Stochastic Gradient Descent

Stefanos Baros

1 Introduction

One of the advantages of Stochastic Gradient Descent (SGD) for machine learning is that it allows us to minimize the risk function $L_D(w)$ directly using a gradient procedure without knowing its gradient. Thus, it allows to go beyond standard empirical risk minimization (ERM) where we only deal with the empirical error $L_S(h)$. In general, in stochastic gradient descent, we take a step along a random direction as long as the expected value of the direction is the negative of the gradient. SGD is efficient and easy to implement, and it enjoys the same sample complexity as the Regularized Loss Minimization (RLM).

2 Gradient Descent

2.1 Method

We consider a function $f : \mathbb{R}^d \to \mathbb{R}$ with gradient $\nabla f(w) = (\frac{\partial f(w)}{\partial w[1]}, ..., \frac{\partial f(w)}{\partial w[d]})$. Then, we initialize $w^{(1)} = 0$ and update w according to the formula:

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

where $\eta > 0$. After T iterations, we output the average of the $w^{(t)}$ values specifically:

$$\overline{w} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)}. \tag{2}$$

In general, the output could also be $w^{(T)}$, the last vector. When T is large, $\overline{w} \approx w^{\star}$.

2.2 Convergence of Gradient Descent

In this section, we prove that gradient descent converges i.e., that $f(\overline{w}) - f(w^{\star})$ is bounded. To do that, we use two properties of the f function. The first one, is that f is convex and the second one that is ρ-Lipschitz. We begin by stating the main theorem related to the convergence of GD.

Theorem 1 ([1]). *Let f be a convex, ρ-Lipschitz function and let $w^\star \in \operatorname{argmin}_{w:\|w\|\leq B} f(w)$. If we run the GD algorithm on f for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output vector $\overline{w}$ satisfies:*

$$f(\overline{w}) - f(w^\star) \leq \frac{B\rho}{\sqrt{T}} \tag{3}$$

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

$$T \geq \frac{B^2\rho^2}{\varepsilon^2}. \tag{4}$$

Proof. We start by writing:

$$f(\overline{w}) - f(w^\star) = f(\frac{1}{T}\sum_{t=1}^{T} w^{(t)}) - f(w^\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^\star), \tag{5}$$

where we used *Jensen's inequality* to obtain the last inequality. Thus, we have shown that:

$$f(\overline{w}) - f(w^\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^\star). \tag{6}$$

Since f is *convex* it holds:

$$f(w^{(t)}) - f(w^\star) \leq \left\langle w^{(t)} - \omega^\star, \nabla f(w^{(t)}) \right\rangle. \tag{7}$$

Using this inequality, we further obtain from (6):

$$f(\overline{w}) - f(w^\star) \leq \frac{1}{T}\sum_{t=1}^{T} \left\langle w^{(t)} - \omega^\star, \nabla f(w^{(t)}) \right\rangle. \tag{8}$$

We will now bound the right-hand side of (8).

$$\begin{aligned}\frac{1}{\eta}\left\langle w^{(t)} - \omega^\star, \eta\nabla f(w^{(t)}) \right\rangle &= -\frac{1}{2\eta}\|w^{(t)} - \omega^\star - \eta\nabla f(w^{(t)})\|^2 + \frac{1}{2\eta}\|w^{(t)} - \omega^\star\|^2 \\ &\quad + \frac{\eta}{2}\|\nabla f(w^{(t)})\|^2 \\ &= \frac{1}{2\eta}\|w^{(t+1)} - \omega^\star\|^2 + \frac{1}{2\eta}\|w^{(t)} - \omega^\star\|^2 + \frac{\eta}{2}\|\nabla f(w^{(t)})\|^2\end{aligned} \tag{9}$$

Taking the sum yields:

$$\sum_{t=1}^{T}\left\langle w^{(t)} - \omega^\star, \nabla f(w^{(t)}) \right\rangle = \frac{1}{2\eta}\|w^{(1)} - \omega^\star\|^2 - \frac{1}{2\eta}\|w^{(T+1)} - \omega^\star\|^2 + \frac{\eta}{2}\|\nabla f(w^{(t)})\|^2. \tag{10}$$

By taking into account that $w^{(1)} = 0$ and that f is ρ-Lipschitz i.e., that $\|\nabla f(w^{(t)})\| \leq \rho$, we finally obtain:

$$\sum_{t=1}^{T} \left\langle w^{(t)} - \omega^\star, \nabla f(w^{(t)}) \right\rangle \leq \frac{1}{2\eta}\|\omega^\star\|^2 + \frac{\eta}{2}\rho^2 T. \tag{11}$$

