(w^{(t)})$
  update $w^{(t+1)} = w^{(t)} - \eta v_t$
output $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$
Figure 14.3 An illustration of the gradient descent algorithm (left) and the stochastic gradient descent algorithm (right). The function to be minimized is $1.25(x + 6)^2 + (y - 8)^2$. For the stochastic case, the black line depicts the averaged value of $\mathbf{w}$. 
THEOREM 14.8 Let $B, \rho > 0$. Let $f$ be a convex function and let $\mathbf{w}^* \in \text{argmin}_{\mathbf{w} : \|\mathbf{w}\| \leq B} f(\mathbf{w})$. Assume that SGD is run for $T$ iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume also that for all $t$, $\|\mathbf{v}_t\| \leq \rho$ with probability 1. Then,

$$\mathbb{E} [f(\mathbf{w})] - f(\mathbf{w}^*) \leq \frac{B \rho}{\sqrt{T}}.$$ 

Therefore, for any $\epsilon > 0$, to achieve $\mathbb{E} [f(\mathbf{w})] - f(\mathbf{w}^*) \leq \epsilon$, it suffices to run the SGD algorithm for a number of iterations that satisfies

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}.$$
Learning with SGD

Stochastic Gradient Descent (SGD) for minimizing $L_D(w)$

parameters: Scalar $\eta > 0$, integer $T > 0$
initialize: $w^{(1)} = 0$
for $t = 1, 2, \ldots, T$
    sample $z \sim \mathcal{D}$
    pick $v_t \in \partial \ell(w^{(t)}, z)$
    update $w^{(t+1)} = w^{(t)} - \eta v_t$
output $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}$
Corollary 14.12  Consider a convex-Lipschitz-bounded learning problem with parameters $\rho, B$. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_D(w)$ with a number of iterations (i.e., number of examples)

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

and with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output of SGD satisfies

$$\mathbb{E} [L_D(\bar{w})] \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon.$$
COROLLARY 14.14 Consider a convex-smooth-bounded learning problem with parameters $\beta, B$. Assume in addition that $\ell(0, z) \leq 1$ for all $z \in Z$. For every $\epsilon > 0$, set $\eta = \frac{1}{\beta(1+3/\epsilon)}$. Then, running SGD with $T \geq 12B^2\beta/\epsilon^2$ yields

$$\mathbb{E}[L_D(\mathbf{w})] \leq \min_{\mathbf{w} \in \mathcal{H}} L_D(\mathbf{w}) + \epsilon.$$