Gradient Descent. Convergence rates

Daniil Merkulov

Optimization methods. MIPT

Previously

- Gradient Descent

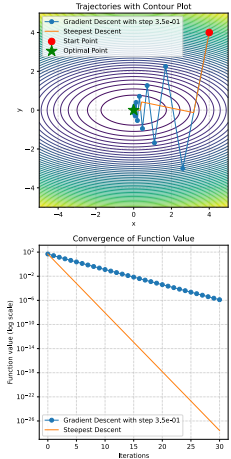


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent

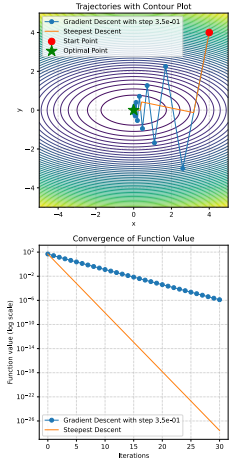


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent
- Convergence rates (no proof)

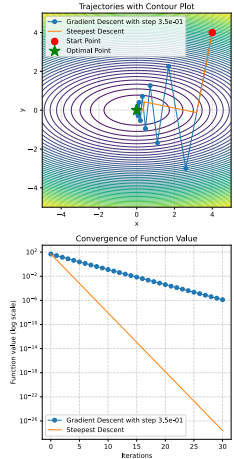


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent
- Convergence rates (no proof)
- If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth then for all $x, y \in \mathbb{R}^d$

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

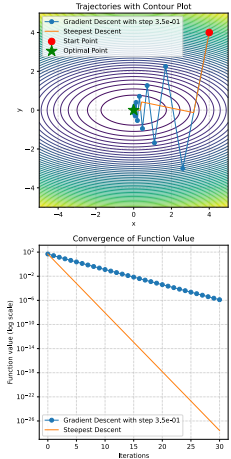


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent
- Convergence rates (no proof)
- If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth then for all $x, y \in \mathbb{R}^d$

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

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable L-smooth function. Then, for all $x \in \mathbb{R}^d$, for every eigenvalue λ of $\nabla^2 f(x)$, we have

$$|\lambda| \leq L.$$

$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

$\lambda_{max}(\nabla^2 f) = L$

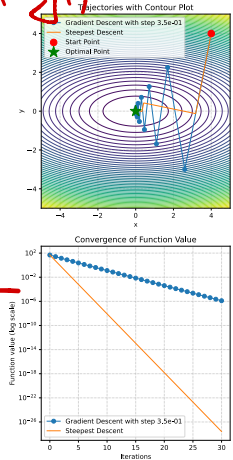


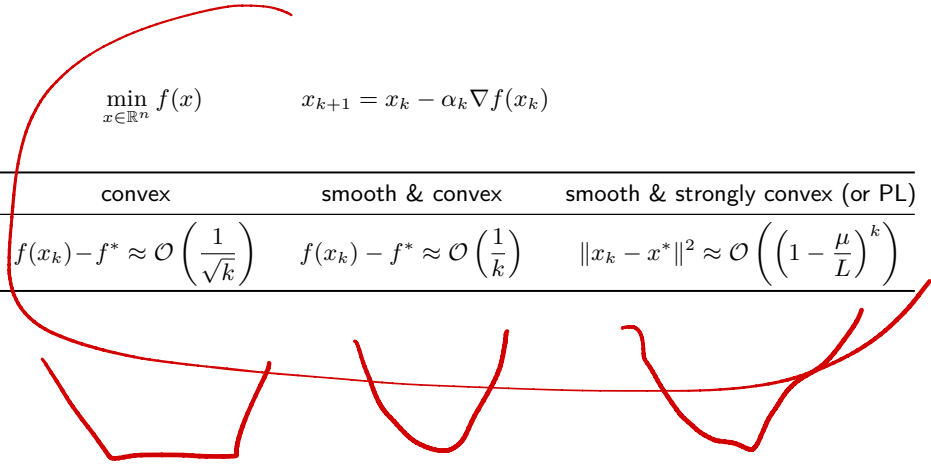
Figure 1: Steepest Descent

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



General quadratic problem

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

для начала $\alpha^k = \text{const}$

$$x^{k+1} = x^k - \alpha(Ax^k - b)$$

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2} x^T A x - b^T x + c$$

если

$$\nabla f(x) = \frac{1}{2}(A + A^T)x - b$$

$$A \geq 0 \Rightarrow A = A^T \succ 0$$

хочу:

