Gradient Descent. Convergence rates

Daniil Merkulov

Optimization methods. MIPT

Previously

- Gradient Descent

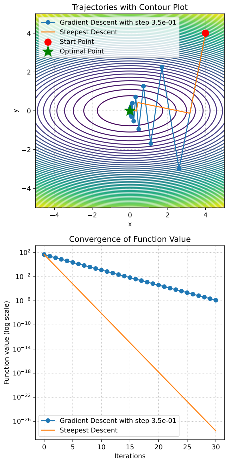


Figure 1: Steepest Descent

Open In Colab ♣

Previously

- Gradient Descent
- Steepest descent

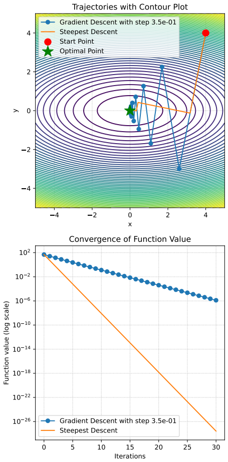


Figure 1: Steepest Descent

Open In Colab ♣

Previously

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

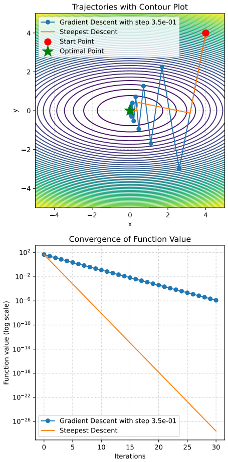


Figure 1: Steepest Descent

Open In Colab ♣

Previously

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

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

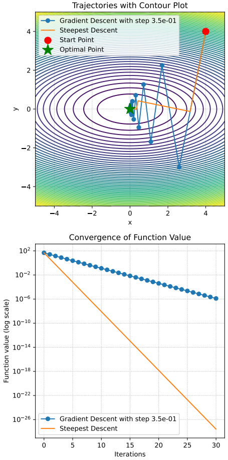


Figure 1: Steepest Descent

Open In Colab ♣

Previously

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

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

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

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

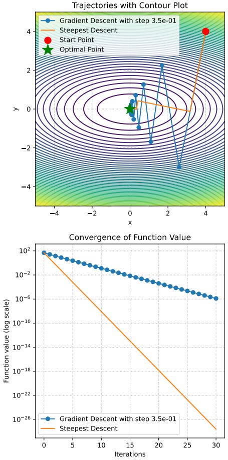


Figure 1: Steepest Descent

Open In Colab ♣

Convergence rates

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

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

Coordinate shift for strongly convex quadratics

Consider the following quadratic optimization problem:

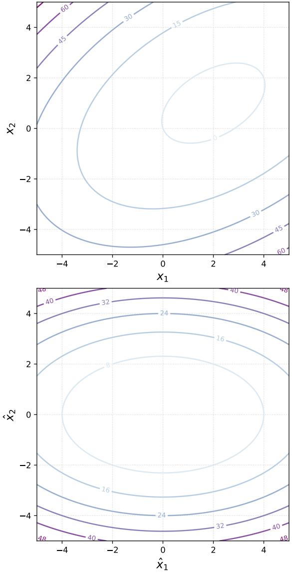
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

Coordinate shift for strongly convex quadratics

Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.



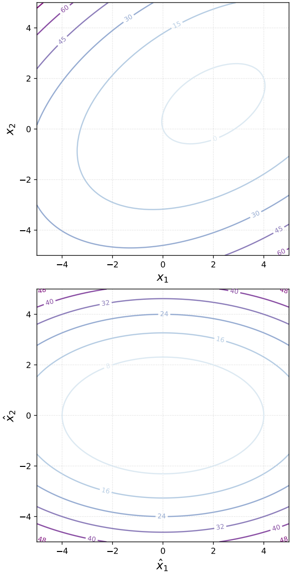
Coordinate shift for strongly convex quadratics

Consider the following quadratic optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{2}x^\top Ax - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q\Lambda Q^T$$



Coordinate shift for strongly convex quadratics

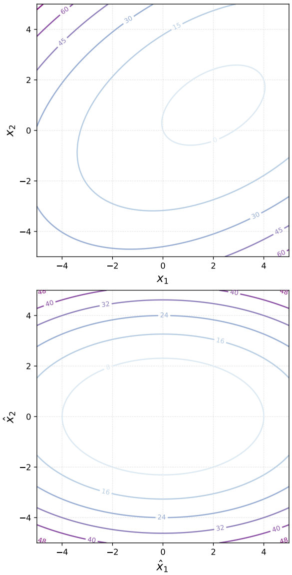
Consider the following quadratic optimization problem:

$$
\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.
$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$
A = Q \Lambda Q^T
$$

- Let's show, that we can switch coordinates in order to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.