From (8), we can finally obtain:

$$f(\overline{w}) - f(w^\star) \leq \frac{1}{2\eta T}\|\omega^\star\|^2 + \frac{\eta}{2}\rho^2. \tag{12}$$

Letting $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ yields:

$$f(\overline{w}) - f(w^\star) \leq \frac{B\rho}{\sqrt{T}}. \tag{13}$$

The second statement follows by letting the right-hand side being less than ε and solving for T. □

3 Subgradients

Gradient descent can be applied to nondifferentiable functions as well by simply using the subgradient of $f(w)$ at $w^{(t)}$ in place of the gradient. We provide the definition of subgradient below. First, recall that, for a *convex* function it holds:

$$\forall u, \quad f(u) \geq f(w) + \langle u - w, \nabla f(w) \rangle . \tag{14}$$

which means that the tangent at any point w lies below f. This inequality can be generalized in the case of subgradients as follows:

Lemma 1 ([1]). *Let S be an open convex set. A function $f : S \to R$ is convex if and only if for every $w \in S$ there exists v such that:*

$$\forall u, \quad f(u) \geq f(w) + \langle u - w, v \rangle . \tag{15}$$

Therefore, f does not have to be differentiable to be convex. As long as a line exists that touches the function at a point w and is not above the function elsewhere f will be convex. The slope of this line is the subgradient. Consider now the following definition of a subgradient.

Definition 1 ([1]). A vector v that satisfies (15) is called a subgradient of f at w. The set of subgradients of f at w is called the differential set and denoted by $\partial f(w)$.

The following claim holds for differentiable functions.

Claim 1 ([1]). *If f is differentiable at w then $\partial f(w)$ contains a single element—the gradient of f at w, $\nabla f(w)$.*

An interesting result holds for functions defined as the maximum of other functions.

Claim 2. *Let $g(w) = \max_{i\in[r]} g_i(w)$ for r convex differentiable functions $g_1, .., g_r$. Given some w let $j \in \operatorname{argmax}_i g_i(w)$. Then, $\nabla g_j(w) \in \partial g(w)$.*

We now go through an example to illustrate where this result can be useful.

Example-Hinge loss. We can use this result to compute the subgradient of the hinge loss function $f(w) = \max\{0, 1 - y\langle w, x\rangle\}$. First, it is important to notice that:

$$1 - y\langle w, x\rangle \leq 0, \quad f(w) = g_1(w) = 0, \tag{16}$$
$$1 - y\langle w, x\rangle \geq 0, \quad f(w) = g_2(w) = 1 - y\langle w, x\rangle . \tag{17}$$

Given that, and by employing the above claim, we can easily compute the *subgradient* v as:

$$v = \begin{cases} \nabla g_1(w) = 0, & 1 - y\langle w, x\rangle \leq 0, \\ \nabla g_2(w) = 0, & 1 - y\langle w, x\rangle \geq 0. \end{cases} \tag{18}$$

Another interesting result says that a function is Lipschitz if its subgradient is bounded. More formally:

Lemma 2. *[1] 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 if and only if for all $w \in A$ and $v \in \partial f(w)$ we have that $\|v\| \leq \rho$.*

Next, we introduce the *Stochastic Gradient Descent (SGD)* algorithm.

4 Stochastic Gradient Descent

4.1 Method

As we mentioned before, the update direction in SGD is not required to be exactly the gradient but can be any random vector whose expected value at each iteration matches the gradient direction. The *Stochastic Gradient Descent method* is described below.

- Parameters: scalar $\eta > 0$, integer $T > 0$.
- Initialize: $w^{(1)} = 0$.
- For $t = 1, 2, ..., T$
  - Choose v_t at random from a distribution such that $\mathbb{E}[v_t|w^{(t)}] \in \partial f(w^{(t)})$

– Update $w^{(t+1)} = w^{(t)} - \eta v_t$

Output: $\overline{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$

Next, we state the main result related to the convergence of SGD together with its proof.

4.2 Convergence of Stochastic Gradient Descent

We now state the main convergence result related to SGD.

