1 Introduction to Gradient Descent

1.1 Introduction

Modern methods of numerical optimisation are largely based on the method of gradient descent.

Suppose that a vector space over $\mathbb{R}^n$ is given with a standard inner product.

We will need two norms: a standard Euclidean norm $\|x\|_p = (\sum_{k=1}^n |x_k|^p)^{\frac{1}{p}}$, and its limit, $\|x\|_\infty = \max_{k=\overline{1,n}} |x_k|$.

Note that $\langle x, y\rangle \leq \|x\|_p \|y\|_q$, where $p, q \in \mathbb{R}$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Such norms $\|\cdot\|_p$ and $\|\cdot\|_q$ are called *dual* norms. Denote a norm dual to the given $\|\cdot\|$ as $\|\cdot\|_*$.

We define a function as *convex*, if for all $y, x \in \mathbb{R}^n$ we have

$$f(x) + \langle \nabla f(x), y - x\rangle \leq f(y),$$

where ∇ is a *subgradient* operator, which is a specialisation of the gradient $(\frac{\delta}{\delta x_1} f, \ldots, \frac{\delta}{\delta x_n} f)$ to convex functions.

1.2 Gradient Descent

We will look at the functions f such that, for x, y arising in the iteration, the smoothness condition $\left\|\nabla f(y) - \nabla f(X)\right\|_* \leq M_\nu \|y - x\|^\nu$ is satisfied, where $\nu \in [0, 1]$ and M_ν is a constant dependent on ν. For example, when $\nu = 1$, $M_\nu = L$, the Lipschitz constant.

We can show that such functions also satisfy $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{M_\nu}{1+\nu} \|y - x\|^{\nu+1}$.

The principal idea of the gradient descent is as follows.

Suppose that $x^{k+1} = \text{argmin}_x \{f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2} \left\|x - x^k\right\|_2^2\}$, where x^k is the vector after the iteration k.

Note that $0 = \nabla f(x^k) + L(x - x^k)$, which means that $x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$.

This is the main property of the gradient descent. First of all, it is convex and its gradient is Lipschitz, which means that we can use both of the results noted above.

The choice of L is important for the efficiency of the gradient descent.

1.3 Non-Convex Cases

The main problem of the non-convex case has to do with falling into the local, rather than global minimum.

To counteract this problem, we check the *stopping criterion*:

$$\left\|\nabla f(x^N)\right\|_2 \leq \epsilon$$

It can be shown that the gradient descent guarantees that we find the solution satisfying the stopping criterion in $N \leq \frac{2L\delta f}{\epsilon}$.

1.4 Finding a Global Minimum

Approximation of the global extremum is a much more daunting task. Unfortunately, we can show that the number steps required is $N \sim \frac{1}{\epsilon^2}$.

1.5 Convex Case

We can look at so-called μ-superconvex functions, such that $f(y) \geq f(x) + \langle \nabla f(x), y \rangle + \frac{\mu}{2}\|y - x\|_2^2$.