Coordinate shift for strongly convex quadratics

Consider the following quadratic optimization problem:

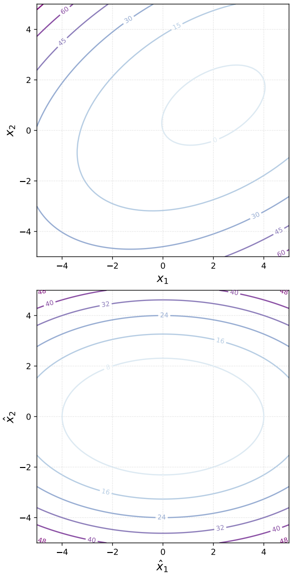
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q \Lambda Q^T$$

- Let's show, that we can switch coordinates in order to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.

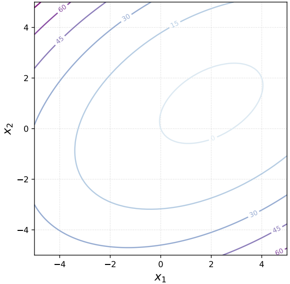
$$f(\hat{x}) = \frac{1}{2}(Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*)$$



Coordinate shift for strongly convex quadratics

Consider the following quadratic optimization problem:

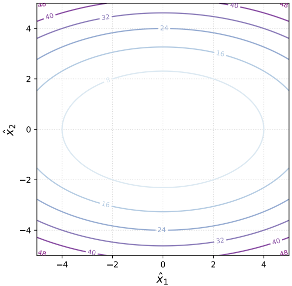
$$
\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.
$$



- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$
A = Q \Lambda Q^T
$$

- Let's show, that we can switch coordinates in order to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.



$$
\begin{aligned}
f(\hat{x}) &= \frac{1}{2}(Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\
&= \frac{1}{2} \hat{x}^T Q^T A Q \hat{x} + (x^*)^T A Q \hat{x} + \frac{1}{2} (x^*)^T A (x^*)^T - b^T Q \hat{x} - b^T x^*
\end{aligned}
$$

Coordinate shift for strongly convex quadratics

Consider the following quadratic optimization problem:

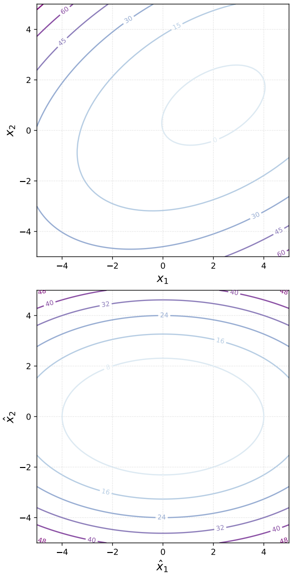
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q \Lambda Q^T$$

- Let's show, that we can switch coordinates in order to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.

$$\begin{aligned}
f(\hat{x}) &= \frac{1}{2}(Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\
&= \frac{1}{2}\hat{x}^T Q^T A Q \hat{x} + (x^*)^T A Q \hat{x} + \frac{1}{2}(x^*)^T A (x^*)^T - b^T Q \hat{x} - b^T x^* \\
&= \frac{1}{2}\hat{x}^T \Lambda \hat{x}
\end{aligned}$$



Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

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

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T \Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

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

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate}
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

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

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1 \\
&\alpha < \frac{2}{\mu} \qquad \alpha\mu > 0
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T \Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1 \\
&\alpha < \frac{2}{\mu} \qquad \alpha\mu > 0
\end{aligned}
\qquad\qquad
|1 - \alpha L| < 1
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 - \alpha\mu &< 1 \\
\alpha < \frac{2}{\mu} \quad\quad & \alpha\mu > 0
\end{aligned}
\qquad\qquad
\begin{aligned}
|1 - \alpha L| &< 1 \\
-1 < 1 - \alpha L &< 1
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1 \\
&\alpha < \frac{2}{\mu} \quad\quad \alpha\mu > 0
\end{aligned}
\qquad\qquad
\begin{aligned}
&|1 - \alpha L| < 1 \\
&-1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{L} \quad\quad \alpha L > 0
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \qquad \alpha\mu > 0 && \alpha < \frac{2}{L} \qquad \alpha L > 0
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad\quad \alpha\mu > 0 && \alpha < \frac{2}{L} \quad\quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1 \\
&\alpha < \frac{2}{\mu} \quad\quad \alpha\mu > 0
\end{aligned}
\qquad\qquad
\begin{aligned}
&|1 - \alpha L| < 1 \\
&-1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{L} \quad\quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\rho^* = \min_\alpha \rho(\alpha)
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 - \alpha\mu &< 1 \\
\alpha < \frac{2}{\mu} \quad & \quad \alpha\mu > 0
\end{aligned}
\qquad\qquad
\begin{aligned}
|1 - \alpha L| &< 1 \\
-1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{L} \quad & \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\rho^* = \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}|
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad \alpha\mu > 0 && \alpha < \frac{2}{L} \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\
&= \min_{\alpha} \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\}
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 - \alpha\mu &< 1 \\
\alpha < \frac{2}{\mu} \quad &\quad \alpha\mu > 0
\end{aligned}
\qquad
\begin{aligned}
|1 - \alpha L| &< 1 \\
-1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{L} \quad &\quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha \lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \\
\alpha^* :& \quad 1 - \alpha^* \mu = \alpha^* L - 1
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 & &|1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 & &-1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \qquad \alpha\mu > 0 & &\alpha < \frac{2}{L} \qquad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\
&= \min_{\alpha} \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \\
\alpha^* :& \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L}
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 & &|1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 & &-1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad \alpha\mu > 0 & &\alpha < \frac{2}{L} \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \{|1 - \alpha\mu|, |1 - \alpha L|\}
\end{aligned}
$$