Theorem 2 ([1]). *Let $B, \rho > 0$, f be a convex function and let $w^\star \in \operatorname{argmin}_{w:\|w\|\leq B} f(w)$. Assume that we run the SGD algorithm on f for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume also that, for all t, $\|v_t\| \leq \rho$ with probability 1. Then:*

$$\mathbb{E}[f(\overline{w})] - f(w^\star) \leq \frac{B\rho}{\sqrt{T}}. \tag{19}$$

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

$$T \geq \frac{B^2\rho^2}{\varepsilon^2}. \tag{20}$$

Proof. As in the proof of GD, the inequality:

$$f(\overline{w}) - f(w^\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^\star), \tag{21}$$

holds. Taking expectation w.r.t. random vectors $v_1, ..., v_T$ yields:

$$\mathbb{E}_{v_{1:T}}[f(\overline{w}) - f(w^\star)] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{v_{1:T}}[f(w^{(t)}) - f(w^\star)]. \tag{22}$$

The right-hand side follows from the linearity of expectation $\mathbb{E}[\sum_{t=1}^{T}\Phi(t)] = \sum_{t=1}^{T}\mathbb{E}[\Phi(t)]$. On the other hand, the bound that we proved for GD also holds here but instead of $\nabla f(w^{(t)})$ we have v_t:

$$\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)} - w^\star, v_t\right\rangle \leq \frac{B\rho}{\sqrt{T}}. \tag{23}$$

Taking expectation here leads to:

$$\mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)} - w^\star, v_t\right\rangle\right] \leq \frac{B\rho}{\sqrt{T}}. \tag{24}$$

Thus, we only need to show that:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{v_{1:T}}[f(w^{(t)})-f(w^{\star})] \leq \mathbb{E}_{v_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)}-w^{\star}, v_t\right\rangle]. \tag{25}$$

The key difference though between the proof of SGD and the proof of GD is that here we cannot use the convexity property of f as before because:

$$f(w^{(t)})-f(w^{\star}) \leq \left\langle w^{(t)}-w^{\star}, v_t\right\rangle, \tag{26}$$

does not necessarily hold as v_t is *not* the gradient/subgradient of f, its expected value is! Thus, it only holds:

$$f(w^{(t)})-f(w^{\star}) \leq \left\langle w^{(t)}-w^{\star}, \mathbb{E}[v_t]\right\rangle. \tag{27}$$

To proceed, we consider the law of total expectation that says that $\mathbb{E}_a[g(a)] = \mathbb{E}_\beta \mathbb{E}_a[g(a)|\beta]$. This allows us to obtain:

$$\mathbb{E}_{v_{1:t}}\left[\left\langle w^{(t)}-w^{\star}, v_t\right\rangle\right] = \mathbb{E}_{v_{1:t-1}}\mathbb{E}_{v_{1:t}}\left[\left\langle w^{(t)}-w^{\star}, v_t\right\rangle|v_{t-1}\right]. \tag{28}$$

But $w^{(t)}$ is known when v_t is known so:

$$\begin{aligned}\mathbb{E}_{v_{1:t}}\left[\left\langle w^{(t)}-w^{\star}, v_t\right\rangle\right] &= \mathbb{E}_{v_{1:t-1}}\left[\left\langle w^{(t)}-w^{\star}, \mathbb{E}_{v_{1:t}}[v_t|v_{t-1}\right\rangle\right] \\ &= \mathbb{E}_{v_{1:t-1}}\left[\left\langle w^{(t)}-w^{\star}, \nabla f(w^{(t)})\right\rangle\right].\end{aligned} \tag{29}$$

Using the convexity property of f we obtain that:

$$\mathbb{E}_{v_{1:t}}\left[\left\langle w^{(t)}-w^{\star}, v_t\right\rangle\right] \geq \mathbb{E}_{v_{1:t-1}}[f(w^{(t)}-f(w^{\star}))] = \mathbb{E}_{v_{1:T}}[f(w^{(t)}-f(w^{\star}))]. \tag{30}$$

The right-hand side equality follows as the extra terms in expectation $v_t, ..., v_T$ do not affect $w^{(t)}$. Finally we that we obtain:

$$\mathbb{E}_{v_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)}-w^{\star}, v_t\right\rangle] \geq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{v_{1:T}}[f(w^{(t)})-f(w^{\star})]. \tag{31}$$

which is the inequality (25) we wanted to prove. □

5 Variants

In GD and SGD we require that $w^{\star} \in H = \{w : \|w\| \leq B\}$. However, GD or SGD may force w to step out of this bound during iterations. A way to

guarantee that this will not happen, is to use a projection operator. This leads to the *modified SGD* update rule:

$$w^{(t+\frac{1}{2})} = w^{(t)} - \eta v_t, \tag{32}$$

$$w^{(t+1)} = \underset{w \in H}{\operatorname{argmin}} \|w - w^{(t+\frac{1}{2})}\|. \tag{33}$$

