Convexity and smoothness. Gradient descent. Newton's method

Optimization in ML

Aleksandr Beznosikov

Skoltech

21 November 2023

Local/global minimum

Let's consider the unconditional optimization task: $\min_{x \in \mathbb{R}^d} f(x)$.

Local/global minimum

Let's consider the unconditional optimization task: $\min_{x \in \mathbb{R}^d} f(x)$.

Local minimum

Point x^* is called local minimum of function f in $\mathbb{R}^d$ (local solution of f minimization task in $\mathbb{R}^d$), if there exists such $r > 0$, that every $y \in B_2^d(r, x^*) = \{y \in \mathbb{R}^d \mid \|y - x^*\|_2 \leq r\}$ implies $f(x^*) \leq f(y)$.

Local/global minimum

Let's consider the unconditional optimization task: $\min_{x \in \mathbb{R}^d} f(x)$.

Local minimum

Point x^* is called local minimum of function f in $\mathbb{R}^d$ (local solution of f minimization task in $\mathbb{R}^d$), if there exists such $r > 0$, that every $y \in B_2^d(r, x^*) = \{y \in \mathbb{R}^d \mid \|y - x^*\|_2 \leq r\}$ implies $f(x^*) \leq f(y)$.

Global minimum

Point x^* is called global minimum of function f in $\mathbb{R}^d$ (global solution of f minimization task in $\mathbb{R}^d$), if $y \in \mathbb{R}^d$ implies $f(x^*) \leq f(y)$.

Local/global minimum

Let's consider the unconditional optimization task: $\min_{x \in \mathbb{R}^d} f(x)$.

Local minimum

Point x^* is called local minimum of function f in $\mathbb{R}^d$ (local solution of f minimization task in $\mathbb{R}^d$), if there exists such $r > 0$, that every $y \in B_2^d(r, x^*) = \{y \in \mathbb{R}^d \mid \|y - x^*\|_2 \leq r\}$ implies $f(x^*) \leq f(y)$.

Global minimum

Point x^* is called global minimum of function f in $\mathbb{R}^d$ (global solution of f minimization task in $\mathbb{R}^d$), if $y \in \mathbb{R}^d$ implies $f(x^*) \leq f(y)$.

The definition can also be generalized to a local/global minimum in the set $\mathcal{X}$, i.e. for a task of the form $\min_{x \in \mathcal{X}} f(x)$. In this case, one should consider $y \in B_2^d(r, x^*) \cap \mathcal{X}$ и $y \in \mathcal{X}$ in the appropriate definitions.

Optimality condition: general case

Necessary condition of local minimum

Let x^* be a local minimum of f in $\mathbb{R}^d$. Then differentiability of f implies $\nabla f(x^*) = 0$.

Optimality condition: general case

Proof

We will prove from the contrary and suggest $\nabla f(x^*) \neq 0$.Let's write Taylor series in a neighbourhood of local minimum:

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + o(\|x - x^*\|_2),$$

where $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.

Optimality condition: general case

Proof

We will prove from the contrary and suggest $\nabla f(x^*) \neq 0$.Let's write Taylor series in a neighbourhood of local minimum:

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + o(\|x - x^*\|_2),$$

where $\lim_{x \to x^*} \frac{o(\|x-x^*\|_2)}{\|x-x^*\|_2} = 0$.
Let's look at $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Our goal is to choose such λ, that $\tilde{x}$ is in appropriate neighbourhood from the definition of local minimum.

Optimality condition: general case

Proof

We will prove from the contrary and suggest $\nabla f(x^*) \neq 0$.Let's write Taylor series in a neighbourhood of local minimum:

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + o(\|x - x^*\|_2),$$

where $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.
Let's look at $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Our goal is to choose such λ, that $\tilde{x}$ is in appropriate neighbourhood from the definition of local minimum. It is obvious, that such λ exists.

Optimality condition: general case

Proof

We will prove from the contrary and suggest $\nabla f(x^*) \neq 0$.Let's write Taylor series in a neighbourhood of local minimum:

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + o(\|x - x^*\|_2),$$

where $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.

Let's look at $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Our goal is to choose such λ, that $\tilde{x}$ is in appropriate neighbourhood from the definition of local minimum. It is obvious, that such λ exists. On the one hand:

$$f(\tilde{x}) \geq f(x^*), \quad \text{and}$$

Optimality condition: general case

Proof

We will prove from the contrary and suggest $\nabla f(x^*) \neq 0$.Let's write Taylor series in a neighbourhood of local minimum:

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + o(\|x - x^*\|_2),$$

where $\lim_{x \to x^*} \frac{o(\|x-x^*\|_2)}{\|x-x^*\|_2} = 0$.

Let's look at $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Our goal is to choose such λ, that $\tilde{x}$ is in appropriate neighbourhood from the definition of local minimum. It is obvious, that such λ exists. On the one hand:

$$f(\tilde{x}) \geq f(x^*), \quad \text{and}$$

$$\begin{aligned} f(\tilde{x}) &= f(x^*) + \langle \nabla f(x^*), \tilde{x} - x^* \rangle + o(\|\tilde{x} - x^*\|_2) \\ &= f(x^*) - \lambda \|\nabla f(x^*)\|^2 + o(\lambda \|\nabla f(x^*)\|_2) \end{aligned}$$

Optimality condition: general case

Proof

Let's throw another restriction on "smallness" of λ. Now consider $|o(\lambda\|\nabla f(x^*)\|_2)| \leq \frac{\lambda}{2}\|\nabla f(x^*)\|_2^2$. Then on the other hand we have $\lambda > 0$

$$f(\tilde{x}) \leq f(x^*) - \frac{\lambda}{2}\|\nabla f(x^*)\|^2$$

We came to the contradiction with definition of x^*.

Local/global minimum

- Our goal is a global minimum (or point, which is close to it in some sense).
- It became clear that it is pointless to look for a global minimum without additional assumptions.

Convexity: definition

Definition of convex function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable function. It is convex, if every $x, y \in \mathbb{R}^d$ implies

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

Convexity: definition

Definition of convex function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable function. It is convex, if every $x, y \in \mathbb{R}^d$ implies

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

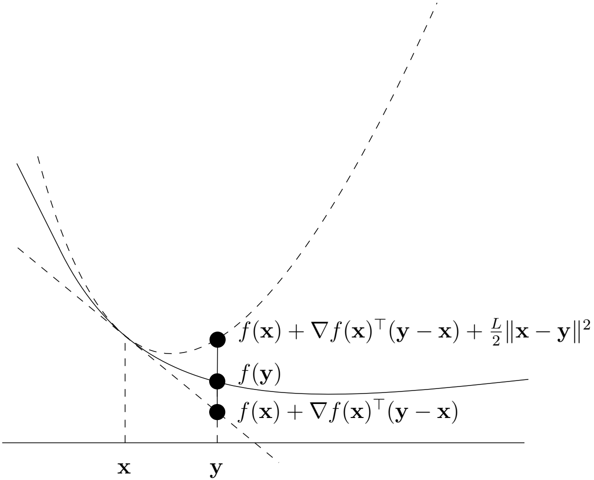
Another definition, which is equivalent in the case of differentiable functions.

Definition of convex function

It is convex, if every $x, y \in \mathbb{R}^d$ and every $\lambda \in [0; 1]$ implies

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

Convexity



Restriction from below on function behavior.

Strong convexity: definition

Definition of μ-strongly convex function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable function. It is μ-strongly convex ($\mu > 0$), if every $x, y \in \mathbb{R}^d$ implies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

Strong convexity: definition

Definition of μ-strongly convex function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable function. It is μ-strongly convex ($\mu > 0$), if every $x, y \in \mathbb{R}^d$ implies

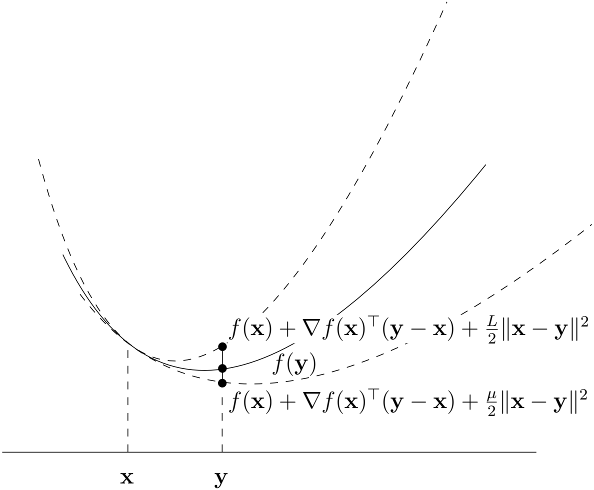
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

Definition of μ-strongly convex function

It is μ-strongly convex, if every $x, y \in \mathbb{R}^d$ and every $\lambda \in [0; 1]$ implies

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x - y\|_2^2$$

Strong convexity



Stronger restriction from below on behavior.

Minima of convex functions

Theorem on minima of convex functions

Consider a problem:

$$\min_{x \in \mathcal{X}} f(x),$$

where f – convex, $\mathcal{X}$ - convex. Then every local minimum of f in $\mathcal{X}$ is global.

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

where x is an arbitrary point in $\mathcal{X}$.

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

where x is an arbitrary point in $\mathcal{X}$. **Question:** what could be said about x_λ?

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1-\lambda)x^*,$$

where x is an arbitrary point in $\mathcal{X}$. **Question:** what could be said about x_λ? $x_\lambda \in \mathcal{X}$ Because of convexity of $\mathcal{X}$.

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1 - \lambda) x^*,$$

where x is an arbitrary point in $\mathcal{X}$. **Question:** what could be said about x_λ? $x_\lambda \in \mathcal{X}$ Because of convexity of $\mathcal{X}$. Choose such small $\lambda > 0$, that x_λ is in a neighbourhood, where x^* is a local minimum.

