Foundations for Machine Learning

L. Y. Stefanus

TU Dresden, June-July 2018

Slides 09

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 $\nabla f(\mathrm{w})$, is the vector of partial derivatives of f, namely,

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

- Gradient descent is an iterative algorithm. We start with an initial value of w (say, $\mathrm{w}^{(1)} = 0$). 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

$$\mathrm{w}^{(t+1)} = \mathrm{w}^{(t)} - \eta \nabla f\left(\mathrm{w}^{(t)}\right), \qquad (14.1)$$

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, $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{w}^{(t)}$, the last vector, $\mathrm{w}^{(T)}$, or the best performing vector, $\mathrm{argmin}_{t\in[T]} f(\mathrm{w}^{(t)})$. 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, ρ-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 $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output vector $\bar{w}$ satisfies

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

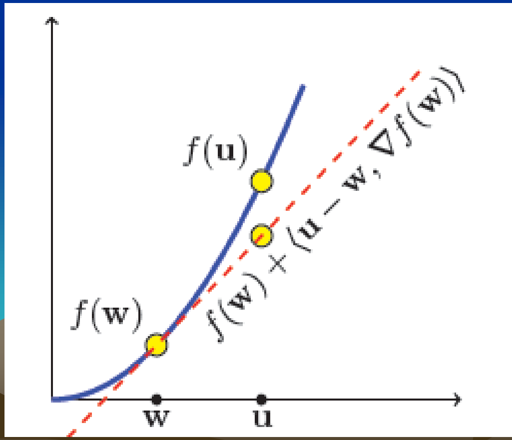
Furthermore, for every $\epsilon > 0$, to achieve $f(\bar{w}) - f(w^*) \leq \epsilon$, it suffices to run the GD algorithm for a number of iterations that satisfies

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

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,

$$\forall \mathrm{u}, f(\mathrm{u}) \geq f(\mathrm{w}) + < \mathrm{u} - \mathrm{w}, \nabla f(\mathrm{w}) >.$$



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.

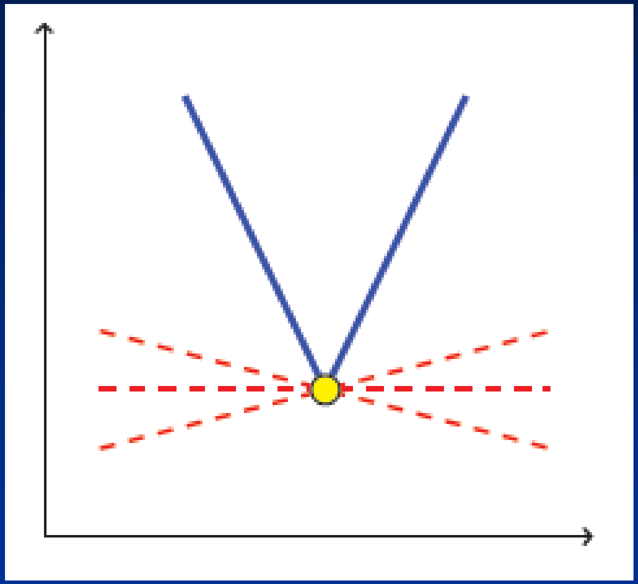
LEMMA 14.3 *Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff for every $\mathbf{w} \in S$ there exists $\mathbf{v}$ such that*

$$\forall \mathbf{u} \in S, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle. \tag{14.8}$$

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 ∂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 $f: \mathrm{A} \rightarrow \mathbb{R}$ is ρ-Lipschitz if for all $u, v \in$ A:

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

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

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

Proof Assume that for all $\mathbf{v} \in \partial f(\mathbf{w})$ we have that $\|\mathbf{v}\| \leq \rho$. Since $\mathbf{v} \in \partial f(\mathbf{w})$ we have

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle.$$

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

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle \leq \|\mathbf{v}\| \, \|\mathbf{w} - \mathbf{u}\| \leq \rho \, \|\mathbf{w} - \mathbf{u}\|.$$

An analogous argument can show that $f(\mathbf{u}) - f(\mathbf{w}) \leq \rho \, \|\mathbf{w} - \mathbf{u}\|$. Hence f is ρ-Lipschitz.

Now assume that f is ρ-Lipschitz. Choose some $\mathbf{w} \in A, \mathbf{v} \in \partial f(\mathbf{w})$. Since A is open, there exists $\epsilon > 0$ such that $\mathbf{u} = \mathbf{w} + \epsilon \mathbf{v}/\|\mathbf{v}\|$ belongs to A. Therefore, $\langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \epsilon \|\mathbf{v}\|$ and $\|\mathbf{u} - \mathbf{w}\| = \epsilon$. From the definition of the subgradient,

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle = \epsilon \|\mathbf{v}\|.$$

On the other hand, from the Lipschitzness of f we have

$$\rho \, \epsilon = \rho \, \|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w}).$$