Basically, the second equality ensures that $w^{(t+1)} \in H$ as it is the projection of $w^{(t+\frac{1}{2})}$ on H. The most important property of the projection operator that allows one to extend the convergence proof of the SGD to this case is its *non-expansiveness*. Simply put, a projection operator P is *nonexpansive* if, for any two points z_2, z_1 holds:

$$\|P(z_2) - P(z_1)\| \leq \|z_2 - z_1\|. \tag{34}$$

5.1 Convergence of SGD with Projection

The inequalities (22) and (25) that we proved for SGD are still valid here. We only need to show that, when we use projection, the inequality (24) still holds. To do that, we use (33) to get:

$$\frac{1}{\eta}\left\langle w^{(t)} - \omega^\star, \eta v_t)\right\rangle = -\frac{1}{2\eta}\|w^{(t)} - \omega^\star - \eta v_t\|^2 + \frac{1}{2\eta}\|w^{(t)} - \omega^\star\|^2 + \frac{\eta}{2}\|v_t)\|^2 \tag{35}$$

$$= \frac{1}{2\eta}\|w^{(t+\frac{1}{2})} - \omega^\star\|^2 + \frac{1}{2\eta}\|w^{(t)} - \omega^\star\|^2 + \frac{\eta}{2}\|v_t\|^2. \tag{36}$$

Using the *nonexpansiveness* property of the projection operator we get:

$$\|w^{(t+1)} - w^\star\|^2 \leq \|w^{(t+\frac{1}{2})} - w^\star\|^2. \tag{37}$$

Equivalently:

$$-\|w^{(t+\frac{1}{2})} - w^\star\|^2 \leq -\|w^{(t+1)} - w^\star\|^2. \tag{38}$$

With this, we finally obtain

$$\frac{1}{\eta}\left\langle w^{(t)} - \omega^\star, \eta v_t)\right\rangle \leq \frac{1}{2\eta}\|w^{(t+1)} - \omega^\star\|^2 + \frac{1}{2\eta}\|w^{(t)} - \omega^\star\|^2 + \frac{\eta}{2}\|v_t\|^2. \tag{39}$$

As in the proof of GD, this allows us to show that:

$$\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)} - w^\star, v_t\right\rangle \leq \frac{B\rho}{\sqrt{T}}. \tag{40}$$

Taking expectation here leads to:

$$\mathbb{E}_{v_{1:T}}[\frac{1}{T}\sum_{t=1}^{T}\left\langle w^{(t)} - w^\star, v_t\right\rangle] \leq \frac{B\rho}{\sqrt{T}}. \tag{41}$$

Thus, we have shown that (24) still holds and the rest of the proof is identical to the proof of SGD.

6 Learning with SGD

We now show how we can apply SGD to standard machine learning problems. First, recall that:

$$L_D(w) = \mathbb{E}_{z\sim D}[\ell(w, z)]. \tag{42}$$

In ERM, we minimize the empirical error L_S as an estimate of minimizing $L_D(w)$. With SGD, we can minimize $L_D(w)$ directly. Since we do not know the distribution D we cannot simply calculate $\nabla L_D(w^{(t)})$ and minimize it with *gradient descent*. With SGD however, we can. We only need a random vector whose conditional expected value is $\nabla L_D(w^{(t)})$ i.e., an unbiased estimate of $\nabla L_D(w^{(t)})$. The following process describes how we can minimize the true risk using SGD.

Process

- First we sample $z \sim D$
- Then, we define v_t to be the gradient of $\ell(w, z)$ with respect to w at $w^{(t)}$.
- The gradient of the loss function $\ell(w, z)$ at $w^{(t)}$ is therefore an unbiased estimate of the gradient of the risk function, $\nabla L_D(w^{(t)})$.

The last statement is easy to see by considering that:

$$\mathbb{E}[v_t|w^{(t)}] = \mathbb{E}_{z\sim D}[\nabla \ell(w^{(t)}, z)] = \nabla\, \mathbb{E}_{z\sim D}[\ell(w^{(t)}, z)] = \nabla L_D(w^{(t)}). \tag{43}$$

With this in mind, we describe the *SGD method for minimizing* $L_D(w)$ below.

- Parameters: scalar $\eta > 0$, integer $T > 0$.
- Initialize: $w^{(1)} = 0$.
- For $t = 1, 2, ..., T$
  - sample $z \sim D$
  - choose $v_t \in \partial l(w^{(t)}, z)$
  - update $w^{(t+1)} = w^{(t)} - \eta v_t$

  Output: $\overline{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$

We now have the following result.

