Gradient Descent

Daniil Merkulov

Optimization methods. MIPT

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$
f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)
$$

We want h to be a decreasing direction:

$$
f(x + \alpha h) < f(x)
$$

$$
f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)
$$

and going to the limit at $\alpha \to 0$:

$$
\langle f'(x), h \rangle \leq 0
$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Also from Cauchy–Bunyakovsky–Schwarz inequality:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Also from Cauchy–Bunyakovsky–Schwarz inequality:

$$\begin{split} |\langle f'(x), h \rangle| &\leq \|f'(x)\|_2 \|h\|_2 \\ \langle f'(x), h \rangle &\geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2 \end{split}$$

Thus, the direction of the antigradient

$$h = -\frac{f'(x)}{\|f'(x)\|_2}$$

gives the direction of the **steepest local** decreasing of the function f.

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Also from Cauchy–Bunyakovsky–Schwarz inequality:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Thus, the direction of the antigradient

$$h = -\frac{f'(x)}{\|f'(x)\|_2}$$

gives the direction of the **steepest local** decreasing of the function f. The result of this method is

$$x_{k+1} = x_k - \alpha f'(x_k)$$

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \tag{GF}$$

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$
\frac{dx}{dt} = -f'(x(t)) \tag{GF}
$$

and discretize it on a uniform grid with α step:

$$
\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),
$$

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \tag{GF}$$

and discretize it on a uniform grid with α step:

$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),$$

where $x_k \equiv x(t_k)$ and $\alpha = t_{k+1} - t_k$ - is the grid step.
From here we get the expression for x_{k+1}

$$x_{k+1} = x_k - \alpha f'(x_k),$$

which is exactly gradient descent.
Open In Colab ♣

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \tag{GF}$$

and discretize it on a uniform grid with α step:

$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),$$

where $x_k \equiv x(t_k)$ and $\alpha = t_{k+1} - t_k$ - is the grid step.
From here we get the expression for x_{k+1}

$$x_{k+1} = x_k - \alpha f'(x_k),$$

which is exactly gradient descent.
Open In Colab ♣

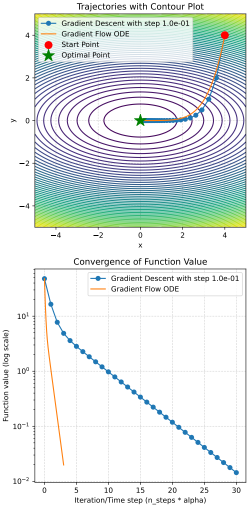


Figure 1: Gradient flow trajectory

Necessary local minimum condition

$$
\begin{aligned}
&f'(x) = 0 \\
&-\eta f'(x) = 0 \\
&x - \eta f'(x) = x \\
&x_k - \eta f'(x_k) = x_{k+1}
\end{aligned}
$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2,$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$

$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2,$$

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,$$

$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

$$
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2,
$$

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$
\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,
$$
$$
\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.
$$

Then

$$
\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.
$$

Now, if we have global upper bound on the function, in a form of parabola, we can try to go directly to its minimum.

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2,$$

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,$$

$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Now, if we have global upper bound on the function, in a form of parabola, we can try to go directly to its minimum.

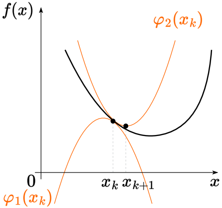


Figure 2: Illustration

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2,$$

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,$$

$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Now, if we have global upper bound on the function, in a form of parabola, we can try to go directly to its minimum.

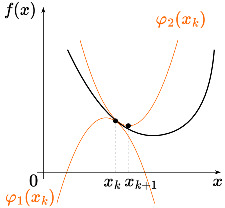


Figure 2: Illustration

$$\nabla \phi_2(x) = 0$$

$$\nabla f(x_0) + L(x^* - x_0) = 0$$

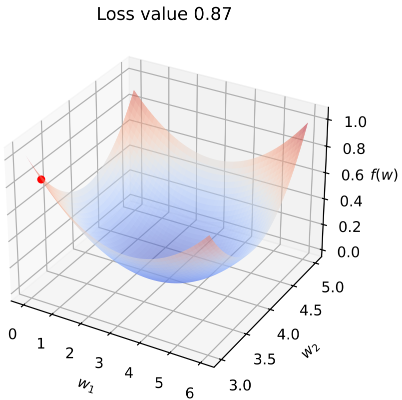
$$x^* = x_0 - \frac{1}{L} \nabla f(x_0)$$

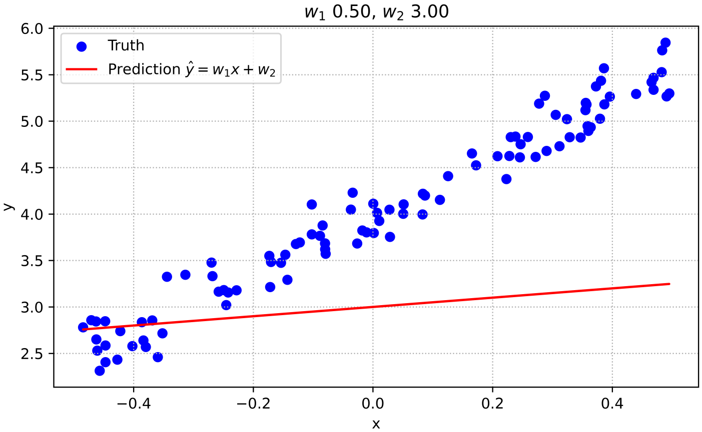
$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$$

This way leads to the $\frac{1}{L}$ stepsize choosing. However, often the L constant is not known.

Convergence of Gradient Descent algorithm

Heavily depends on the choice of the learning rate α:





Exact line search aka steepest descent

$$\alpha_k = \text{arg}\min_{\alpha \in \mathbb{R}^+} f(x_{k+1}) = \text{arg}\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

More theoretical than practical approach. It also allows you to analyze the convergence, but often exact line search can be difficult if the function calculation takes too long or costs a lot. Interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

$$\alpha_k = \text{arg}\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

Optimality conditions:

$$\nabla f(x_{k+1})^\top \nabla f(x_k) = 0$$

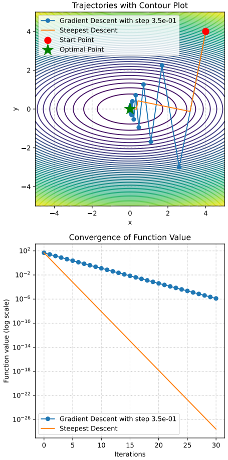


Figure 3: Steepest Descent

Open In Colab ♣

Convergence rates

$$\min_{x \in \mathbb{R}^n} f(x) \qquad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

smooth	convex	smooth & convex	smooth & strongly convex (or PL)
$\|\nabla f(x_k)\|^2 \approx \mathcal{O}\left(\frac{1}{k}\right)$	$f(x_k) - f^* \approx \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$f(x_k) - f^* \approx \mathcal{O}\left(\frac{1}{k}\right)$	$\|x_k - x^*\|^2 \approx \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$

Gradient Descent convergence. Smooth convex case

Gradient Descent convergence. Smooth μ-strongly convex case

Gradient Descent convergence. Polyak-Lojasiewicz case