

Gradient Descent. Convergence rates

Daniil Merkulov

Optimization methods. MIPT

Previously

- Gradient Descent

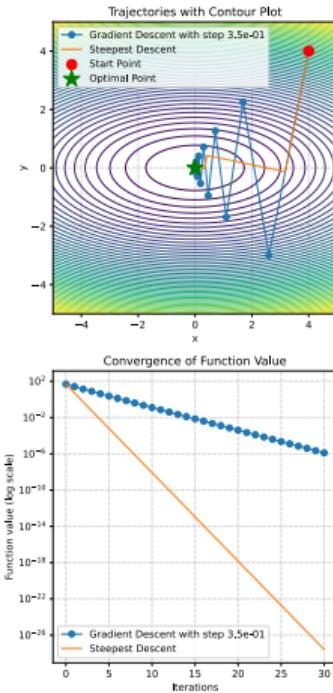


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent

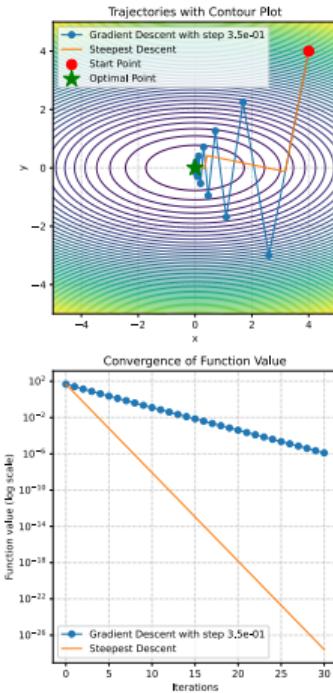


Figure 1: Steepest Descent

Open In Colab ♣

Previously

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

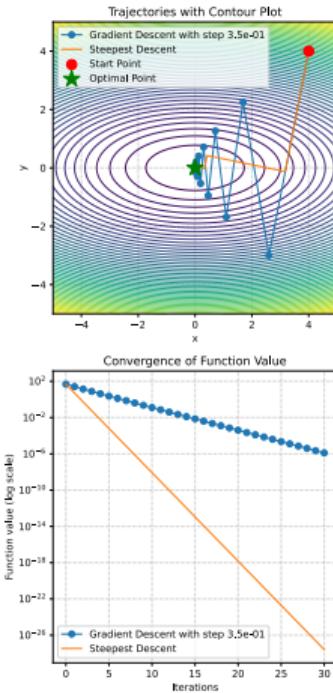


Figure 1: Steepest Descent

Open In Colab ♣

Previously

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

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

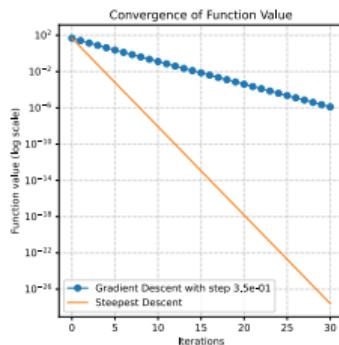
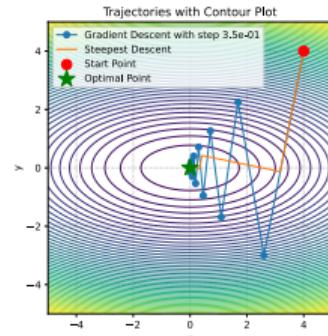


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent
- Convergence rates (no proof)
- If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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

$$|\lambda| \leq L. \quad \rightarrow \quad \lambda_{\max}(\nabla^2 f) = L$$

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

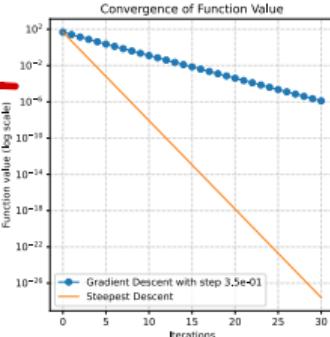
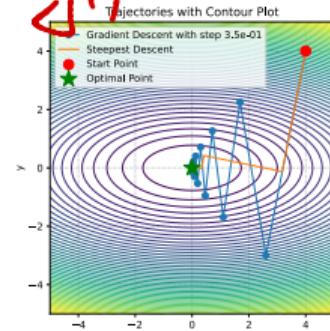