Combining the two inequalities we conclude that $\|\mathbf{v}\| \leq \rho$. □

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

$$f(\mathbf{w})$$

parameters: Scalar $\eta > 0$, integer $T > 0$

initialize: $\mathbf{w}^{(1)} = \mathbf{0}$

for $t = 1, 2, \ldots, T$

choose $\mathbf{v}_t$ at random from a distribution such that $\mathbb{E}[\mathbf{v}_t \,|\, \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$

update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$

output $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$

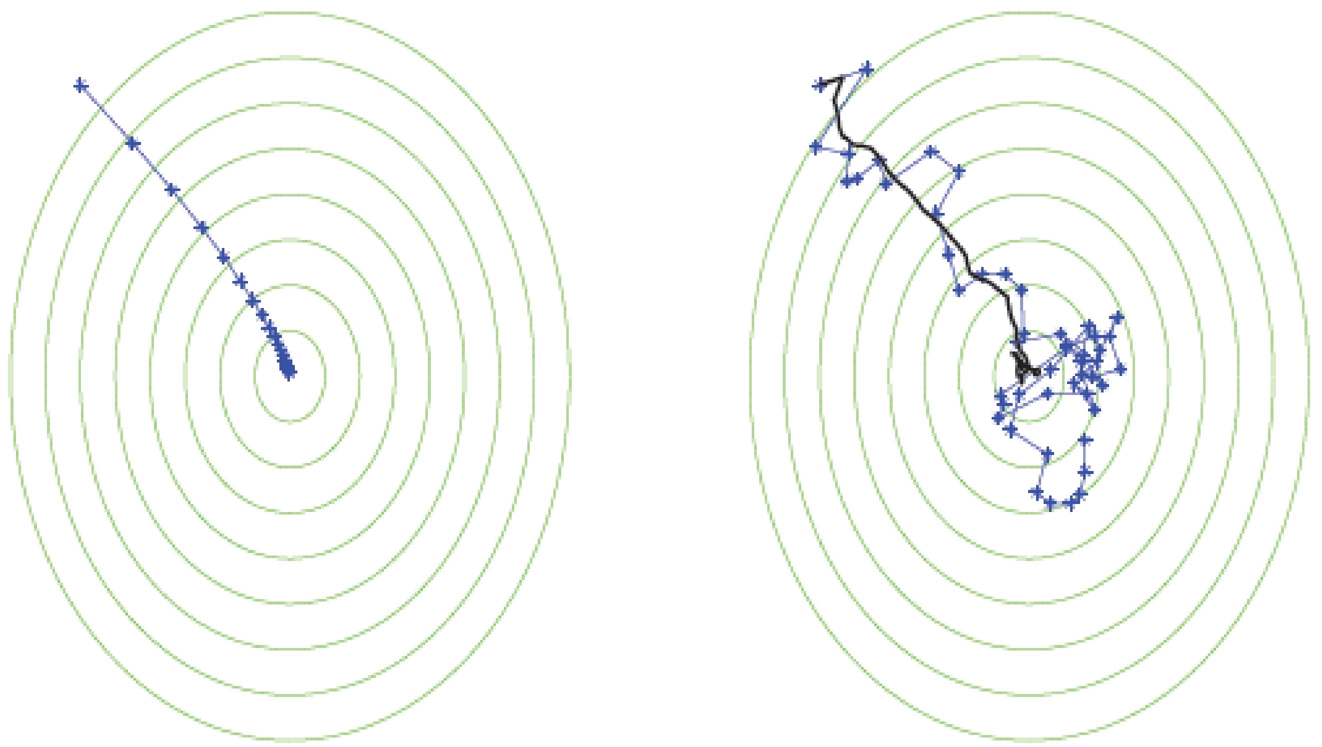


Figure 14.3 An illustration of the gradient descent algorithm (left) and the stochastic gradient descent algorithm (right). The function to be minimized is $1.25(x+6)^2 + (y-8)^2$. For the stochastic case, the black line depicts the averaged value of $\mathbf{w}$.

Bound on the expected output of stochastic gradient descent

THEOREM 14.8 *Let $B, \rho > 0$. Let f be a convex function and let $\mathbf{w}^\star \in \text{argmin}_{\mathbf{w}:\|\mathbf{w}\| \le B} f(\mathbf{w})$. Assume that SGD is run for T iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume also that for all t, $\|\mathbf{v}_t\| \le \rho$ with probability 1. Then,*

$$\mathbb{E}\left[f(\bar{\mathbf{w}})\right] - f(\mathbf{w}^\star) \le \frac{B\,\rho}{\sqrt{T}}.$$

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

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

Learning with SGD

Stochastic Gradient Descent (SGD) for minimizing $L_{\mathcal{D}}(\mathbf{w})$

parameters: Scalar $\eta > 0$, integer $T > 0$

initialize: $\mathbf{w}^{(1)} = \mathbf{0}$

for $t = 1, 2, \ldots, T$

  sample $z \sim \mathcal{D}$

  pick $\mathbf{v}_t \in \partial \ell(\mathbf{w}^{(t)}, z)$

  update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$

output $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$

COROLLARY 14.12 *Consider a convex-Lipschitz-bounded learning problem with parameters ρ, B. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_{\mathcal{D}}(\mathbf{w})$ 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

$$\mathbb{E}\left[L_{\mathcal{D}}(\bar{\mathbf{w}})\right] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

COROLLARY 14.14 *Consider a convex-smooth-bounded learning problem with parameters* β, B. *Assume in addition that* $\ell(\mathbf{0}, z) \leq 1$ *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*

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