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1-\lambda)x^*,$$

where x is an arbitrary point in $\mathcal{X}$. **Question:** what could be said about x_λ? $x_\lambda \in \mathcal{X}$ Because of convexity of $\mathcal{X}$. Choose such small $\lambda > 0$, that x_λ is in a neighbourhood, where x^* is a local minimum.Then because of convexity of f we have

$$f(x^*) \leq f(x_\lambda) \leq \lambda f(x) + (1-\lambda) f(x^*).$$

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

where x is an arbitrary point in $\mathcal{X}$. **Question:** what could be said about x_λ? $x_\lambda \in \mathcal{X}$ Because of convexity of $\mathcal{X}$. Choose such small $\lambda > 0$, that x_λ is in a neighbourhood, where x^* is a local minimum.Then because of convexity of f we have

$$f(x^*) \le f(x_\lambda) \le \lambda f(x) + (1 - \lambda)f(x^*).$$

Question:what did we get?

Minima of convex functions

Proof

Let x^* be local minimum. Let's look at

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

where x is an arbitrary point in $\mathcal{X}$. **Question:** what could be said about x_λ? $x_\lambda \in \mathcal{X}$ Because of convexity of $\mathcal{X}$. Choose such small $\lambda > 0$, that x_λ is in a neighbourhood, where x^* is a local minimum.Then because of convexity of f we have

$$f(x^*) \leq f(x_\lambda) \leq \lambda f(x) + (1 - \lambda) f(x^*).$$

Question:what did we get? $f(x) \geq f(x^*)$. By virtue of arbitrariness $x \in \mathcal{X}$ we have global minimum.

Minima of convex functions

Theorem on minima of convex functions

Consider a problem

$$\min_{x \in \mathcal{X}} f(x),$$

where f is convex, $\mathcal{X}$ is convex. Then the set of minimum points $\mathcal{X}^*$ is convex.

Minima of convex functions

Proof

An empty set and a set of 1 points are convex.

Minima of convex functions

Proof

An empty set and a set of 1 points are convex. Now let $x_1^*, x_2^* \in \mathcal{X}^*$. Have a look at $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, where $\lambda \in [0;1]$. $x_\lambda^* \in \mathcal{X}$ By virtue of $\mathcal{X}$ convexity.

Minima of convex functions

Proof

An empty set and a set of 1 points are convex. Now let $x_1^*, x_2^* \in \mathcal{X}^*$. Have a look at $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, where $\lambda \in [0;1]$. $x_\lambda^* \in \mathcal{X}$ By virtue of $\mathcal{X}$ convexity.
By virtue of f convexity:

$$f^* \leq f(x_\lambda^*) \leq \lambda f(x_1^*) + (1-\lambda) f(x_2^*) = f^*.$$

Minima of convex functions

Proof

An empty set and a set of 1 points are convex. Now let $x_1^*, x_2^* \in \mathcal{X}^*$. Have a look at $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, where $\lambda \in [0; 1]$. $x_\lambda^* \in \mathcal{X}$ By virtue of $\mathcal{X}$ convexity.
By virtue of f convexity:

$$f^* \leq f(x_\lambda^*) \leq \lambda f(x_1^*) + (1-\lambda)f(x_2^*) = f^*.$$

Thus, $f(x_\lambda^*) = f^*$, which implies $x^* \in \mathcal{X}^*$.

Minima of convex functions

Theorem on minima of convex functions

Consider a problem

$$\min_{x \in \mathcal{X}} f(x),$$

where f is *strongly* convex, $\mathcal{X}$ is convex. Then the set of minimum points $\mathcal{X}^*$ consists of only one element.

Minima of convex functions

Proof

From the contrary: Let $x_1^* \neq x_2^* \in \mathcal{X}^*$. Have a look at $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, where $\lambda \in (0;1)$. Again, $x_\lambda^* \in \mathcal{X}$ because of $\mathcal{X}$ convexity.

Minima of convex functions

Proof

From the contrary: Let $x_1^* \neq x_2^* \in \mathcal{X}^*$. Have a look at $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, where $\lambda \in (0;1)$. Again, $x_\lambda^* \in \mathcal{X}$ because of $\mathcal{X}$ convexity.
But now due to the strong convexity of the function f:

$$
\begin{aligned}
f^* \leq f(x_\lambda^*) &\leq \lambda f(x_1^*) + (1-\lambda) f(x_2^*) - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2 \\
&= f^* - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2.
\end{aligned}
$$

Minima of convex functions

Proof

From the contrary: Let $x_1^* \neq x_2^* \in \mathcal{X}^*$. Have a look at $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, where $\lambda \in (0;1)$. Again, $x_\lambda^* \in \mathcal{X}$ because of $\mathcal{X}$ convexity.

But now due to the strong convexity of the function f:

$$
\begin{aligned}
f^* \leq f(x_\lambda^*) &\leq \lambda f(x_1^*) + (1-\lambda)f(x_2^*) - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2 \\
&= f^* - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2.
\end{aligned}
$$

The last term < 0 due to choice $x_1^* \neq x_2^*$ and $\lambda \in (0;1)$. Contradiction.

Minima of convex functions

Theorem on minima of convex functions

Consider a problem

$$\min_{x \in \mathcal{X}} f(x),$$

where f – *strongly* convex, $\mathcal{X}$ - convex. Then the set of minimum points $\mathcal{X}^*$ consists of only one element.

- For a strongly convex function, it can be proved that the solution is strictly unique (i.e., add existence to the previous theorem). This follows from the fact that we are always propped up by a parabola from below. See the proof in the manual.

Strong convexity: more facts

A theorem on another equivalent definition of strong convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable in $\mathbb{R}^d$. Then f is μ-strongly convex if and only if every $x, y \in \mathbb{R}^d$ implies

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2.$$

Strong convexity: more facts

A theorem on another equivalent definition of strong convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable in $\mathbb{R}^d$. Then f is μ-strongly convex if and only if every $x, y \in \mathbb{R}^d$ implies

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2.$$

Strong convexity criterion theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable in $\mathbb{R}^d$. Then f is μ-strongly convex if and only if every $x \in \mathbb{R}^d$ implies

$$\nabla^2 f(x) \succeq \mu I.$$

Strong convexity: more facts

A theorem on another equivalent definition of strong convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable in $\mathbb{R}^d$. Then f is μ-strongly convex if and only if every $x, y \in \mathbb{R}^d$ implies

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2.$$

Strong convexity criterion theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable in $\mathbb{R}^d$. Then f is μ-strongly convex if and only if every $x \in \mathbb{R}^d$ implies

$$\nabla^2 f(x) \succeq \mu I.$$

Both facts are proved in the manual. The second one will be useful for HW.

Smoothness: definition

definition of L-smooth function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable in $\mathbb{R}^d$. We will say that this function has L-lipschitz gradient (it is L-smooth), if every $x, y \in \mathbb{R}^d$ implies

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

Smoothness: definition

definition of L-smooth function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable in $\mathbb{R}^d$. We will say that this function has L-lipschitz gradient (it is L-smooth), if every $x, y \in \mathbb{R}^d$ implies

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

definition of L-smoothness can also be written in a non-Euclidean norm. Therefore, formally, in the previous definition, it is possible to indicate L-smoothness in terms of $\| \cdot \|_2$.

Smoothness: properties

Theorem (property of L-smooth function)

Consider L - smooth function $f : \mathbb{R}^d \to \mathbb{R}$. Than every $x, y \in \mathbb{R}^d$ implies

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|x - y\|_2^2.$$

Smoothness: properties

Proof

Let's start with the Newton-Leibniz formula

$$f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau$$

$$= \langle \nabla f(x), y - x \rangle + \int\limits_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau$$

Smoothness: properties

Proof

Let's start with the Newton-Leibniz formula

$$f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + \tau(y-x)), y - x \rangle d\tau$$

$$= \langle \nabla f(x), y - x \rangle + \int\limits_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle d\tau$$

Then

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| = \left| \int\limits_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle d\tau \right|$$

$$\leq \int\limits_0^1 |\langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle| d\tau$$

Smoothness: properties

Proof

Let's apply Cauchy-Schwartz inequality:

$$
\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &\leq \int\limits_0^1 |\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle| d\tau \\
&\leq \int\limits_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_2 \|y - x\|_2
\end{aligned}
$$

Smoothness: properties

Proof

Let's apply Cauchy-Schwartz inequality:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \int_0^1 |\langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x\rangle| d\tau$$

$$\leq \int_0^1 \|\nabla f(x + \tau(y-x)) - \nabla f(x)\|_2 \|y - x\|_2$$

Now apply definition of L-smoothness:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq L\|y - x\|_2^2 \int_0^1 \tau d\tau$$

$$= \frac{L}{2}\|x - y\|_2^2$$

Smoothness: properties

Theorem (properties of L - smooth convex function)

Consider L - smooth *convex* function $f : \mathbb{R}^d \to \mathbb{R}$. Then every $x, y \in \mathbb{R}^d$ implies

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|_2^2$$

и

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y).$$

Smoothness: properties

Theorem (properties of L - smooth convex function)

Consider L - smooth *convex* function $f : \mathbb{R}^d \to \mathbb{R}$. Then every $x, y \in \mathbb{R}^d$ implies

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|x - y\|_2^2$$

и

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y).$$

Proof

The proof of the first fact follows from convexity and the previous smoothness property: the submodule expression is valid because of convexity.

Smoothness: properties

Proof

Consider $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Question:** Is it L_ϕ-smooth? convex?

