Gradient descent on Ordinary Least Squares

Samuel Vaiter

Created: 2023-11-08.
Last update: 2023-11-08.
Status: inprogress.

GD on OLS: Ordinary Least Squares

Let $n, m \geq 1$, $A \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$.

Ordinary Least Squares (OLS)

$$x^\star \in \arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n} \|Ax - y\|_2^2. \tag{1}$$

Gradient

$$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p,$$

Hessian matrix (constant) $\rightarrow$ covariance matrix

$$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}.$$

First order condition: $Hx^\star = \frac{1}{n} A^T y$

GD on OLS: Convergence in value and in iterates

Gradient descent

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

(OLS)	$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n} \|Ax - y\|_2^2$
Gradient	$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p$
Hessian matrix	$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}$
Minimizers	$Hx^\star = \frac{1}{n} A^T y$

Convergence in norms of the iterates. Distance to a minimizer:

$$r^{(t)} = \|x^{(t)} - x^\star\| \leq \varepsilon$$

Convergence in value. Infimum value $f^* = \inf_{x \in \mathbb{R}^n} f(x)$

$$r^{(t)} = f(x^{(t)}) - f^* \leq \varepsilon$$

GD on OLS: Explicit expression

Gradient descent

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

(OLS)	$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n} \|Ax - y\|_2^2$
Gradient	$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p$
Hessian matrix	$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}$
Minimizers	$Hx^\star = \frac{1}{n} A^T y$

Proposition

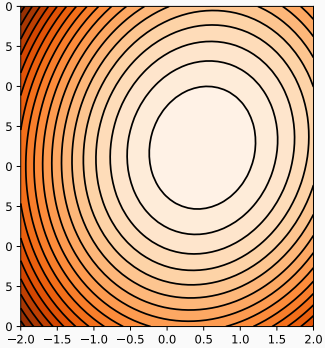
For all $t \in \mathbb{N}$, we have

$$x^{(t)} - x^\star = (I - \eta H)^t (x^{(0)} - x^\star). \tag{2}$$

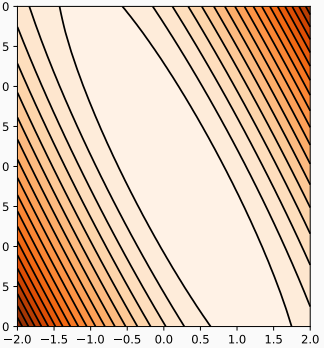
Proof idea: show that $x^{(t)} - x^\star = (I - \eta H)(x^{(t-1)} - x^\star)$

GD on OLS: Geometrical insight

μ smallest eigenvalue of H, L biggest eigenvalue of $H = \frac{1}{n}A^\top A$, $0 \leq \mu \leq L$



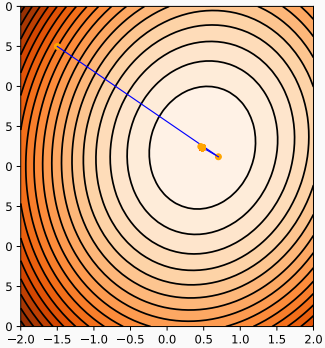
$$\kappa = \frac{L}{\mu} \approx 1$$

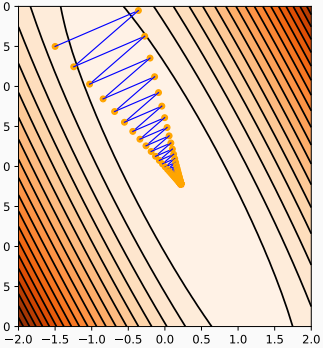


$$\kappa = \frac{L}{\mu} \gg 1$$

GD on OLS: Geometrical insight

μ smallest eigenvalue of H, L biggest eigenvalue of $H = \frac{1}{n}A^\top A$, $0 \leq \mu \leq L$





$$\kappa = \frac{L}{\mu} \approx 1$$

$$\kappa = \frac{L}{\mu} \gg 1$$

GD on OLS: Convergence in norm

Closed-form expression

$$x^{(t)} - x^\star = (\mathrm{Id} - \eta H)^t (x^{(0)} - x^\star)$$

(OLS)	$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n}\|Ax - y\|_2^2$
Gradient	$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p$
Hessian matrix	$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}$
Minimizers	$Hx^\star = \frac{1}{n} A^T y$

