14 Stochastic Gradient Descent

In stochastic gradient descent (SGD) we try to minimize the risk function $L_D(w)$ directly using a gradient descent procedure. Gradient descent is an iterative optimization procedure where at each step the solution is improved by taking a step along the negative of the gradient. But, the function to optimize is not known since the true distribution is unknown. SGD circumvents this by taking a step in a random direction as long as the expected value of the direction is the negative of the gradient. SGD is an efficient algorithm that has the same sample complexity as regularized risk minimization.

14.1 Gradient descent

The gradient of a differentiable function $f(w)$ is denoted as $\nabla f(w)$. It is a vector of partial derivatives (one derivative per dimension). Gradient descent is an iterative algorithm: We start with some value for w. Then for each subsequent iteration we move by

$$w^{t+1} = w^t - \eta \nabla f(w^t)$$

The gradient points in the direction of the greatest increase around w for function f, we move in the opposite direction to decrease the value of the function (we wish to decrease the value of the loss). After T iterations, the algorithm outputs the averaged vector $w = \frac{1}{T}\sum_{t=1}^{T} w^t$. The output can also be the last weight, or the best one, but the average is useful when dealing with nondifferentiable functions and deal with the stochastic case.

Alternatively, we can use the Taylor approximation of f around w: $f(u) = f(w) + \langle u - w, \nabla f(w)\rangle$ Having f convex, that approximation has a lower bound: $f(u) \geq f(w) + \langle u - w, \nabla f(w)\rangle$. This implies we can minimize the approximation of $f(w)$, but since the distance between w and w_t might become large we will jointly minimize the distance between w_t and w and the approximation $f(w)$.

$$w^{t+1} = arg\min_w \frac{1}{2}||w - w^t||^2 + \eta(f(w^t) + \langle w - w^t, \nabla f(w^t)\rangle)$$

Deriving expression by w and introducing $w = w_{t+1}$ we arrive to the update rule.

14.1.1 Analysis of GD for Convex-Lipschitz Functions

Let w^* be any vector and let B be an upper bound on $||w^*||$. We wish to upper bound suboptimality of w^*. Now we derive the difference $f(w) - f(w^*)$, arriving to

$$f(w) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T}\langle w^t - w^*, \nabla f(w^t)\rangle$$

$$f(w) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(w^t) - f(w^*))$$

The function is convex, therefore

$$f(w^t) - f(w^*) \leq \langle w^t - w^*, \nabla f(w^t)\rangle$$

Lemma 14.1 Let $v_1, ..., v_T$ be a sequence of vectors. Any algorithm with an initialization $w_1 = 0$ and the GD update rule satisfies

$$\sum_{t=1}^{T}\langle w_t - w^*, v_t\rangle \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}||v_t||^2$$

This means that the suboptimality of the solution is upper bounded by the norm B, norm ρ and inversely proportional to the number of averaged steps T.

14.2. Subgradients

A subgradient is any vector v that satisfies the convexity property where the vector is placed instead of the derivation of the function:

$$f(u) \geq f(w) + \langle u - w, v\rangle$$

The set of subgradients of f at w is called the differential set of subgradients denoted as $\partial f(w)$.

14.2.1 Calculating subgradients

Claim 14.5 If a claim is differentiable at w then $\partial f(w)$ contains a single element, the gradient of f at $w, \nabla f(w)$.

The absolute function has subgradients 1 for $x > 0$ and -1 for $x < 0$, whereas for $x = 0$ its subgradients are $[-1, 1]$.

Claim 14.6 Let $g(w) = \max_{i\in[r]} g_i(w)$ for r convex differentiable functions. Given some w, let $j \in argmax_i g_i(w)$. Then $\nabla g_j(w) \in \partial g(w)$.

14.2.2 Subgradients of Lipschitz Functions

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 $w \in A$ and $v \in \partial f(w)$ we have that $||v|| \leq \rho$.

14.2.3 Subgradient Descent

The gradient descent can be generalized to nondifferentiable functions by using subgradient of $f(w)$ at w instead of the gradient. The convergence rate stays the same since equation 14.3 is still valid.

14.3 Stochastic gradient descent (SGD)

In SGD we allow the direction to vary (instead of it being the negative gradient) as long as the expectation is in the negative gradient. The expected value of the random vector is a subgradient of the function: $E[v_t|w^t] \in \partial f(w^t)$

14.3.1 Analysis of SGD for Convex-Lipchitz-Bounded Functions

Only the expectation of $v_t \in \partial f(w)$, therefore equation 14.3 cannot be applied. But, since the expected value of v_t is a subgradient at $f(w)$ a similar bound can be derived based on the expected output of gradient descent.

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

To derive the proof, we use that
w_t depends on merely v_{t-1} according to the update rule, therefore expectation is constant

14.4 Variants

14.4.1 Adding a projection step

To ensure that weights modified during gradient descent always end up within the bounds, that is each w is $||w|| \leq B$, we add a projection step to the GD algorithm. Thus we have a two step update rule:

$$w^{t+\frac{1}{2}} = w^t - \eta v_t$$
$$w^{t+1} = arg\min_{w \in \mathbb{H}} ||w - w^{t+\frac{1}{2}}||$$

This projection step replaces the gradiently descent weight with the nearest one in the allowed space.

14.5 Learning with SGD

14.5.1 SGD for Risk Minimization

We wish to minimize the risk function $L_D(w) = \mathbb{E}_{z\sim D}[l(w,z)]$ SGD minimizes $L_D(w)$ directly instead of minimizing $L_S(w)$. With SGD we find an unbiased estimate of the gradient $\nabla L_D(w)$. First, we consider differntiable functions. First, we sample $z \sim D$. Then we define random vector v_t to be the gradient of the function $l(w,z)$ with respect w, at the point w^t. Then by linearity of the gradient

$$\mathbb{E}[v_t|w_t] = \mathbb{E}_{z\sim D}[\nabla l(w^t, z)] = \nabla L_D(w)$$

The same argument holds for nondifferentiable loss functions. v_t is a subgradient of $l(w,z)$ at w^t. Then for every u

$$l(u,z) - l(w^t, z) \geq \langle u - w^t, v_t \rangle$$