Smoothness: properties

Proof

Consider $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Question:** Is it L_ϕ-smooth? convex?
Yes to both questions. $L_\phi = L$ (by definition).

Smoothness: properties

Proof

Consider $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Question:** Is it L_ϕ-smooth? convex? Yes to both questions. $L_\phi = L$ (by definition). One also can observe that $y^* = x$ is minimum. **Question:** Why?

Smoothness: properties

Proof

Consider $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Question:** Is it L_ϕ-smooth? convex? Yes to both questions. $L_\phi = L$ (by definition). One also can observe that $y^* = x$ is minimum. **Question:** Why? $\nabla \phi(y^*) = \nabla \phi(x) = 0$. Let's use the first statement of theorem: $f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|_2^2$ c $\left(y = y - \frac{1}{L}\nabla \phi(y), x = y, f = \phi \right)$. Then

$$\phi\left(y - \frac{1}{L}\nabla\phi(y)\right) - \phi(y) - \left\langle \nabla\phi(y), -\frac{1}{L}\nabla\phi(y) \right\rangle \leq \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

After a little rearrangement:

$$\phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Smoothness: properties

Proof

Then we get, knowing that $y^* = x$ is the minimum:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Smoothness: properties

Proof

Then we get, knowing that $y^* = x$ is the minimum:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Substituting ϕ:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Smoothness: properties

Proof

Then we get, knowing that $y^* = x$ is the minimum:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Substituting ϕ:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

It remains to rearrange:

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

Smoothness: properties

Proof

Then we get, knowing that $y^* = x$ is the minimum:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Substituting ϕ:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

It remains to rearrange:

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

Question: have we used convexity here at all?

Smoothness: properties

Proof

Then we get, knowing that $y^* = x$ is the minimum:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Substituting ϕ:

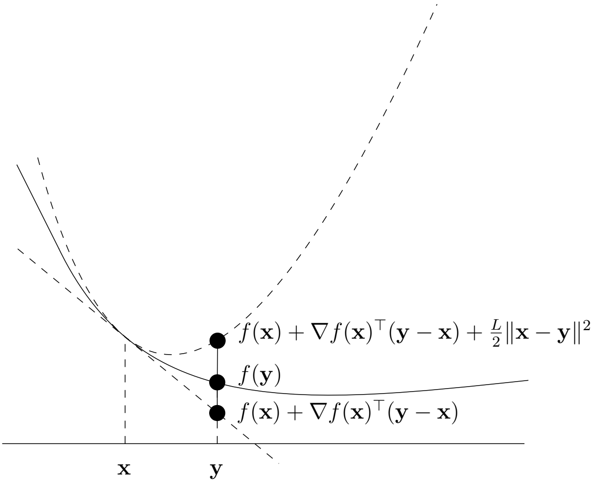
$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

It remains to rearrange:

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

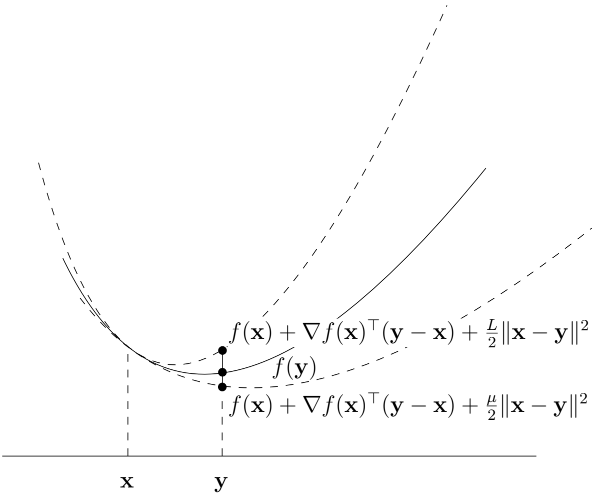
Question: have we used convexity here at all?Yes, $\nabla(y^*) = 0 \;\Rightarrow\; y^*$ is minimum

Smoothness: physical meaning



The restriction from above on behavior (growth) – does not grow too fast.

Smoothness: physical meaning



Gradient descent

- **Problem:** find a solution to unconditional optimization:

$$\min_{x \in \mathbb{R}^d} f(x). \tag{1}$$

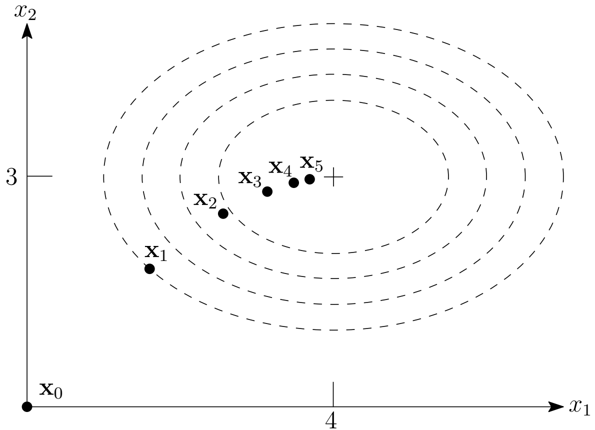
Algorithm 1 Gradient descent

Input: stepsizes $\{\gamma_k\}_{k=0} > 0$, start point $x^0 \in \mathbb{R}^d$, number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Estimate $\nabla f(x^k)$
3: $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$
4: **end for**

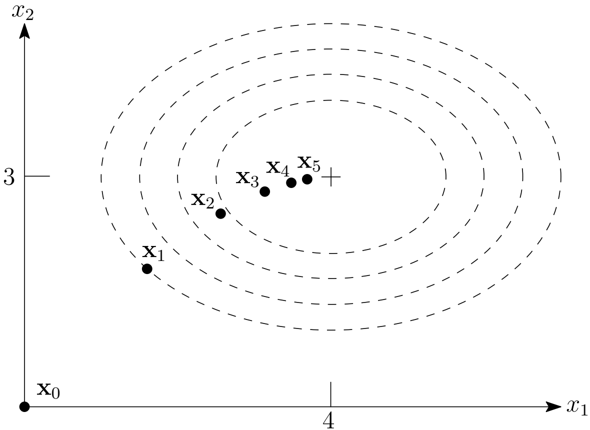
Output: x^K

Example



Question: what is the gradient direction at point x_1?

Example



Question: what is the gradient direction at point x_1? growth direction

Convergence: L-smooth и μ-strongly convex functions

Proof

We know that for strongly convex functions the solution is unique, let us try to estimate how the distance to the solution changes. Let us substitute the iteration:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Convergence: L-smooth и μ-strongly convex functions

Proof

We know that for strongly convex functions the solution is unique, let us try to estimate how the distance to the solution changes. Let us substitute the iteration:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Question: what is next?

Convergence: L-smooth и μ-strongly convex functions

Proof

We know that for strongly convex functions the solution is unique, let us try to estimate how the distance to the solution changes. Let us substitute the iteration:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Question: what is next? Remembering that we have smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$.

Convergence: L-smooth и μ-strongly convex functions

Proof

We know that for strongly convex functions the solution is unique, let us try to estimate how the distance to the solution changes. Let us substitute the iteration:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Question: what is next? Remembering that we have smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$. It is enough just to recall the optimality condition $\nabla f(x^*) = 0$.

Convergence: L-smooth и μ-strongly convex functions

Proof

We know that for strongly convex functions the solution is unique, let us try to estimate how the distance to the solution changes. Let us substitute the iteration:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Question: what is next? Remembering that we have smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$. It is enough just to recall the optimality condition $\nabla f(x^*) = 0$.

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

Convergence: L-smooth и μ-strongly convex functions

Proof

Smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2\|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|_2^2$:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - x^*\|_2^2 - 2\gamma_k\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle \\
&+ \gamma_k^2\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \\
\leq& \|x^k - x^*\|_2^2 - 2\gamma_k\mu\|x^k - x^*\|_2^2 + \gamma_k^2 L^2\|x^k - x^*\|_2^2 \\
=& (1 - 2\gamma_k\mu + \gamma_k^2 L^2)\|x^k - x^*\|_2^2
\end{aligned}
$$

Convergence: L-smooth и μ-strongly convex functions

Proof

Smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2\|x-y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \mu \|x-y\|_2^2$:

$$
\begin{aligned}
\|x^{k+1}-x^*\|_2^2 =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle \\
& + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \\
\leq& \|x^k - x^*\|_2^2 - 2\gamma_k \mu \|x^k - x^*\|_2^2 + \gamma_k^2 L^2 \|x^k - x^*\|_2^2 \\
=& (1 - 2\gamma_k \mu + \gamma_k^2 L^2) \|x^k - x^*\|_2^2
\end{aligned}
$$

Question: what do we want now?

Convergence: L-smooth и μ-strongly convex functions

Proof

Smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\
&+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \\
\leq& \|x^k - x^*\|_2^2 - 2\gamma_k \mu \|x^k - x^*\|_2^2 + \gamma_k^2 L^2 \|x^k - x^*\|_2^2 \\
=& (1 - 2\gamma_k \mu + \gamma_k^2 L^2) \|x^k - x^*\|_2^2
\end{aligned}
$$

Question: what do we want now? $(1 - 2\gamma_k \mu + \gamma_k^2 L^2) < 1$. How to choose it?

Convergence: L-smooth и μ-strongly convex functions

Proof

Smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$:

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\
&+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \\
\leq& \|x^k - x^*\|_2^2 - 2\gamma_k \mu \|x^k - x^*\|_2^2 + \gamma_k^2 L^2 \|x^k - x^*\|_2^2 \\
=& (1 - 2\gamma_k \mu + \gamma_k^2 L^2) \|x^k - x^*\|_2^2
\end{aligned}$$

Question: what do we want now? $(1 - 2\gamma_k \mu + \gamma_k^2 L^2) < 1$. How to choose it? $\arg\min_{\gamma_k} (1 - 2\gamma_k \mu + \gamma_k^2 L^2)$?

