

Handout 2: Gradient descent

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim. To introduce gradient descent, its motivation, description, practical tricks, analysis in the convex scenario, and implementation.

****This handout is still Under construction**

1. MOTIVATIONS

Note 1. Consider a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$. Learning may involve the computation of the minimizer $h^* \in \mathcal{H}$, where \mathcal{H} is a class of hypotheses, of the empirical risk function (ERF) $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h, z_i)$ given a finite sample $\{z_i; i = 1, \dots, n\}$ generated from the data generating model $g(\cdot)$ and loss $\ell(\cdot)$; that is

$$(1.1) \quad w^* = \arg \min_{h \in \mathcal{H}} (\hat{R}(h)) = \arg \min_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \ell(h, z_i) \right)$$

If analytical minimization of (1.1) is impossible or impractical, numerical procedures can be applied; eg Gradient Descent (GD) algorithms. Such approaches introduce additional errors in the solution.

2. DESCRIPTION

Notation 2. For the sake of notation simplicity and generalization, we will present Gradient Descent (GD) in the following minimization problem

$$(2.1) \quad w^* = \arg \min_{w \in \mathcal{H}} (f(w))$$

where here $f: \mathbb{R}^d \rightarrow \mathbb{R}$, and $w \in \mathcal{H} \subseteq \mathbb{R}^d$. Eg, f can be an empirical risk function.

Assumption 3. Assume (for now) that $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is differential function.

Definition 4. The Gradient Descent (GD) algorithm for the solution of the minimization problem (2.1) is given in Algorithm 1

Algorithm 1 Gradient descent algorithm; learning rate $\{\eta_t\}$

For $t = 1, 2, 3, \dots, T$, iterate:

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

where

$$\nabla f(w) = \left(\frac{\partial}{\partial x_1} f(w), \dots, \frac{\partial}{\partial x_d} f(w) \right)^\top$$

is the gradient of f at w .

Remark 5. As a motivation, consider the (1st order) Taylor polynomial for the approximation of $f(w)$ in a small area around u (i.e. $\|v - u\| = \text{small}$)

$$f(u) \approx P(u) = f(w) + \langle u - w, \nabla f(w) \rangle$$

Assuming convexity for f , it is

$$f(u) \geq \underbrace{f(w) + \langle u - w, \nabla f(w) \rangle}_{=P(u)}$$

meaning that P lower bounds f . Hence we could design an update mechanism producing $w^{(t+1)}$ that is nearby $w^{(t)}$ (small steps) and

$$(2.2) \quad f(w) \approx f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle.$$

Hence we could recursively minimize the approximation (2.2) and the distance between the current state $w^{(t)}$ and the next w value to produce $w^{(t+1)}$; namely

$$\begin{aligned} w^{(t+1)} &= \arg \min_{\forall w} \left(\frac{1}{2} \|w - w^{(t)}\|^2 + \eta P(w) \right) \\ &= \arg \min_{\forall w} \left(\frac{1}{2} \|w - w^{(t)}\|^2 + \eta \left(f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle \right) \right) \\ &= w^{(t)} + \eta_t \nabla f(w^{(t)}) \end{aligned}$$

where parameter $\eta > 0$ controls the trade off.

Remark 6. The output of GD can be (but not a exclusively), the average

$$(2.3) \quad w_{\text{GD}}^{(T)} = \frac{1}{T} \sum_{t=1}^T f(w^{(t)})$$

after discarding the first few iterations of $w^{(t)}$ for stability reasons, or the best value discovered

$$w_{\text{GD}}^{(T)} = \arg \min_{\forall w_t} \left(f(w^{(t)}) \right)$$

or the last value discovered

$$w_{\text{GD}}^{(T)} = w^{(t)}$$

Note 7. GD algorithm converges to a local minimum, $w_{\text{GD}}^{(T)} \rightarrow w_*$ (in some sense), under different sets of regularity conditions (some are weaker other stronger). Section 4 has a brief discussion.

Remark 8. The parameter η_t is called learning rate, step size, or gain. $\{\eta_t\}$ is a non-negative sequence and it is chosen by the practitioner. Regularity conditions (Note 7) often imply restrictions on the decay of $\{\eta_t\}$ which guide the practitioner to parametrise it properly. Some popular choices of learning rate η_t are:

- (1) constant; $\eta_t = \eta$, where η is a small value. The rational is that GD chain $\{w_t\}$ performs constant small steps towards the (local) minimum w_* and then oscillate around it.
- (2) decreasing and converging to zero; e.g. $\eta_t = \left(\frac{C}{t}\right)^\varsigma$ where $\varsigma \in (0.5, 1]$ and $C > 0$. The rational is that GD algorithm at the beginning starts by performing larger step to explore the area

for discovering possible minima while reducing the size of the steps with the iterations to converge to a possible w_* value.

- (3) decreasing and converging to a tiny value τ_* ; e.g. $\eta_t = \left(\frac{C}{t}\right)^\varsigma + \tau_*$ where $\varsigma \in (0.5, 1]$, $C > 0$, and $\tau_* \approx 0$. Same as previously, but the algorithm aims at oscillating around the detected local minimum.
 - (4) constant until an iteration T_0 and then decreasing; eg $\eta_t = \left(\frac{C}{\max(t, T_0)}\right)^\varsigma$, where $\varsigma \in (0.5, 1]$ and $C > 0$. The rational is that in the first stage of the iterations the algorithm may need constant larger stems for a significant number of iterations in order to explore the domain and hence the chain $\{w_t\}$ to reach the area around the (local) minimum w_* . In the second stage the chain $\{w_t\}$ may be in a close proximity to the (local) minimum w_* and hence the algorithm needs to perform smaller steps to exploit (converge to w_*). The first stage is called burn-in and it is discarded from the output of the GD algorithm.
- Parameters C, ς, τ_* may be chosen based on pilot runs.
 - A set of regularity conditions restricting the choices of $\{\eta_t\}$ are those of Robbins–Monro algorithm (to be seen and discussed later)

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty$$

E.g., they are satisfied by choices in items 2 and 4 .

3. EXAMPLES

4. ANALYSIS

5. GD FOR NON-DIFFERENTIAL FUNCTIONS

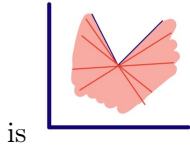
In several learning problems the functions to be minimized is not differentiable. GD can be extended to address such problems with the use of subgradients.

Definition 9. Vector v is called subgradient of a function $f : S \rightarrow \mathbb{R}$ at $w \in S$ iff

$$(5.1) \quad \forall u \in S, \quad f(u) \geq f(w) + \langle u - w, v \rangle$$

Notation 10. The set of subgradients of function $f : S \rightarrow \mathbb{R}$ at $w \in S$ denoted as $\partial f(w)$.

Note 11. Essentially there may be more than one subgradients of the function at a specific point. As seen by (5.1), subgradients are all the lines passing from w and been under the function f that



Theorem 12. If function $f : S \rightarrow \mathbb{R}$ is differentiable at w then the only subgradient of f at w is the gradient $\nabla f(w)$, and (5.1) is equality; i.e. $\partial f(w) = \{\nabla f(w)\}$.