Modern Optimization

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

1. Gradient descent algorithms

Outline

Definition 2.1

A **local numerical optimization** *algorithm is an iterative algorithm where*

$$x^{k+1} = x^k + \alpha_k d_k$$

assuming a starting point x^0 is provided.
The algorithm is characterized by

- *A choice of direction d_k at each iteration.*
- *A choice of step size α_k at each iteration.*

Example 2.1

We have in our mathematical journey already seen some iterative local optimization algorithms:

- Gradient descent: assumes the objective function of an unconstrained problem is differentiable and choose the steepest descent direction: $d_k = -\nabla f_0(x^k)$.
- Newton-like algorithms: assumes a twice differentiable function and pick $d_k = -\nabla^2 f_0(x^k)^{-1} \nabla f_0(x^k)$.
- Quasi-Newton type: approximate the (inverse) Hessian, pick $d_k = -B_k \nabla f_0(x_k)$ where $B_k \approx \nabla^2 f_0(x_k)^{-1}$ (SR1 and BFGS are great candidates)

Example 2.2

They are various ways of selecting the step size

- Constant step – Works in the convex settings, if you know a lot about your function. It should be avoided in most cases
- α_k satisfies the Goldstein conditions – We'll talk about it later. Roughly speaking, it makes sure that the next step decreases the objective value sufficiently.
- α_k satisfies the (weak/strong) Wolfe conditions – We'll talk about it later. Roughly speaking, it makes sure that we decrease the function sufficiently, and that decrease at the next point is not as big as at the previous.
- Backtracking α_k: go somewhat far from x^k and reduce slightly the step size until enough decrease is noticed.

We will give more details to why these strategies work fine in later chapters.

Remark 2.1

This is not the whole story and we will scratch only parts of the problem:

- Line search methods: define a search direction and find a good step size along this direction
- Trust region methods: define a search region and find a good direction within this region.
- One may accelerate the updates ...

Definition 2.2 (Globally convergent algorithms)

An algorithm is said to be **globally convergent** *if*

$$\|\nabla f_0(x^k)\| \to 0.$$

Example 2.3

Note that globally convergence only means convergence to a stationary point. As a counter example think of

$$x \mapsto x^3.$$

Proposition 2.1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. The steepest decrease from a point x^k is done in the direction of the negative gradient.

Definition 2.3 (Vanilla gradient descent: the convex case)

The vanilla gradient descent is characterized by the following iterations

$$x^{k+1} = x^k - \gamma \nabla f_0(x^k),$$

for a constant $\gamma > 0$.
Note: From now on, we will write ∇f_k *or even* g_k *for the gradient evaluated at point* x^k.

Exercise 2.1

Some care should be taken though. Consider the univariate function $f(x) = \frac{1}{2}x^2$ and the step size $\gamma = 2$.
Show that for any given starting point x^0, the vanilla gradient descent will not converge to the optimal point.
This shows that some care should be taken when using the gradient descent method.

Exercise 2.2 (in-class)

Let $A \in \mathbb{R}^{m \times d}$ with $m \geq n$ be a full rank matrix. Let $b \in \mathbb{R}^m$. Let $x^0 \in \mathbb{R}^d$ be a given starting point.
We want to solve the following optimization problem:

$$\min_x \|Ax - b\|_2^2$$

1. Solve the original problem exactly.
2. Show that the gradient descent with step $0 < \gamma < \|A^T A\|^{-1}$ converges to the optimal solution.
3. Consider an adaptive step size and more specifically the exact line search. Compute the value of the optimal step size at every iteration.

Proposition 2.2 (GD: The convex case)

Let f_0 be a convex function with a global minimum x^*. Then, using a fixed stepsize $\gamma > 0$ and starting at any initial point x^0 will yield an error averaged over K steps in the sequence of iterates fulfilling

$$\sum_{k=0}^{K} \left(f(x^k) - f(x^*) \right) \leq \frac{\gamma}{2} \sum_{k=0}^{K} \|g_k\|^2 + \frac{1}{2\gamma} \|x^0 - x^*\|^2.$$

Remark 2.2

It is important to remark:

- We cannot hope much more than this. All we have used here is convexity and differentiability.
- The dependence on $\|x^0 - x^*\|$ makes sense: the further away you start the longer you'll have to work.
- This gradient isn't quite the nicest thing ever.