Convergence: L-smooth и μ-strongly convex functions

Proof

Smoothness $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ and a strong convexity in the form of $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\
& + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \\
\leq& \|x^k - x^*\|_2^2 - 2\gamma_k \mu \|x^k - x^*\|_2^2 + \gamma_k^2 L^2 \|x^k - x^*\|_2^2 \\
=& (1 - 2\gamma_k \mu + \gamma_k^2 L^2) \|x^k - x^*\|_2^2
\end{aligned}
$$

Question: what do we want now? $(1 - 2\gamma_k \mu + \gamma_k^2 L^2) < 1$. How to choose it? $\arg\min_{\gamma_k} (1 - 2\gamma_k \mu + \gamma_k^2 L^2)$? $\gamma_k = \frac{\mu}{L^2}$ и $(1 - 2\gamma_k \mu + \gamma_k^2 L^2) = 1 - \frac{\mu^2}{L^2}$.

Convergence: L-smooth и μ-strongly convex functions

Proof

Totally:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Convergence: L-smooth и μ-strongly convex functions

Proof

Totally:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Let us run the recursion:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Convergence: L-smooth и μ-strongly convex functions

Proof

Totally:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Let us run the recursion:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Question: what is that type of convergence rate?

Convergence: L-smooth и μ-strongly convex functions

Proof

Totally:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Let us run the recursion:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Question: what is that type of convergence rate? Linear.

Convergence: L-smooth и μ-strongly convex functions

Proof

Totally:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Let us run the recursion:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Question: what is that type of convergence rate? Linear. And how do we get an estimate for the number of iterations?

Convergence: L-smooth и μ-strongly convex functions

Proof

Totally:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Let us run the recursion:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Question: what is that type of convergence rate? Linear. And how do we get an estimate for the number of iterations? (Here we just need to remember the exponent's decomposition into a Taylor series)

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2$$

Convergence: L-smooth и μ-strongly convex functions

Proof

From the previous slide:

$$\|x^K - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2$$

Convergence: L-smooth и μ-strongly convex functions

Proof

From the previous slide:

$$\|x^K - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2$$

We want to guarantee that

$$\|x^K - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2 \leq \varepsilon^2$$

Then we logarithmize and obtain

$$K \geq \frac{L^2}{\mu^2} \log\left(\frac{\|x^0 - x^*\|_2^2}{\varepsilon^2}\right)$$

Convergence: L-smooth и μ-strongly convex functions

Proof

From the previous slide:

$$\|x^K - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2$$

We want to guarantee that

$$\|x^K - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2 \leq \varepsilon^2$$

Then we logarithmize and obtain

$$K \geq \frac{L^2}{\mu^2} \log\left(\frac{\|x^0 - x^*\|_2^2}{\varepsilon^2}\right)$$

Totally: Not great, not terrible – it can be better. An example of how in getting top grades we can «load it up».

Convergence: L-smooth и μ-strongly convex functions

Proof

We start the same way:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\
& + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

Convergence: L-smooth и μ-strongly convex functions

Proof

We start the same way:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\
& + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

But let us make it thinner. A strong convexity in the form of:
$-\langle \nabla f(x), x - y \rangle \leq -\left(\frac{\mu}{2} \|x - y\|_2^2 + f(x) - f(y) \right)$:

Convergence: L-smooth и μ-strongly convex functions

Proof

We start the same way:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\ =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\ =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

But let us make it thinner. A strong convexity in the form of: $-\langle \nabla f(x), x - y \rangle \leq -\left(\frac{\mu}{2}\|x - y\|_2^2 + f(x) - f(y)\right)$:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq& \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

Convergence: L-smooth и μ-strongly convex functions

Proof

Further smoothness, but in the form of:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Question:** is everything true about this property?

Convergence: L-smooth и μ-strongly convex functions

Proof

Further smoothness, but in the form of:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Question:** is everything true about this property? Yes, it's used that $\nabla f(x^*) = 0$. We get

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) \\
= & (1 - \gamma_k \mu)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f(x^*))
\end{aligned}
$$

Convergence: L-smooth и μ-strongly convex functions

Proof

Further smoothness, but in the form of:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Question:** is everything true about this property? Yes, it's used that $\nabla f(x^*) = 0$. We get

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) \\
= & (1 - \gamma_k \mu)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f(x^*))
\end{aligned}
$$

Question: what is left?

Convergence: L-smooth и μ-strongly convex functions

Proof

Further smoothness, but in the form of: $\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Question:** is everything true about this property? Yes, it's used that $\nabla f(x^*) = 0$. We get

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k\left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) \\
= & (1 - \gamma_k\mu)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f(x^*))
\end{aligned}
$$

Question: what is left? $(\gamma_k L - 1) \leq 0$. Which means $\gamma_k \leq \frac{1}{L}$.

$$
\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma_k\mu)\|x^k - x^*\|_2^2
$$

Convergence: L-smooth и μ-strongly convex functions

Proof

From the previous slide:

$$\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma_k \mu)\|x^k - x^*\|_2^2$$

Running recursion:

$$\|x^K - x^*\|_2^2 \leq \prod_{k=0}^{K-1} (1 - \gamma_k \mu)\|x^0 - x^*\|_2^2$$

With a constant stepsize $\gamma_k = \gamma = \frac{1}{L}$:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \|x^0 - x^*\|_2^2$$

Convergence: L-smooth и μ-strongly convex functions

The gradient descent convergence theorem for L-smooth and μ-strongly convex functions

Let the unconditional optimization problem (1) with L-smooth, μ-strongly convex objective function f is solved using gradient descent. Then the following convergence estimate is valid:

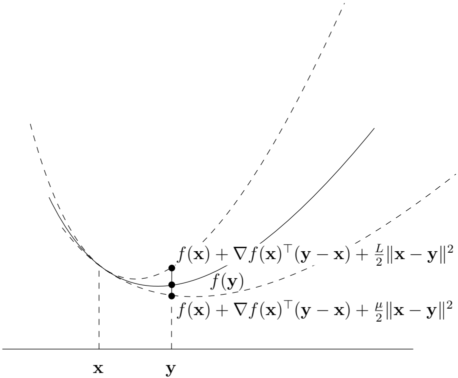
$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \|x^0 - x^*\|_2^2.$$

Moreover, in order to achieve the accuracy of ε on the argument, it is necessary to

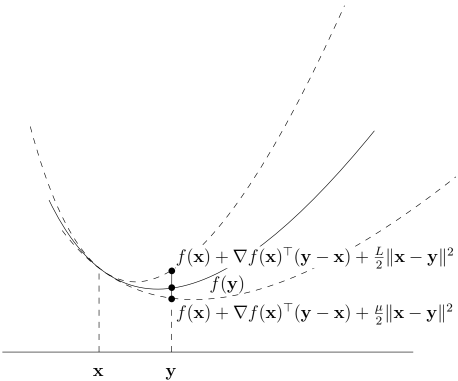
$$K = O\left(\frac{L}{\mu} \log \frac{\|x^0 - x^*\|_2}{\varepsilon}\right) = \tilde{O}\left(\frac{L}{\mu}\right) \text{ iterations.}$$

We will use O-notation to "remove" numerical factors and $\tilde{O}$-notation to remove log-factors as well.

A bit of proof intuition



A bit of proof intuition



We walk based on the properties of the upper boundary (L) – to ensure that we don't "fly away", and move in the worst case based on the properties of the lower boundary (μ).

Convergence

	μ-strongly convex	convex	nonconvex
L-smooth	$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|_2}{\varepsilon}\right)$	$O\left(\frac{L\|x^0-x^*\|_2^2}{\varepsilon}\right)$	$O\left(\frac{L(f(x^0)-f^*)}{\varepsilon^2}\right)$
M-Lipschitz	$O\left(\frac{M^2}{\mu^2\varepsilon}\right)$	$O\left(\frac{M^2\|x^0-x^*\|_2^2}{\varepsilon^2}\right)$	grid search

- In the strongly convex case by argument: $\|x - x^*\|_2 \leq \varepsilon$,
- In the convex case on the function (the solution of x^* may not be unique): $f(x) - f^* \leq \varepsilon$,
- In the non-convex case (convergence to some stationary point): $\|\nabla f(x)\|_2 \leq \varepsilon$.

Convergence

	μ-strongly convex	convex	nonconvex
L-smooth	$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|_2}{\varepsilon}\right)$	$O\left(\frac{L\|x^0-x^*\|_2^2}{\varepsilon}\right)$	$O\left(\frac{L(f(x^0)-f^*)}{\varepsilon^2}\right)$
M-Lipschitz	$O\left(\frac{M^2}{\mu^2\varepsilon}\right)$	$O\left(\frac{M^2\|x^0-x^*\|_2^2}{\varepsilon^2}\right)$	grid search

- In the strongly convex case by argument: $\|x - x^*\|_2 \leq \varepsilon$,
- In the convex case on the function (the solution of x^* may not be unique): $f(x) - f^* \leq \varepsilon$,
- In the non-convex case (convergence to some stationary point): $\|\nabla f(x)\|_2 \leq \varepsilon$.
- Gradient descent is optimal (**question:** what does that mean?) in the nonsmooth case as well as in the smooth nonconvex case.
- Our analysis of gradient descent in the strongly convex case is unimprovable with numerical multipliers.
- In the smooth convex and strongly convex cases, improvements are possible, but this requires a different method.

Stepsize selection: Polak-Shore

We already have

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2} \|x^k - x^*\|_2^2 + f(x^k) - f(x^*) \right) \\
& + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Stepsize selection: Polak-Shore

We already have

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*) \right) \\
& + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Question: how can we pick γ_k optimally in this situation?