Distance to a minimizer

$$\|x^{(t)} - x^\star\|^2 = \langle x^{(0)} - x^\star, (\mathrm{Id} - \eta H)^{2t} (x^{(0)} - x^\star) \rangle$$

Eigenvalues of $(\mathrm{Id} - \eta H)^{2t}$

$$\mathrm{eigen}(\mathrm{Id} - \eta H)^{2t} = \left\{ (1 - \eta\lambda)^{2t} \ : \ \lambda \in \mathrm{eigen}(H) \right\}$$

$$\Downarrow \text{ (using the fact } \lambda \in [\mu, L])$$

$$\rho \in \mathrm{eigen}(\mathrm{Id} - \eta H)^{2t} \implies |\rho| \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t}$$

GD on OLS: Convergence in norm

Closed-form expression

$$x^{(t)} - x^\star = (\mathrm{Id} - \eta H)^t (x^{(0)} - x^\star)$$

(OLS)	$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n}\|Ax - y\|_2^2$
Gradient	$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p$
Hessian matrix	$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}$
Minimizers	$Hx^\star = \frac{1}{n} A^T y$

Upper bounds for the distance to a minimizer

$$\|x^{(t)} - x^\star\|^2 = \langle x^{(0)} - x^\star, (\mathrm{Id} - \eta H)^{2t} (x^{(0)} - x^\star) \rangle$$

$$+$$

$$\rho \in \mathrm{eigen}(\mathrm{Id} - \eta H)^{2t} \implies |\rho| \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t}$$

$$\Downarrow \text{ (spectral bound)}$$

$$\|x^{(t)} - x^\star\|^2 \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t} \|x^{(0)} - x^\star\|_2^2.$$

GD on OLS: Convergence in norm

Closed-form expression

$$x^{(t)} - x^\star = (\mathrm{Id} - \eta H)^t (x^{(0)} - x^\star)$$

(OLS)	$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n}\|Ax - y\|_2^2$
Gradient	$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p$
Hessian matrix	$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}$
Minimizers	$Hx^\star = \frac{1}{n} A^T y$

Upper bounds for the distance to a minimizer

$$\|x^{(t)} - x^\star\|^2 = \langle x^{(0)} - x^\star, (\mathrm{Id} - \eta H)^{2t} (x^{(0)} - x^\star) \rangle$$

$$+$$

$$\rho \in \mathrm{eigen}(\mathrm{Id} - \eta H)^{2t} \implies |\rho| \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t}$$

$$\Downarrow \text{ (spectral bound)}$$

$$\|x^{(t)} - x^\star\|^2 \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t} \|x^{(0)} - x^\star\|_2^2.$$

What if $\mu = 0$?

GD on OLS: Convergence in norm

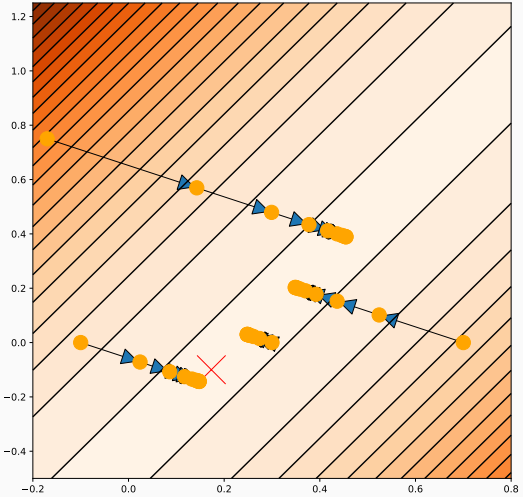
Case $\mu = 0$
$\Leftrightarrow H$ is non-invertible.