$$g(\tilde{x}) = \frac{1}{2}\tilde{x}^T \tilde{A} \tilde{x}$$

$$\boxed{\nabla f = Ax - b}$$

$$x^* = A^{-1} b$$

Замена:

$$\tilde{x} = Q^T(x - x^*) \rightarrow \tilde{x} = x - x^* \rightarrow \boxed{x = Qx + x^*}$$

Метод: $x^{k+1} = (I - \alpha A)x^k + \alpha A x^* \mid -x^*$

$$x^{k+1} - x^* = (I - \alpha A)(x^k - x^*)$$

Восп. $A = Q \Lambda Q^T$ $\quad Q^T Q = I$

$Q = (q_1 \cdot q_2 \dots q_n)$ $\quad \Lambda = \text{diag}(\lambda_1, \dots \lambda_n)$

подставим $Q \Lambda Q^T = A$

General quadratic problem

$$x^{k+1} - x^* = \left(I - \alpha Q \Lambda Q^T\right)(x^k - x^*) \quad | \cdot Q^T$$

$$Q^T(x^{k+1} - x^*) = \left(Q^T - \alpha \Lambda Q^T\right)(x^k - x^*) = (I - \alpha\Lambda) Q^T (x^k - x^*)$$

$$\underset{n\times 1}{\tilde{x}^{k+1}} = \underset{n\times n}{(I - \alpha\Lambda)} \underset{n\times 1}{\tilde{x}_k}$$

← град. спуск для сильно вып. квадр. f

i-ая коорд:

$$\tilde{x}_i^{k+1} = (1 - \alpha\lambda_i)\tilde{x}_i^k$$

← ск-ть $\max_i |1 - \alpha\lambda_i| < 1$

ск-ть ск-ти

$$\alpha < \frac{2}{L}$$

$\lambda_{min} = \mu$, $\lambda_{max} = L$

$|1 - \alpha\mu| < 1$

$1 - \alpha\mu < 1$

$\alpha\mu - 1 < 1$ $\quad \alpha < \frac{2}{\mu}$

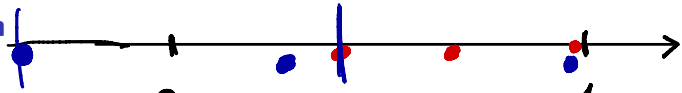
$|1 - \alpha L| < 1$

$\alpha - \alpha L < 1$

$\alpha L - 1 < 1$

$\alpha < \frac{2}{L}$

General quadratic problem



$$\rho = \max_i |1-\alpha\lambda_i| = \max(|1-\alpha\mu|, |1-\alpha L|)$$

выберем α так, чтобы $\rho \to \min$ →

$$1-\alpha^{opt}\mu = \alpha^{opt}L - 1$$

$$\alpha^{opt} = \frac{2}{\mu+L}$$

$|1-\alpha\mu|, |1-\alpha L|$

$\mu = 1$, $L = 10$

$|1-\alpha|, |1-10\alpha|$

оптим. ρ:

$$\rho^{opt} = \frac{L-\mu}{L+\mu} =$$

$$\varkappa = \frac{L}{\mu} \geqslant 1$$

$$= \frac{\varkappa - 1}{\varkappa + 1}$$

по аргументу

$$\text{no } f\left(\frac{\varkappa-1}{\varkappa+1}\right)^2$$

General quadratic problem

$\varkappa = 1.1$	$\rho^2 = 0.0023$	кол-во итер. для уменьш. знач. f в 10 р
		1
3	0.25	2
10	0.67	6
100	0.96	58
200	0.98	116
400	0.99	231