Stepsize selection: Polak-Shore

We already have

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2} \|x^k - x^*\|_2^2 + f(x^k) - f(x^*) \right) \\
& + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}$$

Question: how can we pick γ_k optimally in this situation?
$\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \right)$?

Stepsize selection: Polak-Shore

We already have

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*) \right) \\
& + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Question: how can we pick γ_k optimally in this situation?
$\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \right)$?

$$
\gamma_k = \frac{f(x^k) - f(x^*)}{\|\nabla f(x^k)\|_2^2}
$$

Question: what problems do we see?

Stepsize selection: Polak-Shore

We already have

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\ & + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\ \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Question: how can we pick γ_k optimally in this situation?
$\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2\right)$?

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\|\nabla f(x^k)\|_2^2}$$

Question: what problems do we see? $f(x^*)$ – is sometimes known and sometimes can be estimated.

Stepsize selection

- Polak-Shore stepsize:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(have to be chosen)}$$

Stepsize selection

- Polak-Shore stepsize:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(have to be chosen)}$$

- Steepest descent:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Stepsize selection

- Polak-Shore stepsize:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(have to be chosen)}$$

- Steepest descent:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Question: how to solve?

Stepsize selection

- Polak-Shore stepsize:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(have to be chosen)}$$

- Steepest descent:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Question: how to solve? but when there is an explicit formula, you have to solve a one-dimensional problem..

Stepsize selection

- Polak-Shore stepsize:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(have to be chosen)}$$

- Steepest descent:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

  Question: how to solve? but when there is an explicit formula, you have to solve a one-dimensional problem..

- Armijo, Wolfe and Goldstein rules.
- Adaptive selection, e.g., online estimation of the local constant L.

Root finding problem

- Consider following problem of finding the root of the function:

$$\text{Find } t^*, \text{ s.t. } \varphi(t^*) = 0,$$

where $\varphi : \mathbb{R} \to \mathbb{R}$.

Root finding problem

- Consider following problem of finding the root of the function:

$$\text{Find } t^*, \text{ s.t. } \varphi(t^*) = 0,$$

  where $\varphi : \mathbb{R} \to \mathbb{R}$.
- Let's take the point t^0 and we want to find such Δt that $t^0 + \Delta t \approx t^*$.

Root finding problem

- Consider following problem of finding the root of the function:

$$\text{Find } t^*, \text{ s.t. } \varphi(t^*) = 0,$$

  where $\varphi : \mathbb{R} \to \mathbb{R}$.
- Let's take the point t^0 and we want to find such Δt that $t^0 + \Delta t \approx t^*$.
- Expand in series:

$$\varphi(t^0 + \Delta t) = \varphi(t^0) + \varphi'(t^0)\Delta t + o(\Delta t).$$

Root finding problem

- Consider following problem of finding the root of the function:

$$\text{Find } t^*, \text{ s.t. } \varphi(t^*) = 0,$$

where $\varphi : \mathbb{R} \to \mathbb{R}$.

- Let's take the point t^0 and we want to find such Δt that $t^0 + \Delta t \approx t^*$.
- Expand in series:

$$\varphi(t^0 + \Delta t) = \varphi(t^0) + \varphi'(t^0)\Delta t + o(\Delta t).$$

- As we want $t^0 + \Delta t \approx t^*$, consider

$$\varphi(t^0 + \Delta t) \approx \varphi(t^*) = 0 \;\Rightarrow\; \varphi(t^0) + \varphi'(t^0)\Delta t \approx 0.$$

Root finding problem: Newton's method

- From $\varphi(t^0) + \varphi'(t^0)\Delta t \approx 0$ received:

$$\Delta t \approx -\frac{\varphi(t^0)}{\varphi'(t^0)}.$$

Root finding problem: Newton's method

- From $\varphi(t^0) + \varphi'(t^0)\Delta t \approx 0$ received:

$$\Delta t \approx -\frac{\varphi(t^0)}{\varphi'(t^0)}.$$

- So, we get the new point $t^1 = t^0 + \Delta t$ with following iterative procedure:

$$\boxed{t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)}}$$

- This method called Newton's method. It was introduced in the second half of the 17th century by exactly that Newton.

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method?

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root?

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$.

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$. From where, Newton's method iteration
$$t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)} = -(t^k)^3.$$

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$. From where, Newton's method iteration
$$t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)} = -(t^k)^3.$$
- **Question:** what can we say about the convergence to the solution?

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$. From where, Newton's method iteration
$$t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)} = -(t^k)^3.$$
- **Question:** what can we say about the convergence to the solution?
  - $|t^0| < 1$ — there is convergence

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$. From where, Newton's method iteration
$$t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)} = -(t^k)^3.$$
- **Question:** what can we say about the convergence to the solution?
  - $|t^0| < 1$ — there is convergence
  - $|t^0| = 1$ — fluctuating between -1 and 1

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$

  Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$. From where, Newton's method iteration

$$t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)} = -(t^k)^3.$$

- **Question:** what can we say about the convergence to the solution?
  - $|t^0| < 1$ — there is convergence
  - $|t^0| = 1$ — fluctuating between -1 and 1
  - $|t^0| > 1$ — divergence

Newton's method: local convergence

- **Question:** what are the questions to the intuition of getting an iteration of Newton's method? It's important that t^0 is from «good neighbourhood» of t^*.
- Consider
$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}.$$
Question: what's the root? $t^* = 0$.
- Derivative: $\varphi'(t) = \frac{1}{(1+t^2)^{3/2}}$. From where, Newton's method iteration
$$t^{k+1} = t^k - \frac{\varphi(t^k)}{\varphi'(t^k)} = -(t^k)^3.$$
- **Question:** what can we say about the convergence to the solution?
    - $|t^0| < 1$ — there is convergence
    - $|t^0| = 1$ — fluctuating between -1 and 1
    - $|t^0| > 1$ — divergence
- The key point of Newton's method is local convergence (only in the

Newton's method: optimisation

- Consider the unconditional optimization problem with a convex, twice continuously differentiable objective function:

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

- **Question:** for such a task, we are also looking for 0, but what?

Newton's method: optimisation

- Consider the unconditional optimization problem with a convex, twice continuously differentiable objective function:

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

- **Question:** for such a task, we are also looking for 0, but what? $\nabla f(x^*) = 0$.

Newton's method: optimisation

- Consider the unconditional optimization problem with a convex, twice continuously differentiable objective function:

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

- **Question:** for such a task, we are also looking for 0, but what? $\nabla f(x^*) = 0$. Where does Newton's method come from for the optimization problem

---

Algorithm 4 Newton's method

Input: starting point $x^0 \in \mathbb{R}^d$, amount of iterations K

1: **for** $k = 0, 1, \dots, K-1$ **do**
2: Evaluate $\nabla f(x^k)$, $\nabla^2 f(x^k)$
3: $x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$
4: **end for**

Output: x^K

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k)\rangle.$$

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k)\rangle.$$

We minimize the quadratic approximation on x:

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k) \rangle.$$

We minimize the quadratic approximation on x:
$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$.

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k)\rangle.$$

We minimize the quadratic approximation on x:
$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$. From where we get the next point of the method: $x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$.

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k)\rangle.$$

  We minimize the quadratic approximation on x:
  $\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$. From where we get the next point of the method: $x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$.

- Newton's method uses a second-order oracle: it requires the calculation of the Hessian.

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle x - x^k, \nabla^2 f(x^k)(x - x^k) \rangle.$$

  We minimize the quadratic approximation on x:
  $\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$. From where we get the next point of the method: $x^{k+1} = x^k - \left(\nabla^2 f(x^k) \right)^{-1} \nabla f(x^k)$.
- Newton's method uses a second-order oracle: it requires the calculation of the Hessian.
- The iteration cost increases significantly (compared to gradient descent) not only because of the hessian, but also its reversal.

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:
$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k)\rangle.$$
We minimize the quadratic approximation on x:
$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$. From where we get the next point of the method: $x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$.
- Newton's method uses a second-order oracle: it requires the calculation of the Hessian.
- The iteration cost increases significantly (compared to gradient descent) not only because of the hessian, but also its reversal.
Question: in how many iterations will Newton's method converge for a quadratic problem with a positive definite matrix?

Newton's method and gradient descent

- Gradient descent works with linear approximation at the current point, Newton's method — with quadratic:
$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k)\rangle.$$
We minimize the quadratic approximation on x:
$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0.$ From where we get the next point of the method: $\quad x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k).$
- Newton's method uses a second-order oracle: it requires the calculation of the Hessian.
- The iteration cost increases significantly (compared to gradient descent) not only because of the hessian, but also its reversal.
Question: in how many iterations will Newton's method converge for a quadratic problem with a positive definite matrix? in 1 (but expensive).

Newton's method: Convergence

- The fact that for a quadratic problem, Newton's method converges in 1 iteration suggests that for all its disadvantages (local convergence, high cost of iteration), the key advantage is the speed of convergence.

Newton's method: Convergence

- The fact that for a quadratic problem, Newton's method converges in 1 iteration suggests that for all its disadvantages (local convergence, high cost of iteration), the key advantage is the speed of convergence.
- Let the objective function in the unconditional minimization problem be twice continuously differentiable, μ-strongly convex and has M-Lipschitz Hessian, that is for any $x, y \in \mathbb{R}^d$ the following is true:

$$\nabla^2 f(x) \succeq \mu I, \quad \|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M\|x - y\|_2.$$

In the case of the matrix $\|\cdot\|_2$ is a spectral norm (a consistent norm with Euclidean for vectors).

