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 $f(w)$.

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

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

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