Polyak- Lojasiewicz condition. Linear convergence of gradient descent without convexity

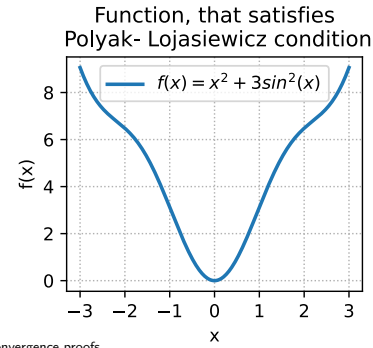
PL inequality holds if the following condition is satisfied for some $\mu > 0$

$$\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \forall x$$

It is interesting, that Gradient Descent algorithm has

The following functions satisfy the PL-condition, but are not convex. Link to the code

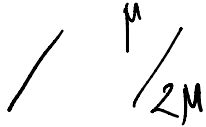
$$f(x) = x^2 + 3\sin^2(x)$$



Polyak- Lojasiewicz condition. Linear convergence of gradient descent without convexity

PL inequality holds if the following condition is satisfied for some $\mu > 0$,

$$\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \forall x$$



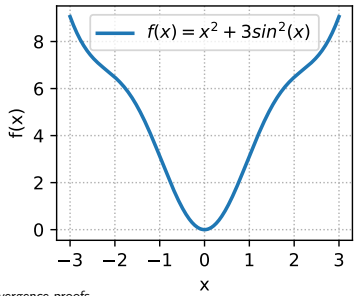
It is interesting, that Gradient Descent algorithm has

The following functions satisfy the PL-condition, but are not convex. Link to the code

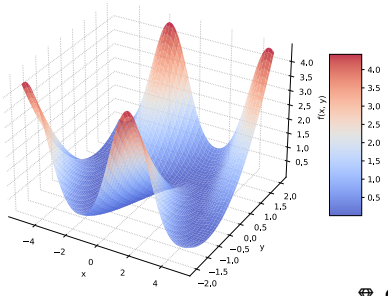
$$f(x) = x^2 + 3\sin^2(x)$$

$$f(x, y) = \frac{(y - \sin x)^2}{2}$$

Function, that satisfies Polyak- Lojasiewicz condition



Non-convex PL function



Gradient Descent convergence. Polyak-Lojasiewicz case

Theorem

Consider the Problem

$$f(x) \to \min_{x \in \mathbb{R}^d}$$

and assume that f is μ-Polyak-Łojasiewicz and L-smooth, for some $L \geq \mu > 0$.
Consider $(x^t)_{t \in \mathbb{N}}$ a sequence generated by the gradient descent constant stepsize algorithm, with a stepsize satisfying $0 < \alpha \leq \frac{1}{L}$. Then:

$$f(x^t) - f^* \leq (1 - \alpha\mu)^t (f(x^0) - f^*).$$

Gradient Descent convergence. Polyak-Lojasiewicz case

We can use L-smoothness, together with the update rule of the algorithm, to write

$$
\begin{aligned}
f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \\
&= f(x^t) - \alpha \|\nabla f(x^t)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^t)\|^2 \\
&= f(x^t) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^t)\|^2 \\
&\leq f(x^t) - \frac{\alpha}{2}\|\nabla f(x^t)\|^2,
\end{aligned}
$$

where in the last inequality we used our hypothesis on the stepsize that $\alpha L \leq 1$.

Gradient Descent convergence. Polyak-Lojasiewicz case

We can use L-smoothness, together with the update rule of the algorithm, to write

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \\ &= f(x^t) - \alpha\|\nabla f(x^t)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^t)\|^2 \\ &= f(x^t) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^t)\|^2 \\ &\leq f(x^t) - \frac{\alpha}{2}\|\nabla f(x^t)\|^2, \end{aligned}$$

where in the last inequality we used our hypothesis on the stepsize that $\alpha L \leq 1$.