Newton's method: Convergence

- Proof of the convergence

Newton's method: Convergence

- Proof of the convergence We will study how the distance to the solution changes:

$$x^{k+1} - x^* = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) - x^*.$$

Newton's method: Convergence

- Proof of the convergence We will study how the distance to the solution changes:

$$x^{k+1} - x^* = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) - x^*.$$

- Again, let's recall the Newton-Leibniz formula for the integral along the curve:

$$\nabla f(x^k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau$$

Newton's method: Convergence

- Proof of the convergence We will study how the distance to the solution changes:

$$x^{k+1} - x^* = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) - x^*.$$

- Again, let's recall the Newton-Leibniz formula for the integral along the curve:

$$\nabla f(x^k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau$$

With $\nabla f(x^*) = 0$, we receive

$$x^{k+1} - x^* = x^k - x^* - \left(\nabla^2 f(x^k)\right)^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau.$$

Newton's method: Convergence

- Continue and use the «smart 1»:

$$
\begin{aligned}
x^{k+1} - x^* =& x^k - x^* - \left(\nabla^2 f(x^k)\right)^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau \\
=& \left(\nabla^2 f(x^k)\right)^{-1} \nabla^2 f(x^k)(x^k - x^*) \\
&- \left(\nabla^2 f(x^k)\right)^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau.
\end{aligned}
$$

Newton's method: Convergence

- Continue and use the «smart 1»:

$$
\begin{aligned}
x^{k+1} - x^* =& x^k - x^* - \left(\nabla^2 f(x^k)\right)^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau \\
=& \left(\nabla^2 f(x^k)\right)^{-1} \nabla^2 f(x^k)(x^k - x^*) \\
& - \left(\nabla^2 f(x^k)\right)^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau.
\end{aligned}
$$

- Note that $x^k - x^*$ can be taken out of the integral:

$$
\begin{aligned}
x^{k+1} - x^* =& \left(\nabla^2 f(x^k)\right)^{-1} \nabla^2 f(x^k)(x^k - x^*) \\
& - \left(\nabla^2 f(x^k)\right)^{-1} \left(\int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*)) d\tau\right)(x^k - x^*).
\end{aligned}
$$

Newton's method: Convergence

- Let 's introduce the notation $G_k = \nabla^2 f(x^k) - \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))d\tau$:

$$x^{k+1} - x^* = \left(\nabla^2 f(x^k)\right)^{-1} G_k(x^k - x^*).$$

Newton's method: Convergence

- Let 's introduce the notation $G_k = \nabla^2 f(x^k) - \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))d\tau$:

$$x^{k+1} - x^* = \left(\nabla^2 f(x^k)\right)^{-1} G_k(x^k - x^*).$$

- Let's move on to estimating the distance norm:

$$\|x^{k+1} - x^*\|_2 = \left\|\left(\nabla^2 f(x^k)\right)^{-1} G_k(x^k - x^*)\right\|_2$$

Newton's method: Convergence

- Let 's introduce the notation $G_k = \nabla^2 f(x^k) - \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))d\tau$:

$$x^{k+1} - x^* = \left(\nabla^2 f(x^k)\right)^{-1} G_k(x^k - x^*).$$

- Let's move on to estimating the distance norm:

$$\|x^{k+1} - x^*\|_2 = \left\| \left(\nabla^2 f(x^k)\right)^{-1} G_k(x^k - x^*)\right\|_2$$

- We use that the spectral norm of the matrix is consistent with the Euclidean norm of vector:

$$\begin{aligned}\|x^{k+1} - x^*\|_2 &\le \left\| \left(\nabla^2 f(x^k)\right)^{-1} G_k\right\|_2 \|x^k - x^*\|_2 \\ &\le \left\| \left(\nabla^2 f(x^k)\right)^{-1}\right\|_2 \|G_k\|_2 \|x^k - x^*\|_2.\end{aligned}$$

Newton's method: Convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2 \leq \left\| \left(\nabla^2 f(x^k)\right)^{-1} \right\|_2 \|G_k\|_2 \|x^k - x^*\|_2.$$

Newton's method: Convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2 \leq \left\| \left(\nabla^2 f(x^k)\right)^{-1} \right\|_2 \|G_k\|_2 \|x^k - x^*\|_2.$$

- **Question:** How to estimate $\left\| \left(\nabla^2 f(x^k)\right)^{-1} \right\|_2$?

Newton's method: Convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2 \leq \left\| \left(\nabla^2 f(x^k) \right)^{-1} \right\|_2 \|G_k\|_2 \|x^k - x^*\|_2.$$

- **Question:** How to estimate $\left\| \left(\nabla^2 f(x^k) \right)^{-1} \right\|_2$? We know that $\nabla^2 f(x) \succeq \mu I$, it means $\frac{1}{\mu} I \succeq \left(\nabla^2 f(x^k) \right)^{-1}$,

Newton's method: Convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2 \leq \left\| \left(\nabla^2 f(x^k) \right)^{-1} \right\|_2 \|G_k\|_2 \|x^k - x^*\|_2.$$

- **Question:** How to estimate $\left\| \left(\nabla^2 f(x^k) \right)^{-1} \right\|_2$? We know that $\nabla^2 f(x) \succeq \mu I$, it means $\frac{1}{\mu} I \succeq \left(\nabla^2 f(x^k) \right)^{-1}$, from where $\left\| \left(\nabla^2 f(x^k) \right)^{-1} \right\|_2 \leq \frac{1}{\mu}$ and

$$\|x^{k+1} - x^*\|_2 \leq \frac{1}{\mu} \|G_k\|_2 \|x^k - x^*\|_2.$$

Newton's method: Convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2 \leq \left\| \left(\nabla^2 f(x^k)\right)^{-1} \right\|_2 \|G_k\|_2 \|x^k - x^*\|_2.$$

- **Question:** How to estimate $\left\| \left(\nabla^2 f(x^k)\right)^{-1} \right\|_2$? We know that $\nabla^2 f(x) \succeq \mu I$, it means $\frac{1}{\mu} I \succeq \left(\nabla^2 f(x^k)\right)^{-1}$, from where $\left\| \left(\nabla^2 f(x^k)\right)^{-1} \right\|_2 \leq \frac{1}{\mu}$ and

$$\|x^{k+1} - x^*\|_2 \leq \frac{1}{\mu} \|G_k\|_2 \|x^k - x^*\|_2.$$

- It remains to estimate $\|G_k\|_2$.

Newton's method: Convergence

- Estimate of $\|G_k\|_2$:

$$
\begin{aligned}
\|G_k\|_2 &= \left\| \nabla^2 f(x^k) - \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*)) d\tau \right\|_2 \\
&= \left\| \int_0^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*)) \right) d\tau \right\|_2 \\
&\leq \int_0^1 \left\| \nabla^2 f(x^k) - \nabla^2 f(x^* + \tau(x^k - x^*)) \right\|_2 d\tau \\
&\leq \int_0^1 M(1-\tau)\|x^k - x^*\|_2 d\tau \\
&= M\|x^k - x^*\|_2 \int_0^1 (1-\tau) d\tau = \frac{M}{2}\|x^k - x^*\|_2.
\end{aligned}
$$

Newton's method: Convergence

- We substitute the estimate for $\|G_k\|_2$:

$$\|x^{k+1} - x^*\|_2 \leq \frac{M}{2\mu}\|x^k - x^*\|_2^2.$$

Newton's method: Convergence

- We substitute the estimate for $\|G_k\|_2$:

$$\|x^{k+1} - x^*\|_2 \leq \frac{M}{2\mu}\|x^k - x^*\|_2^2.$$

Theorem on the convergence estimation of Newton's method for μ-strongly convex functions with M-Lipschitz Hessian

Let the problem of unconditional optimization with μ-strongly convex objective function f with M-Lipschitz Hessian be solved by Newton's method. Then the following convergence estimate for 1 iteration is valid

$$\|x^{k+1} - x^*\|_2 \leq \frac{M}{2\mu}\|x^k - x^*\|_2^2.$$

Newton's method: Convergence

- We substitute the estimate for $\|G_k\|_2$:

$$\|x^{k+1} - x^*\|_2 \leq \frac{M}{2\mu}\|x^k - x^*\|_2^2.$$

Theorem on the convergence estimation of Newton's method for μ-strongly convex functions with M-Lipschitz Hessian

Let the problem of unconditional optimization with μ-strongly convex objective function f with M-Lipschitz Hessian be solved by Newton's method. Then the following convergence estimate for 1 iteration is valid

$$\|x^{k+1} - x^*\|_2 \leq \frac{M}{2\mu}\|x^k - x^*\|_2^2.$$

We already know that such estimates give a quadratic convergence rate.

Newton's method: Convergence

- Convergence, as in the case of Newton's original method, is local.

Newton's method: Convergence

- Convergence, as in the case of Newton's original method, is local. Namely, to guarantee $\|x^1 - x^*\|_2 < \|x^0 - x^*\|_2$, we need to assume that
$$\|x^0 - x^*\|_2 < \frac{2\mu}{M}.$$

Newton's method: Convergence

- Convergence, as in the case of Newton's original method, is local. Namely, to guarantee $\|x^1 - x^*\|_2 < \|x^0 - x^*\|_2$, we need to assume that
$$\|x^0 - x^*\|_2 < \frac{2\mu}{M}.$$

- Let's understand how quickly the method converges. Let $M = 2$, $\mu = 1$, a $\|x^0 - x^*\|_2 = \frac{1}{2}$.