$$
\alpha^* : \quad 1 - \alpha^*\mu = \alpha^* L - 1
$$

$$
\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 & |1 - \alpha L| &< 1 \\
-1 < 1 - \alpha\mu &< 1 & -1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{\mu} \quad & \alpha\mu > 0 & \alpha < \frac{2}{L} \quad & \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \\
\alpha^* &: \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu} \\
x^{k+1} &= \left(\frac{L - \mu}{L + \mu} \right)^k x^0
\end{aligned}
$$

Strongly convex quadratics

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \qquad & |1 - \alpha L| &< 1 \\
-1 < 1 - \alpha\mu &< 1 & -1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{\mu} \qquad \alpha\mu &> 0 & \alpha < \frac{2}{L} \qquad \alpha L &> 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to choose α in order to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \\
\alpha^* &: \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu} \\
x^{k+1} &= \left(\frac{L - \mu}{L + \mu} \right)^k x^0 \quad f(x^{k+1}) = \left(\frac{L - \mu}{L + \mu} \right)^{2k} f(x^0)
\end{aligned}
$$

Strongly convex quadratics

So, we have a linear convergence in domain with rate $\frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$, where $\kappa = \frac{L}{\mu}$ is sometimes called *condition number* of the quadratic problem.

κ	ρ	Iterations to decrease domain gap 10 times	Iterations to decrease function gap 10 times
1.1	0.05	1	1
2	0.33	3	2
5	0.67	6	3
10	0.82	12	6
50	0.96	58	29
100	0.98	116	58
500	0.996	576	288
1000	0.998	1152	576

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

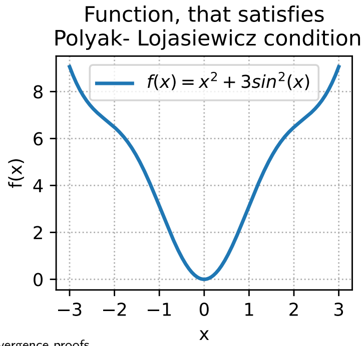
PL inequality holds if the following condition is satisfied for some $\mu > 0$,

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

It is interesting, that Gradient Descent algorithm has

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

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



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

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

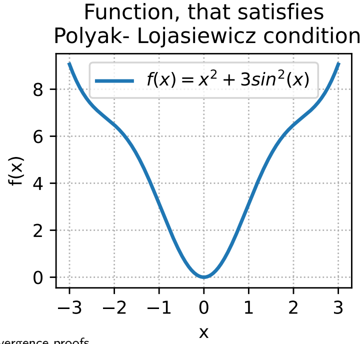
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

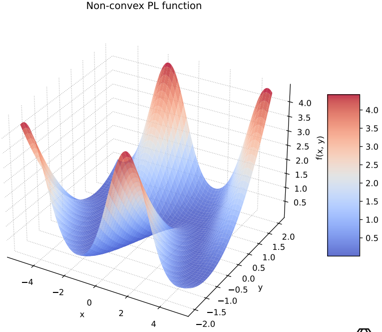
It is interesting, that Gradient Descent algorithm has

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

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

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





Gradient Descent convergence. Polyak-Lojasiewicz case

Theorem

Consider the Problem

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

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

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

Gradient Descent convergence. Polyak-Lojasiewicz case

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

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

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

Gradient Descent convergence. Polyak-Lojasiewicz case

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

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

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

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

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

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

Gradient Descent convergence. Smooth convex case

Theorem

Consider the Problem

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

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

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

Gradient Descent convergence. Smooth convex case

Gradient Descent convergence. Smooth μ-strongly convex case

Theorem

Consider the Problem

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

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

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

Gradient Descent convergence. Smooth μ-strongly convex case

Gradient Descent for Linear Least Squares aka Linear Regression

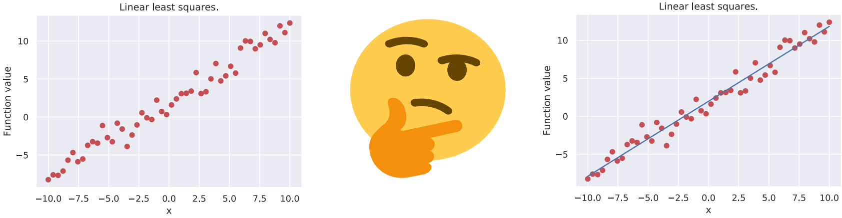


Figure 4: Illustration

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

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

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

Linear Least Squares aka Linear Regression [1]

1. Is this problem convex? Strongly convex?

Linear Least Squares aka Linear Regression [1]

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

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

l_2-regularized Linear Least Squares

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

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

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

Take a look at the code