We can now use the Polyak-Lojasiewicz property to write:

$$f(x^{t+1}) \leq f(x^t) - \frac{\alpha\mu}{2}(f(x^t) - f^*).$$

The conclusion follows after subtracting f^* on both sides of this inequality, and using recursion.

Gradient Descent convergence. Smooth convex case

Theorem

Consider the Problem

$$f(x) \to \min_{x \in \mathbb{R}^d}$$

and assume that f is convex and L-smooth, for some $L > 0$.
Let $(x^t)_{t \in \mathbb{N}}$ be the sequence of iterates generated by the gradient descent constant stepsize algorithm, with a stepsize satisfying $0 < \alpha \leq \frac{1}{L}$. Then, for all $x^* \in \operatorname{argmin} f$, for all $t \in \mathbb{N}$ we have that

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

Gradient Descent convergence. Smooth convex case

$x^{k+1} = x^k - \alpha \nabla f(x^k)$

f - L smooth : $f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 =$

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

$= f(x^k) + \|\nabla f(x^k)\|^2 \left(\frac{\alpha^2 L}{2} - \alpha \right)$ $\alpha^{opt} = \frac{1}{L}$

$\rightarrow \min \alpha$

$$f(x^k) - f(x^{k+1}) \ge \frac{1}{2L} \|\nabla f(x^k)\|^2$$

$f(x^k) - f(x^{k+1}) \ge 0$

Gradient Descent convergence. Smooth convex case

диф. критерий вып. I: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ — использую выпуклость

$y = x^* \\ x = x^k \Big\} \Rightarrow f(x^k) - f(x^*) \leq \langle \nabla f(x^k), x^k - x^* \rangle$

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{1}{2L}\|\nabla f(x^k)\|^2 =$$

$$= f^* + \left\langle \nabla f(x^k), x^k - x^* - \frac{1}{2L}\nabla f(x^k) \right\rangle =$$

$$= f^* + \frac{L}{2}\left\langle \frac{1}{L}\nabla f(x_k), 2\left(x^k - x^* - \frac{1}{2L}\nabla f(x^k)\right) \right\rangle = \quad (a-b)^T(a+b) = \|a\|^2 - \|b\|^2$$

$$= f^* + \frac{1}{2}\left[\|x^k - x^*\|^2 - \left\|x^k - x^* - \frac{1}{L}\nabla f(x^k)\right\|^2\right]$$

$x^{k+1} - x^k = -\frac{1}{L}\nabla f(x^k)$

$a = x^k - x^*$

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

$a - b = \frac{1}{L}\nabla f(x^k)$

$a + b = 2\left(x^k - x^* - \frac{1}{2L}\nabla f(x^k)\right)$

Gradient Descent convergence. Smooth convex case

$$f(x^{k+1}) \leq f^* + \frac{L}{2}\left[\underbrace{\|x^k - x^*\|^2}_{R_k} - \underbrace{\|x^{k+1} - x^*\|^2}_{R_{k+1}}\right]$$

$$\frac{2}{L}\left(f(x^{k+1}) - f^*\right) \leq R_k - R_{k+1} \quad \leftarrow \quad k = 0, \dots, t$$

$$\sum_{k=0}^{t} \frac{2}{L}\left(f(x^{k+1}) - f^*\right) \leq R_0 - R_t \leq R_0$$

$$\frac{2}{L}\sum_{k=0}^{t-1} f(x^{k+1}) - \frac{2t}{L} \cdot f^* \leq R_0$$

$$\frac{2t}{L} f\left(\sum_{k=0}^{t-1} \frac{1}{t} x^{k+1}\right) \frac{2t}{L} f^*$$

$$f\left(\sum_{k=0}^{t-1} \frac{1}{t} x^{k+1}\right) \leq \frac{1}{t}\sum_{k=0}^{t-1} f(x^{k+1})$$