Newton's method: Convergence

- Convergence, as in the case of Newton's original method, is local. Namely, to guarantee $\|x^1 - x^*\|_2 < \|x^0 - x^*\|_2$, we need to assume that
$$\|x^0 - x^*\|_2 < \frac{2\mu}{M}.$$
- Let's understand how quickly the method converges. Let $M = 2$, $\mu = 1$, a $\|x^0 - x^*\|_2 = \frac{1}{2}$. Then we can guarantee that $\|x^1 - x^*\|_2 \leq \frac{1}{2^2}$,

Newton's method: Convergence

- Convergence, as in the case of Newton's original method, is local. Namely, to guarantee $\|x^1 - x^*\|_2 < \|x^0 - x^*\|_2$, we need to assume that
$$\|x^0 - x^*\|_2 < \frac{2\mu}{M}.$$
- Let's understand how quickly the method converges. Let $M = 2$, $\mu = 1$, a $\|x^0 - x^*\|_2 = \frac{1}{2}$. Then we can guarantee that $\|x^1 - x^*\|_2 \leq \frac{1}{2^2}$, $\|x^2 - x^*\|_2 \leq \frac{1}{(2^2)^2}$ and so on..

Newton's method: Modifications

- We are trying to solve the problem of local convergence. We act by analogy with gradient descent. **Question:** Any ideas?

Newton's method: Modifications

- We are trying to solve the problem of local convergence. We act by analogy with gradient descent. **Question:** Any ideas?
- Idea number 1 – step:

$$x^{k+1} = x^k - \gamma_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k).$$

This method is called the damped Newton method.

Newton's method: Modifications

- We are trying to solve the problem of local convergence. We act by analogy with gradient descent. **Question:** Any ideas?
- Idea number 1 – step:

$$x^{k+1} = x^k - \gamma_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k).$$

This method is called the damped Newton method. **Question:** Which γ_k to take?

Newton's method: Modifications

- We are trying to solve the problem of local convergence. We act by analogy with gradient descent. **Question:** Any ideas?
- Idea number 1 – step:

$$x^{k+1} = x^k - \gamma_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k).$$

This method is called the damped Newton method. **Question:** Which γ_k to take? There are many different ways, for example, linear search: $\arg\min_\gamma f(x^k + \gamma p_k)$, где $p_k = -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$.

Newton's method: Modifications

- The second idea is «upper-bound estimates». The analysis of gradient descent was based on the optimization of «upper-bound estimates» on the function:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \|x - x^k\|_2^2 \right).$$

Newton's method: Modifications

- The second idea is «upper-bound estimates». The analysis of gradient descent was based on the optimization of «upper-bound estimates» on the function:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \|x - x^k\|_2^2 \right).$$

Question: what is x^{k+1} equal to?

Newton's method: Modifications

- The second idea is «upper-bound estimates». The analysis of gradient descent was based on the optimization of «upper-bound estimates» on the function:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|_2^2 \right).$$

Question: what is x^{k+1} equal to? $x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$.

Newton's method: Modifications

- The second idea is «upper-bound estimates». The analysis of gradient descent was based on the optimization of «upper-bound estimates» on the function:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|_2^2 \right).$$

Question: what is x^{k+1} equal to? $x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$. Let's write something similar for the 2nd order approximation:

$$\begin{aligned} x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \Big(& f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \\ & + \frac{1}{2}\langle x - x^k, \nabla^2 f(x^k)(x - x^k) \rangle + \frac{M}{6}\|x - x^k\|_2^3 \Big). \end{aligned}$$

Here M is the Lipschitz constant of the Hessian. This method is called the cubic Newton method.

Quasi-Newton equation

- Let 's write down Newton 's method as follows:

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

Quasi-Newton equation

- Let 's write down Newton 's method as follows:

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

In the case of Newton's method, instead of H_k is $\left(\nabla^2 f(x^k)\right)^{-1}$.

- We want to replace $\left(\nabla^2 f(x^k)\right)^{-1}$ with something cheaper from the point of view of calculations.

Quasi-Newton equation

- Let 's write down Newton 's method as follows:

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

In the case of Newton's method, instead of H_k is $\left(\nabla^2 f(x^k)\right)^{-1}$.
- We want to replace $\left(\nabla^2 f(x^k)\right)^{-1}$ with something cheaper from the point of view of calculations.
- The idea is to extract some properties inherent in the Hessian.

Quasi-Newton equation

- Let 's write down Newton 's method as follows:

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

In the case of Newton's method, instead of H_k is $\left(\nabla^2 f(x^k)\right)^{-1}$.
- We want to replace $\left(\nabla^2 f(x^k)\right)^{-1}$ with something cheaper from the point of view of calculations.
- The idea is to extract some properties inherent in the Hessian.
- Relation of gradient and Hessian:

$$\nabla f(x^k) = \nabla f(x^{k+1}) + \nabla^2 f(x^{k+1})(x^k - x^{k+1}) + o(\|x^{k+1} - x^k\|_2)$$

or $\nabla f(x^k) - \nabla f(x^{k+1}) \approx \nabla^2 f(x^{k+1})(x^k - x^{k+1})$. From where $x^{k+1} - x^k \approx (\nabla^2 f(x^{k+1}))^{-1}(\nabla f(x^{k+1}) - \nabla f(x^k))$.

Quasi-Newton equation

- Let 's write down Newton 's method as follows:

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

In the case of Newton's method, instead of H_k is $\left(\nabla^2 f(x^k)\right)^{-1}$.

- We want to replace $\left(\nabla^2 f(x^k)\right)^{-1}$ with something cheaper from the point of view of calculations.
- The idea is to extract some properties inherent in the Hessian.
- Relation of gradient and Hessian:

$$\nabla f(x^k) = \nabla f(x^{k+1}) + \nabla^2 f(x^{k+1})(x^k - x^{k+1}) + o(\|x^{k+1} - x^k\|_2)$$

or $\nabla f(x^k) - \nabla f(x^{k+1}) \approx \nabla^2 f(x^{k+1})(x^k - x^{k+1})$. From where $x^{k+1} - x^k \approx (\nabla^2 f(x^{k+1}))^{-1}(\nabla f(x^{k+1}) - \nabla f(x^k))$. Let's change $(\nabla^2 f(x^{k+1}))^{-1}$ to H_{k+1}, introduce the notation $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$:

$$s^k = H_{k+1} y^k$$

Quasi-Newton equation

- Quasi-Newton equation:

$$\boxed{s^k = H_{k+1} y^k}$$

Quasi-Newton equation

- Quasi-Newton equation:

$$\boxed{s^k = H_{k+1} y^k}$$

- We also require that H_{k+1} be symmetric: $H_{k+1}^T = H_{k+1}$.

Quasi-Newton equation

- Quasi-Newton equation:

$$\boxed{s^k = H_{k+1} y^k}$$

- We also require that H_{k+1} be symmetric: $H_{k+1}^T = H_{k+1}$.
- **Question:** how many solutions does the system of equations have $s^k = H_{k+1} y^k$ relative to H_{k+1}, provided that $H_{k+1}^T = H_{k+1}$?

Quasi-Newton equation

- Quasi-Newton equation:

$$\boxed{s^k = H_{k+1} y^k}$$

- We also require that H_{k+1} be symmetric: $H_{k+1}^T = H_{k+1}$.
- **Question:** how many solutions does the system of equations have $s^k = H_{k+1} y^k$ relative to H_{k+1}, provided that $H_{k+1}^T = H_{k+1}$? d variables, $d + d(d-1)/2$ equations. You can surely solve it.

Quasi-Newton equation

- Quasi-Newton equation:

$$\boxed{s^k = H_{k+1} y^k}$$

- We also require that H_{k+1} be symmetric: $H_{k+1}^T = H_{k+1}$.
- **Question:** how many solutions does the system of equations have $s^k = H_{k+1} y^k$ relative to H_{k+1}, provided that $H_{k+1}^T = H_{k+1}$?
  d variables, $d + d(d-1)/2$ equations. You can surely solve it. We also need to narrow down the search rules of H_{k+1}.

Quasi-Newton methods: SR1/Broyden

- The first idea is a 1-rank (computationally cheap) additive:

$$H_{k+1} = H_k + \mu_k q^k (q^k)^T,$$

where $\mu_k \in \mathbb{R}$ and $q^k \in \mathbb{R}^d$ should be selected.

Quasi-Newton methods: SR1/Broyden

- The first idea is a 1-rank (computationally cheap) additive:

$$H_{k+1} = H_k + \mu_k q^k (q^k)^T,$$

where $\mu_k \in \mathbb{R}$ and $q^k \in \mathbb{R}^d$ should be selected.

- We select based on the quasi-Newtonian equation:

$$\begin{aligned} s^k =& H_{k+1} y^k = H_k y^k + \mu_k q^k (q^k)^T y^k \\ =& H_k y^k + \mu_k \left((q^k)^T y^k \right) q^k \end{aligned}$$

Quasi-Newton methods: SR1/Broyden

- The first idea is a 1-rank (computationally cheap) additive:

$$H_{k+1} = H_k + \mu_k q^k (q^k)^T,$$

where $\mu_k \in \mathbb{R}$ and $q^k \in \mathbb{R}^d$ should be selected.

- We select based on the quasi-Newtonian equation:

$$\begin{aligned} s^k =& H_{k+1} y^k = H_k y^k + \mu_k q^k (q^k)^T y^k \\ =& H_k y^k + \mu_k \left((q^k)^T y^k \right) q^k \end{aligned}$$

From where

$$\mu_k \left((q^k)^T y^k \right) q^k = s^k - H_k y^k$$

Quasi-Newton methods: SR1/Broyden

- From the previous slide:

$$\mu_k \left((q^k)^T y^k \right) q^k = s^k - H_k y^k$$

- **Question:** what can we say about the vector q^k?

Quasi-Newton methods: SR1/Broyden

- From the previous slide:

$$\mu_k \left((q^k)^T y^k \right) q^k = s^k - H_k y^k$$

- **Question:** what can we say about the vector q^k?
  Collinear to $s^k - H_k y^k$.