Figure 1: Steepest Descent

Open In Colab ♣

Convergence rates

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

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

smooth

convex

smooth & convex

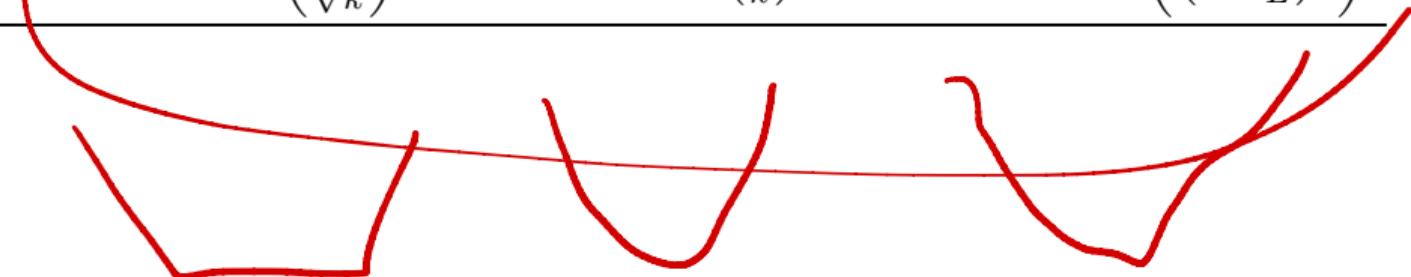
smooth & strongly convex (or PL)

$$\|\nabla f(x_k)\|^2 \approx \mathcal{O}\left(\frac{1}{k}\right)$$

$$f(x_k) - f^* \approx \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$

$$f(x_k) - f^* \approx \mathcal{O}\left(\frac{1}{k}\right)$$

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



General quadratic problem

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

где шаг $\lambda^k = \text{const}$

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

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

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

easy

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

тогда:

$$A \succeq 0 \Rightarrow A = A^T > 0$$

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

$$\nabla f = Ax - b$$

$$x^* = \underline{\tilde{A}^{-1} b}$$

Замена:

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

$$x = Qx + x^*$$

$$\text{Бесн. } A = Q \Lambda Q^T$$

$$Q^T Q = I$$

$$\begin{matrix} \uparrow & \uparrow \\ (q_1, q_2, \dots, q_n) & \text{diag}(\lambda_1, \dots, \lambda_n) \end{matrix}$$

$$\text{Метр: } x^{k+1} = (I - \lambda A)x^k + \lambda A x^* \quad | -x^*$$

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

$$\text{ногоравен } Q \Lambda Q^T = A$$

General quadratic problem $x^{k+1} - x^* = (I - \alpha Q \Lambda Q^T)(x^k - x^*)$ | · Q^T

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

ipaq. счyck gne
свободн. бен. к бeгp.f

i-akkoopgi

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

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

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

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

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

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

$$1 - \alpha L < 1$$

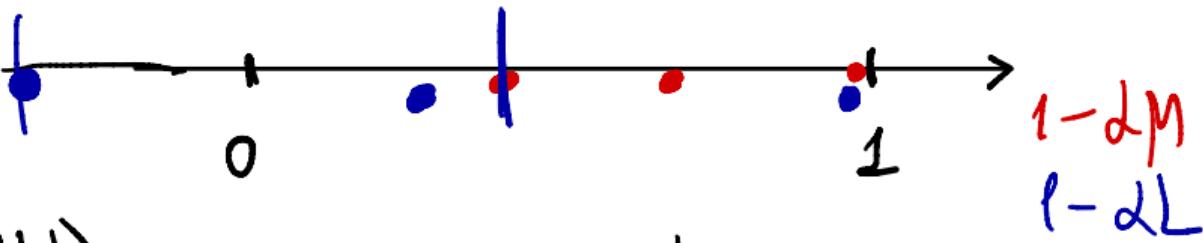
$$\alpha L - 1 < 1 \quad \alpha < \frac{2}{L}$$

ex-T6 $\max_i |1 - \alpha \lambda_i| < 1$

ex-T6 $\alpha < \frac{2}{L}$

ex-T4

General quadratic problem



$$P = \max_i (1 - d\lambda_i) = \max(1 - d\mu, 1 - dL)$$

berechne d tak, wo für

$$|1 - d\mu|, |1 - dL|$$

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

$$\underline{|1 - d|}, \underline{|1 - 10d|}$$

$$P \rightarrow \min \rightarrow$$

$$d^{\text{opt}}_L - 1 = 1 - d^{\text{opt}}_M$$

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

entw. P:

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

$$x = \frac{L}{\mu} \geq 1$$

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

no optimality
not $\left(\frac{x-1}{x+1}\right)^2$

General quadratic problem

$$\lambda = 1.1$$

$$f^2 = 0.0023$$

кон-бо шаг.
где уменьш.
этот. 5 б10
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 $\mu > 0$,

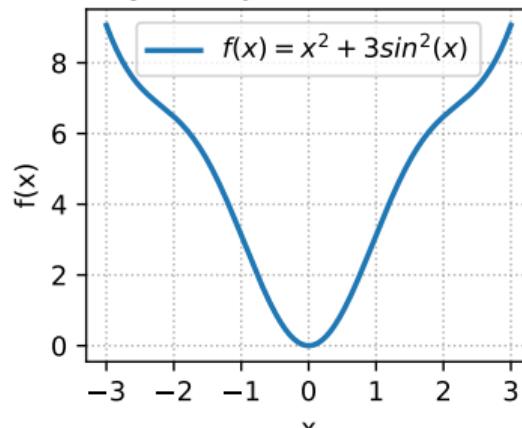
$$\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \forall x$$

It is interesting, that Gradient Descent algorithm has

The following functions satisfy the PL-condition, but are not convex.  [Link to the code](#)

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

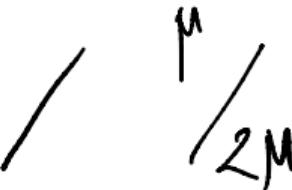
Function, that satisfies
Polyak- Lojasiewicz condition



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

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

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

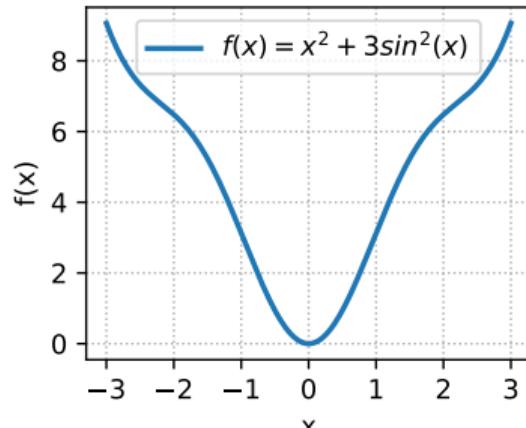


It is interesting, that Gradient Descent algorithm has

The following functions satisfy the PL-condition, but are not convex. [Link to the code](#)

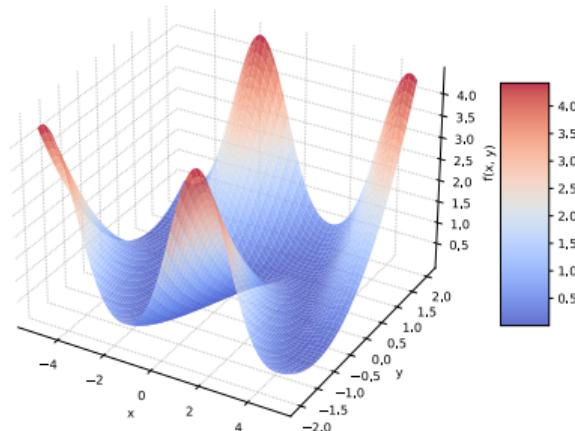
$$f(x) = x^2 + 3 \sin^2(x)$$

Function, that satisfies
Polyak- Lojasiewicz condition



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

Non-convex PL function



Gradient Descent convergence. Polyak-Łojasiewicz case

Theorem

Consider the Problem

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

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:

$$f(x^t) - f^* \leq \left(1 - \frac{\alpha\mu}{2}\right)^t (f(x^0) - f^*).$$

Gradient Descent convergence. Polyak-Lojasiewicz case

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

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

$$\underbrace{\nabla f(x^t)}_{\text{L-smooth}} = x^t - x^{t+1}$$

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

$$= f(x^t) - \alpha \|\nabla f(x^t)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^t)\|^2$$

$$= f(x^t) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^t)\|^2$$

$$\leq f(x^t) - \frac{\alpha}{2} \|\nabla f(x^t)\|^2,$$

L - wagaet6

CD

$$|1 - L\alpha| < 1$$

$$\alpha L \leq 1 \quad \alpha \leq \frac{1}{L}$$

where in the last inequality we used our hypothesis on the stepsize that

Gradient Descent convergence. Polyak-Lojasiewicz case

PL:

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

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

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

$\frac{\alpha}{2} \|\nabla f(x^t)\|^2$

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:

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

$$\begin{aligned} f(x^{t+1}) - f^* &\leq f(x^t) - f^* - \frac{\alpha\mu}{2}(f(x^t) - f^*) \\ f(x^{t+1}) - f^* &= \left(1 - \frac{\alpha\mu}{2}\right)(f(x^t) - f^*) \end{aligned}$$

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

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

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

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

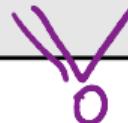
Gradient Descent convergence. Smooth convex case $x^{k+1} = x^k - \lambda \nabla f(x^k)$

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

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

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

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



Gradient Descent convergence. Smooth convex case

гип. критерий

бесн. I:

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

использовано

бывшукость

$$\left. \begin{array}{l} y = x^* \\ x = x^* \end{array} \right\} \Rightarrow f(x^k) - f(x^*) \leq \langle \nabla f(x^k), x^k - x^* \rangle$$



$$\begin{aligned}
 f(x^{k+1}) &\leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|^2 \leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{1}{2L} \|\nabla f(x^k)\|^2 = \\
 &= f^* + \left\langle \nabla f(x^k), x^k - x^* - \frac{1}{2L} \nabla f(x^k) \right\rangle = \\
 &= f^* + \frac{L}{2} \left\langle \frac{1}{L} \nabla f(x^k), 2 \left(x^k - x^* - \frac{1}{2L} \nabla f(x^k) \right) \right\rangle = \boxed{\frac{1}{2} (a - b)(a + b)} = \\
 &= f^* + \frac{L}{2} \left[\|x^k - x^*\|^2 - \left\| x^k - x^* - \frac{1}{2L} \nabla f(x^k) \right\|^2 \right]
 \end{aligned}$$

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

$$a = x^k - x^*$$

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

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

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

Gradient Descent convergence. Smooth convex case

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

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

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

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

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

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

Gradient Descent convergence. Smooth convex case

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

У3-ЗА
 МОНОТОННОСТЬ мережи

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

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

$$\Rightarrow f(x^t) - f^* \leq \frac{L \|x^0 - x^*\|^2}{2t} \quad \frac{LR^2}{2t}$$

Gradient Descent convergence. Smooth μ -strongly convex case

Theorem

Consider the Problem

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

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

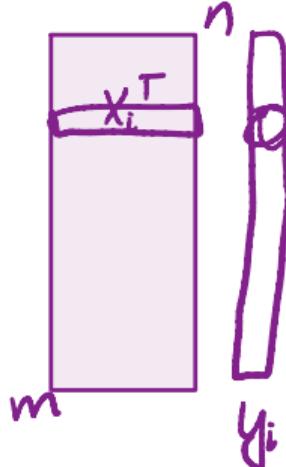
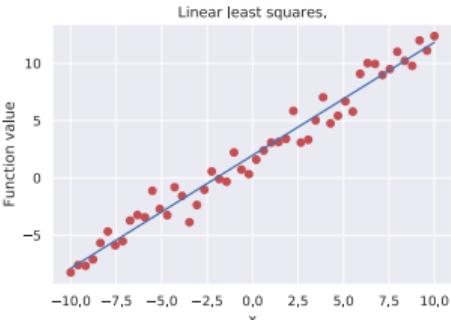
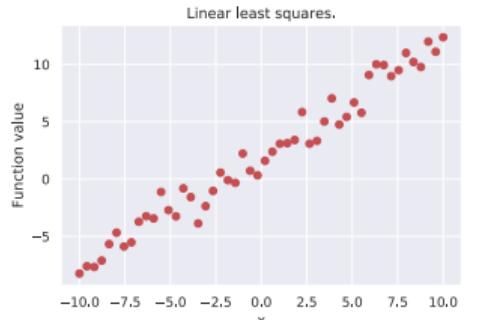


Figure 4: Illustration

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:

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

$$\begin{aligned}\nabla f &= X^\top(X\theta - y) \\ \nabla^2 f &= X^\top X\end{aligned}$$

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

$$\text{CubH0} \quad X^T X \quad m > n$$

X :
 $m \times n$

$$X^T X \succ 0$$

$n \times m \quad m \times n$

1. Is this problem convex? Strongly convex?

$$\text{Bb1n.} \quad X^T X \succeq 0$$

$m < n$

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?

¹Take 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 -penalty, also known as Tikhonov regularization, l_2 -regularization, or weight decay.

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

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

Take a look at the  code