Gradient Descent

The gradient of a differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ at $\vec{w} \in \mathbb{R}^d$, denoted $\nabla f(\vec{w})$, is the vector of partial derivatives of f, namely,

$$\nabla f(\vec{w}) = \left(\frac{\partial f(\vec{w})}{\partial w_1} \quad \cdots \quad \frac{\partial f(\vec{w})}{\partial w_d} \right).$$

Gradient descent is an iterative algorithm.

We start with an initial value of $\vec{w}$, then at each iteration we take a step in the direction of the negative of the gradient at the current point.

That is,

UPDATE RULE

$$w_{t+1} = w_t - \eta \nabla f(w_t) \qquad (\eta > 0).$$

Eventually, after K iterations, algorithm outputs the averaged vector

$$\bar{w} = \frac{1}{K} \sum_{t=1}^{K} w_t$$

Note The output could also be the last vector $\vec{w}_K$ or the best performing vector $\underset{t \in [K]}{\operatorname{argmin}} f(w_t)$, but avg tends to be more useful in extensions.

Another motivation for GD is by relying on Taylor approximation.

The gradient of f at $\vec{w}$ yields the first order Taylor approximation of f around $\vec{w}$ by

$$f(\vec{u}) \approx f(\vec{w}) + \langle \vec{u} - \vec{w}, \nabla f(\vec{w}) \rangle$$

If f is convex, this approximation lower bounds f, that is

$$f(\vec{u}) \geq f(\vec{w}) + \langle \vec{u} - \vec{w}, \nabla f(\vec{w}) \rangle$$

Therefore, for $\vec{w}$ close to $\vec{w}_t$ we have that

$$f(\vec{w}) \approx f(\vec{w}_t) + \langle \vec{w} - \vec{w}_t, \nabla f(\vec{w}_t) \rangle$$

Hence we can minimize the approximation of $f(\vec{w})$.

However, the approximation might become loose for $\vec{w}$, which is far away from $\vec{w}_t$.

Therefore, we would like to minimize jointly the distance between $\vec{w}$ and $\vec{w}_t$ and the approximation of f around $\vec{w}_t$.

If the parameter η controls the tradeoff between the two terms, we obtain the update rule

$$\vec{w}_{t+1} = \operatorname{argmin}_{\vec{w}} \frac{1}{2} \|\vec{w} - \vec{w}_t\| + \eta \left(f(w_t) + \langle \vec{w} - \vec{w}_t, \nabla f(\vec{w}_t) \rangle \right)$$

Solving by taking the derivative with respect to $\vec{w}$ and comparing it to zero yields the same update rule as above!

Analysis of GD for Convex-Lipschitz Functions

Consider the case of f being a convex-Lipschitz function. Let $\vec{w}^*$ be any vector and let B be an upper bound on $\|\vec{w}^*\|$.

Note It is convenient to think of $\vec{w}^*$ as the minimizer of $f(\vec{w})$, but the analysis that follows holds $\forall\, \vec{w}^*$

We want an upper bound on the suboptimality of our solution with respect to $\vec{w}^*$, namely, $f(\bar{w}) - f(\vec{w}^*)$, where $\bar{w} = \frac{1}{K}\sum_{t=1}^{K} w_t$.

Aside: Jensen's Inequality

$$f(tx_1 + (1-t)x_2) \leq t f(x_1) + (1-t) f(x_2)$$

From the definition of $\bar{w}$, and using Jensen's inequality, we have

$$\begin{aligned} f(\bar{w}) - f(\vec{w}^*) &= f\left(\frac{1}{K}\sum_{t=1}^{K} w_t\right) - f(\vec{w}^*) \\ &\leq \frac{1}{K}\sum_{t=1}^{K} f(w_t) - f(\vec{w}^*) \\ &= \frac{1}{K}\sum_{t=1}^{K}\left(f(w_t) - f(\vec{w}^*)\right) \end{aligned}$$

Due to the convexity f, we have

$$f(\vec{w}_t) - f(\vec{w}^*) \leq \langle \vec{w}_t - \vec{w}^* , \nabla f(\vec{w}_t) \rangle \qquad \forall t$$

we thus have

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

We now bound the right-hand side using the following lemma.

Lemma: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an arbitrary sequence of vectors. Any algorithm with an initialization $\vec{w}_1 = 0$ and an update rule of the form

$$\vec{w}_{t+1} = \vec{w}_t - \eta \vec{v}_t$$

satisfies

$$\sum_{t=1}^{K} \langle \vec{w}_t - w^*, \vec{v}_t \rangle \leq \frac{\|\vec{w}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{K} \|\vec{v}_t\|^2$$

In particular, $\forall B, \rho > 0$, if $\forall t$ we have that $\|\vec{v}_t\|^2 \leq \rho$ and if we set

$$\eta = \sqrt{\frac{B^2}{\rho^2 K}}$$

then for every $\vec{w}^*$, with $\|\vec{w}^*\| \leq B$ we have

$$\frac{1}{K} \sum_{t=1}^{K} \langle \vec{w}_t - w^*, \vec{v}_t \rangle \leq \frac{B\rho}{\sqrt{K}}$$

Proof

Completing the square, we get

$$\langle \vec{w}_t - \vec{w}^*, \vec{v}_t \rangle = \frac{1}{\eta} \langle \vec{w}_t - \vec{w}^*, \eta \vec{v}_t \rangle$$

$$= \frac{1}{2\eta} \left(-\|\vec{w}_t - \vec{w}^* - \eta \vec{v}_t\|^2 + \|\vec{w}_t - \vec{w}^*\|^2 + \eta^2 \|\vec{v}_t\|^2 \right)$$

$$= \frac{1}{2\eta} \left(-\|\vec{w}_{t+1} - \vec{w}^*\|^2 + \|\vec{w}_t - \vec{w}^*\|^2 \right) + \frac{\eta}{2} \|\vec{v}_t\|^2$$

Summing the equality over t, we have

$$\sum_{t=1}^{T} \langle \vec{w}_t - \vec{w}^*, \vec{v}_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left(-\|\vec{w}_{t+1} - \vec{w}^*\|^2 + \|\vec{w}_t - \vec{w}^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\vec{v}_t\|^2$$

First sum is a telescopic sum which collapses to

$$\|\vec{w}_1 - \vec{w}^*\|^2 - \|\vec{w}_{T+1} - \vec{w}^*\|^2$$

Plugging this in yiels

$$\sum_{t=1}^{T} \langle \vec{w}_t - \vec{w}^*, \vec{v}_t \rangle = \frac{1}{2\eta} \left(\|\vec{w}_1 - \vec{w}^*\|^2 - \|\vec{w}_{T+1} - \vec{w}^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\vec{v}_t\|^2$$

$$\leq \frac{1}{2\eta} \left(\|\vec{w}_1 - \vec{w}^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\vec{v}_t\|^2$$

$$= \frac{1}{2\eta} \|w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\vec{v}_t\|^2 \qquad (\vec{w}_1 = 0 \text{ by assumption})$$

Bounding $\|\vec{w}^*\|^2$ by B, $\|\vec{v}_t\|$ by ρ, dividing by T, and plugging in for value of η completes second part of lemma.

Q.E.D

The previous lemma applies to GD algorithm w/ $\vec{v}_t = \nabla f(\vec{w}_t)$. If f is ρ-Lipschitz, then $\|\nabla f(\vec{w}_t)\| \leq \rho$. We therefore satisfy the lemma's conditions and achieve the following corollary:

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

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

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

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