Quasi-Newton methods: SR1/Broyden

- From the previous slide:

$$\mu_k \left((q^k)^T y^k \right) q^k = s^k - H_k y^k$$

- **Question:** what can we say about the vector q^k? Collinear to $s^k - H_k y^k$. Consider

$$q^k = s^k - H_k y^k,$$

then

$$\mu_k = \frac{1}{(q^k)^T y^k}.$$

- We get the SR1 method of counting matrices H:

$$\boxed{H_{k+1} = H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^T}{(s^k - H_k y^k)^T y^k}}$$

Quasi-Newton methods: BFGS

- Let's look at the search problem H_{k+1} as a search problem «close» to H_k matrix from the point of view of optimization:

$$
\begin{aligned}
H_{k+1} = \arg \min_{H \in \mathbb{R}^{d \times d}} \|H - H_k\|^2 \\
s.t.\ s^k = Hy^k \\
H^T = H
\end{aligned}
$$

Quasi-Newton methods: BFGS

- Let's look at the search problem H_{k+1} as a search problem «close» to H_k matrix from the point of view of optimization:

$$H_{k+1} = \arg\min_{H \in \mathbb{R}^{d\times d}} \|H - H_k\|^2$$
$$s.t.\ s^k = Hy^k$$
$$H^T = H$$

- The norm in the optimization problem can be any. Depending on the norm, different quasi-Newtonian methods will be obtained.

Quasi-Newton methods: BFGS

- Let's look at the search problem H_{k+1} as a search problem «close» to H_k matrix from the point of view of optimization:

$$H_{k+1} = \arg \min_{H \in \mathbb{R}^{d \times d}} \|H - H_k\|^2$$
$$s.t.\ s^k = Hy^k$$
$$H^T = H$$

- The norm in the optimization problem can be any. Depending on the norm, different quasi-Newtonian methods will be obtained.
- Consider the weighted Frobenius norm $\|A\|_W = \|W^{1/2} A W^{1/2}\|_F$, where $Wy^k = s^k$ should be executed. This choice is given by the BFGS method:

$$\boxed{H_{k+1} = (I - \rho_k s^k (y^k)^T) H_k (I - \rho_k y^k (s^k)^T) + \rho_k s^k (s^k)^T,\ \rho_k = \frac{1}{(y^k)^T s^k}}$$

Quasi-Newton methods: BFGS

- There is another way to reach such a formula.

Quasi-Newton methods: BFGS

- There is another way to reach such a formula. Consider $B_{k+1} = H_{k+1}^{-1}$. For B, the quasi-Newtonian equation looks like
$$B_{k+1} s^k = y^k$$

Quasi-Newton methods: BFGS

- There is another way to reach such a formula. Consider $B_{k+1} = H_{k+1}^{-1}$. For B, the quasi-Newtonian equation looks like
$$B_{k+1} s^k = y^k$$
- For B_{k+1}, you can write SR1 matrix recalculation:
$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$$

Quasi-Newton methods: BFGS

- There is another way to reach such a formula. Consider $B_{k+1} = H_{k+1}^{-1}$. For B, the quasi-Newtonian equation looks like

$$B_{k+1}s^k = y^k$$

- For B_{k+1}, you can write SR1 matrix recalculation:

$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$$

- We look at the form B_{k+1} and make a two-rank change from it:

$$B_{k+1} = B_k + \mu_{k,1} y^k (y^k)^T + \mu_{k,2} B_k y^k (B_k y^k)^T$$

Quasi-Newton methods: BFGS

- There is another way to reach such a formula. Consider $B_{k+1} = H_{k+1}^{-1}$. For B, the quasi-Newtonian equation looks like
$$B_{k+1}s^k = y^k$$
- For B_{k+1}, you can write SR1 matrix recalculation:
$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$$
- We look at the form B_{k+1} and make a two-rank change from it:
$$B_{k+1} = B_k + \mu_{k,1} y^k (y^k)^T + \mu_{k,2} B_k y^k (B_k y^k)^T$$
- As in SR1, you can adjust $\mu_{k,1}$ and $\mu_{k,2}$:
$$B_{k+1} = B_k + \frac{y^k (y^k)^T}{(y^k)^T s^k} + \frac{B_k y^k (B_k y^k)^T}{(s^k)^T B_k s^k}$$

Quasi-Newton methods: BFGS

- There is another way to reach such a formula. Consider $B_{k+1} = H_{k+1}^{-1}$. For B, the quasi-Newtonian equation looks like
$$B_{k+1} s^k = y^k$$
- For B_{k+1}, you can write SR1 matrix recalculation:
$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$$
- We look at the form B_{k+1} and make a two-rank change from it:
$$B_{k+1} = B_k + \mu_{k,1} y^k (y^k)^T + \mu_{k,2} B_k y^k (B_k y^k)^T$$
- As in SR1, you can adjust $\mu_{k,1}$ and $\mu_{k,2}$:
$$B_{k+1} = B_k + \frac{y^k (y^k)^T}{(y^k)^T s^k} + \frac{B_k y^k (B_k y^k)^T}{(s^k)^T B_k s^k}$$
- If we now reverse B_{k+1} (the Sherman-Marrison-Woodberry formula), we get an expression for H_{k+1}

Quasi-Newton methods: BFGS

- **Question:** to calculate a new matrix, you need $O(d^2)$ operations (not counting gradients). It seems that BFGS counting is more expensive (there is a multiplication of three matrices). Is it so?
  $H_{k+1} = (I - \rho_k s^k (y^k)^T) H_k (I - \rho_k y^k (s^k)^T) + \rho_k s^k (s^k)^T$

Quasi-Newton methods: BFGS

- **Question:** to calculate a new matrix, you need $O(d^2)$ operations (not counting gradients). It seems that BFGS counting is more expensive (there is a multiplication of three matrices). Is it so?
  $H_{k+1} = (I - \rho_k s^k (y^k)^T) H_k (I - \rho_k y^k (s^k)^T) + \rho_k s^k (s^k)^T$
- You need to open the brackets in matrix multiplication. In the calculation of $s^k (y^k)^T H_k$ we must first multiply $(y^k)^T H_k$, and then vector by vector. Similarly for $H_k y^k (s^k)^T$. It turns out that the complexity of BFGS is $O(d^2)$ operations (not counting gradients).

Quasi-Newton methods: BFGS

- **Question:** to calculate a new matrix, you need $O(d^2)$ operations (not counting gradients). It seems that BFGS counting is more expensive (there is a multiplication of three matrices). Is it so?
  $H_{k+1} = (I - \rho_k s^k (y^k)^T) H_k (I - \rho_k y^k (s^k)^T) + \rho_k s^k (s^k)^T$
- You need to open the brackets in matrix multiplication. In the calculation of $s^k (y^k)^T H_k$ we must first multiply $(y^k)^T H_k$, and then vector by vector. Similarly for $H_k y^k (s^k)^T$. It turns out that the complexity of BFGS is $O(d^2)$ operations (not counting gradients).
- When initializing, it is enough to take the matrix H_0 equal to I. There are more tricky ways, but it doesn't feel much difference, everything works well.

Newton's method and quasi-Newtonian methods

- Quasi-Newtonian methods do not require Hessian counting and its reversal. The complexity of all arithmetic operations in one iteration is $O(d^2)$, which is cheaper than the Hessian conversion for $O(d^3)$.

Newton's method and quasi-Newtonian methods

- Quasi-Newtonian methods do not require Hessian counting and its reversal. The complexity of all arithmetic operations in one iteration is $O(d^2)$, which is cheaper than the Hessian conversion for $O(d^3)$.
- Quasi-Newtonian methods have a global superlinear convergence rate. This is slower than Newton's method, but you don't need a «good» neighborhood of the solution.

Newton's method and quasi-Newtonian methods

- Quasi-Newtonian methods do not require Hessian counting and its reversal. The complexity of all arithmetic operations in one iteration is $O(d^2)$, which is cheaper than the Hessian conversion for $O(d^3)$.
- Quasi-Newtonian methods have a global superlinear convergence rate. This is slower than Newton's method, but you don't need a «good» neighborhood of the solution.
- Quasi-Newtonian methods use only a gradient, but in theory converge faster than the accelerated gradient method. **Question:** why is this so, because the Nesterov method is optimal?

Newton's method and quasi-Newtonian methods

- Quasi-Newtonian methods do not require Hessian counting and its reversal. The complexity of all arithmetic operations in one iteration is $O(d^2)$, which is cheaper than the Hessian conversion for $O(d^3)$.
- Quasi-Newtonian methods have a global superlinear convergence rate. This is slower than Newton's method, but you don't need a «good» neighborhood of the solution.
- Quasi-Newtonian methods use only a gradient, but in theory converge faster than the accelerated gradient method. **Question:** why is this so, because the Nesterov method is optimal? See the definitions of the class of problems for which the Nesterov method is optimal: vector products are not allowed.

Newton's method and quasi-Newtonian methods

- Quasi-Newtonian methods do not require Hessian counting and its reversal. The complexity of all arithmetic operations in one iteration is $O(d^2)$, which is cheaper than the Hessian conversion for $O(d^3)$.
- Quasi-Newtonian methods have a global superlinear convergence rate. This is slower than Newton's method, but you don't need a «good» neighborhood of the solution.
- Quasi-Newtonian methods use only a gradient, but in theory converge faster than the accelerated gradient method. **Question:** why is this so, because the Nesterov method is optimal? See the definitions of the class of problems for which the Nesterov method is optimal: vector products are not allowed.
- Newton's method and quasi-Newtonian methods in practice quickly find the exact local minimum. They can be safely used as «finish-solvers». Quasi-Newtonian methods as a «starting» method.