Proposition 2.3 (GD: The Lipschitz convex case)

Let $f_0 : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with a global minimum x^* and bounded gradient $\|\nabla f_0(x)\| \leq B$, for all x. Assume moreover that your have a starting point x^0. Then, choosing a constant step size

$$\gamma := \frac{\|x^0 - x^*\|}{B\sqrt{K}}$$

the sequence of iterates generated by the constant-step size gradient descent satisfies

$$\frac{1}{K}\sum_{k=0}^{K-1}\Big(f_0(x^k) - f_0(x^*)\Big) \leq \|x^0 - x^*\| \frac{B}{\sqrt{K}}.$$

Remark 2.3

What does this tell and does not tell us:

1. At one point, one iteration is performing well.
2. It's better to know how to approximate something if you know what you approximate
3. You might not manage to use gradient descent on quadratic functions ....

Definition 2.4 (Smooth convex functions)

A function $f : \mathbb{R}^d \to \mathbb{R}$ *be differentiable on its domain and let* $X \subseteq \mathrm{dom}(f)$ *be a convex subset.* f *is said to be* L *smooth over* X *if*

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in X.$$

It called simply L *smooth if it is smooth over its domain.*

Proposition 2.4

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function. The following statements are equivalent:

- *f is smooth with parameter L.*
- *The gradient of f is L Lipschitz.*

Proposition 2.5

Let f be an L smooth convex differentiable function. Then the gradient step with $\gamma = 1/L$ is a descent direction (i.e. decreases the objective value).

Theorem 2.1 (GD: The smooth convex case)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L smooth convex differentiable function with a global minimum x^*. Then the iterates obtained by gradient descent with step size

$$\gamma = \frac{1}{L}$$

satisfy

$$f(x^K) - f(x^*) \leq \frac{L}{2K} \|x^0 - x^*\|^2.$$

Theorem 2.2

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be a* L *smooth convex differentiable function with a global minimum* x^**. Then the iterates obtained by gradient descent with step size*

$$\gamma = \frac{1}{L}$$

satisfy

$$f(x^K) - f(x^*) \leq \frac{2L}{K+4}\|x^0 - x^*\|^2.$$

Proposition 2.6 (Admitted)

Let $K \leq (d-1)/2$, $x^0 \in \mathbb{R}^d$, *and a Lipschitz constant* $L > 0$. *There exists a smooth convex* L*-Lipschitz function* f *with minimizer* x^* *such that*

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}.$$

Definition 2.5

We define Nesterov's second accelerated gradient descent algorithm (AGM2) for a differentiable and L-smooth function f as

$$
\begin{aligned}
x^{k+1} &= y^k - \frac{1}{L}\nabla f(y^k) \\
z^{k+1} &= z^k - \frac{k+1}{2L}\nabla f(y^k) \\
y^{k+1} &= (1-\tau_{k+1})x_{k+1} + \tau_{k+1}z_{k+1},
\end{aligned}
$$

where the memory parameter is set to $\tau_k = \frac{2}{k+2}$.

Proposition 2.7

Let f be a convex differentiable L smooth function which admits a global minimizer x^*. The iterates generated from (AGM2) satisfy

$$f(x^K) - f(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{K(K+1)}, \quad K \geq 1.$$

Proposition 2.8

Nesterov's acceleragted gradient descent updates are equivalent to the following iterations.

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = \left(1 - \frac{1-\lambda_k}{\lambda_{k+1}}\right) x^{k+1} + \frac{1-\lambda_k}{\lambda_{k+1}} x^k,$$

where λ_k *are defined recursively as*

$$\lambda_0 = 0$$
$$\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}.$$