Gradient Descent convergence. Smooth convex case

$t f(x^t) \le f(x^0) + \dots + f(x^t)$

из-за монотонности метода

$t \cdot f(x^t) \le$

$\sum_{k=0}^{T-1} f(x^{k+1}) - t \cdot f^* \le \frac{R_0 L}{2}$

$t \cdot f(x^t) - t \cdot f^* \le \frac{R_0 \cdot L}{2}$

$$\Rightarrow \boxed{f(x^t) - f^* \le \frac{L \|x^0 - x^*\|^2}{2t}} \qquad \frac{LR^2}{2t}$$

Gradient Descent convergence. Smooth μ-strongly convex case

Theorem

Consider the Problem

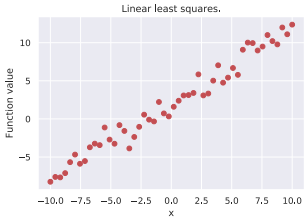
$$f(x) \to \min_{x \in \mathbb{R}^d}$$

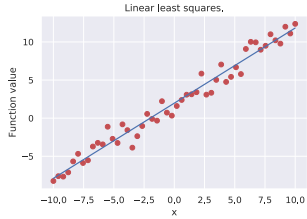
and assume that f is μ-strongly convex and L-smooth, for some $L \geq \mu > 0$. Let $(x^t)_{t \in \mathbb{N}}$ be the sequence of iterates generated by the gradient descent constant stepsize algorithm, with a stepsize satisfying $0 < \alpha \leq \frac{1}{L}$. Then, for $x^* = \operatorname{argmin} f$ and for all $t \in \mathbb{N}$:

$$\|x^{t+1} - x^*\|^2 \leq (1 - \alpha\mu)^{t+1} \|x^0 - x^*\|^2.$$

Gradient Descent convergence. Smooth μ-strongly convex case

Gradient Descent for Linear Least Squares aka Linear Regression





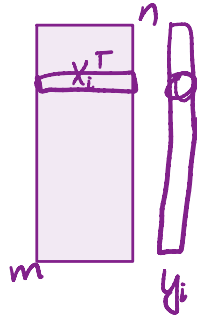


Figure 4: Illustration

In a least-squares, or linear regression, problem, we have measurements $X \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$ and seek a vector $\theta \in \mathbb{R}^n$ such that $X\theta$ is close to y. Closeness is defined as the sum of the squared differences:

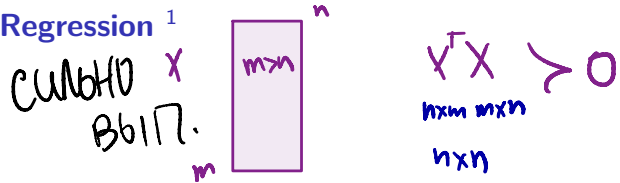
$$\sum_{i=1}^{m} (x_i^\top \theta - y_i)^2 = \|X\theta - y\|_2^2 \to \min_{\theta \in \mathbb{R}^n}$$

$\nabla f = X^T(X\theta - y)$

$\nabla^2 f = X^T X$

For example, we might have a dataset of m users, each represented by n features. Each row $x_i^\top$ of X is the features for user i, while the corresponding entry y_i of y is the measurement we want to predict from $x_i^\top$, such as ad spending. The prediction is given by $x_i^\top \theta$.

Linear Least Squares aka Linear Regression [1]



1. Is this problem convex? Strongly convex?

Linear Least Squares aka Linear Regression [1]

1. Is this problem convex? Strongly convex?
2. What do you think about convergence of Gradient Descent for this problem?

[1] Take a look at the example of real-world data linear least squares problem

l_2-regularized Linear Least Squares

In the underdetermined case, it is often desirable to restore strong convexity of the objective function by adding an l_2-penality, also known as Tikhonov regularization, l_2-regularization, or weight decay.

$$\|X\theta - y\|_2^2 + \frac{\mu}{2}\|\theta\|_2^2 \to \min_{\theta \in \mathbb{R}^n}$$

Note: With this modification the objective is μ-strongly convex again.

Take a look at the code