Corollary 1. Consider a convex-Lipschitz-bounded learning problem with parameters ρ, B. Then, for every $\varepsilon > 0$, if we run the SGD method for minimizing $L_D(w)$ with a number of iterations (i.e., number of examples)

$$T \geq \frac{B^2\rho^2}{\varepsilon^2}, \tag{44}$$

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

$$\mathbb{E}[L_D(\overline{w})] \leq \min_{w \in H} L_D(w) + \varepsilon. \tag{45}$$

We now make the following remark.

> ***Remark***. Number of iterations T here matches the number of samples needed i.e., sample complexity. This is because at every iteration we need a sample instance $z \sim D$ to compute $\nabla \ell(w^{(t)}, z)$. So, for T iterations we need T instances and this defines the training set size.

7 SGD for Regularized Loss Minimization

We now want to explore how we can use SGD to solve the following RLM problem:

$$\min_w (\frac{\lambda}{2}\|w\|^2 + L_S(w)). \tag{46}$$

To use SGD, we first need to find a vector v_t which is an unbiased estimator of:

$$\nabla f(w^{(t)}) = \lambda w^{(t)} + \nabla L_S(w^{(t)}). \tag{47}$$

We consider the vector:

$$v_t' = \lambda w^{(t)} + v_t. \tag{48}$$

where $v_t = \nabla \ell(w^{(t)}, z)$. This gives us:

$$\mathbb{E}[v_t'] = \mathbb{E}_{z \sim S, \text{uniformly}}[\lambda w^{(t)} + \nabla \ell(w^{(t)}, z)] = \lambda w^{(t)} + \nabla \, \mathbb{E}_{z \sim S, \text{uniformly}}[\ell(w^{(t)}, z)]. \tag{49}$$

It is only left to show that:

$$\nabla \, \mathbb{E}_{z \sim S, \text{uniformly}}[\ell(w^{(t)}, z)] = \nabla L_S(w^{(t)}). \tag{50}$$

First, note that, every z_i, from the training set $z_1, ..., z_m$, has probability of being selected $1/m$ and results in a value for the loss function $\ell(w^{(t)}, z_i)$. Immediately, we have:

$$\mathbb{E}[\ell(w^{(t)}, z)] = \frac{1}{m}\ell(w^{(t)}, z_1) + ... + \frac{1}{m}\ell(w^{(t)}, z_m) := L_S(w^{(t)}). \tag{51}$$

From this follows (50). We have hence shown that v_t' is an unbiased estimator of $\nabla f(w^{(t)})$. By choosing $\eta_t = \frac{1}{\lambda t}$ we obtain the SGD update rule as:

$$w^{(t+1)} = w^{(t)} - \frac{1}{\lambda t}(\lambda w^{(t)} + v_t). \tag{52}$$

Expanding further and doing some algebraic manipulations lead us to:

$$w^{(t+1)} = -\frac{1}{\lambda t}\sum_{i=1}^{t} v_i. \tag{53}$$

8 Difference between GD and SGD when Minimizing Empirical Risk

Assume that, we wish to apply gradient descent to minimize $L_S(w)$. The iteration rule should be:

$$w^{(t+1)} = w^{(t)} - \eta \nabla L_S(w^{(t)}). \tag{54}$$

Now, further assuming that $w^{(t)} = [w_1^{(t)}, ..., w_n^{(t)}]^T \in \mathbb{R}^n$ and given that $L_S(w^{(t)}) = \frac{1}{m}\sum_{i=1}^{m} \ell(w^{(t)}, z_i)$, one easily realizes that computing $\nabla L_S(w^{(t)})$ involves the computation of $n \times m$ partial derivatives. On the other hand, the iteration rule with SGD is:

$$w^{(t+1)} = w^{(t)} - \eta v_t. \tag{55}$$

where:

$$v_t = \nabla \ell(w^{(t)}, z_i), \tag{56}$$

where z_i is picked from S with probability $1/m$. Thus, at each step of the SGD update rule, we only need to compute the gradient of the loss function at a single z_i which amounts to calculating n partial derivatives. This distinction between SGD and GD makes SGD *computationally much more efficient* in cases where the training set size is large. Therefore, when the training set contains a large number of examples m, GD will be significantly *slower* than SGD, despite both requiring the same number of iterations for convergence. This will be due to the increased computational effort required at each step when performing GD versus SGD.

Conclusions. In these notes, we introduced Stochastic Gradient Descent, a powerful and popular algorithm for solving machine learning optimization problems. We explored different variations of SGD and analytically proved its convergence.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.