Eigenvector v associated to $\mu = 0$

$$\exists v \in \mathbb{R}^p, \quad (I - \eta H)v = v$$

Choose $x^\star$ a minimizer
Initialization $x^{(0)} = v + x^\star$.

$$\begin{aligned}
&\|x^{(t)} - x^\star\|^2 \\
=&\langle v + x^\star - x^\star, (I - \eta H)^{2t}(v + x^\star - x^\star)\rangle \\
=&\langle v, (I - \eta H)^{2t} v\rangle \\
=&\|v\|_2^2
\end{aligned}$$



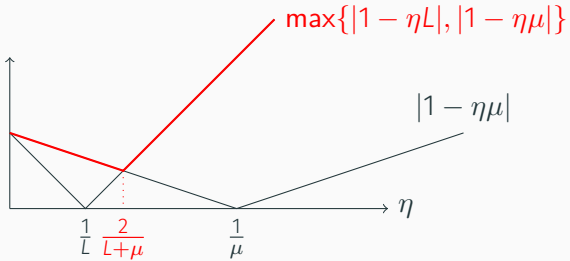
GD on OLS: Convergence in norm

Case $\mu \neq 0$

$$\|x^{(t)} - x^\star\|^2 \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda|\right)^{2t} \|x^{(0)} - x^\star\|_2^2.$$

Optimal step-size: solving

$$\min_{\eta > 0} \max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \implies \eta^\star = \frac{2}{\mu + L} \quad \text{and} \quad \max_{\lambda \in [\mu, L]} |1 - \eta^\star \lambda| = \frac{\kappa - 1}{\kappa + 1} \in (0, 1)$$



GD on OLS: Convergence in norm

Case $\mu \neq 0$

$$\|x^{(t)} - x^\star\|^2 \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t} \|x^{(0)} - x^\star\|_2^2.$$

Choice independant from μ: $\bar{\eta} = 1/L$

$$\max_{\lambda \in [\mu, L]} |1 - \bar{\eta}\lambda| = 1 - \frac{\mu}{L} = 1 - \frac{1}{\kappa} \in (0, 1)$$

GD on OLS: Convergence in norm

Proposition

Assume that $\min_x \|Ax - b\|^2$ *has a unique minimizer* $x^\star$*. Then, the gradient descent iterates* $x^{(t)}$ *with constant step-size* $\eta^{(t)} = \eta = \frac{2}{\mu+L}$ *(resp.* $\eta = \frac{1}{L}$*) converges linearly in norm*

$$\|x^{(t)} - x^\star\|^2 \leq c^{2t}\|x^{(0)} - x^\star\|^2,$$

where $c = \frac{\kappa-1}{\kappa+1}$ *(resp.* $c = 1 - \frac{1}{\kappa}$*.*

⚠ Linear ≡ Exponential ≡ Geometric!

Characteristic time (case $\eta = 1/L$)

$$\left(1 - \frac{1}{\kappa}\right)^{2t} \leq \exp\left(-\frac{2t}{\kappa}\right)$$

GD on OLS: Convergence in value

$$f(x) = \|Ax - b\|^2$$

Proposition

Let $x^\star$ any solution of $\min_x f(x)$. The gradient descent iterates $x^{(t)}$ with constant step-size $\eta^{(t)} = \eta = \frac{1}{L}$ converges with a sublinear $O(1/t)$ rate

$$f(x^{(t)}) - f(x^\star) \leq \frac{1}{4t\eta}\|x^{(0)} - x^\star\|_2^2.$$

If moreover, the solution is unique, the convergence is linear:

$$f(x^{(t)}) - f(x^\star) \leq \left(1 - \frac{1}{\kappa}\right)^{2t} (f(x^{(0)}) - f(x^\star)).$$

GD on OLS: Summary

Closed-form expression

$$x^{(t)} - x^\star = (\mathrm{Id} - \eta H)^t (x^{(0)} - x^\star)$$

(OLS)	$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n}\|Ax - y\|_2^2$
Gradient	$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p$
Hessian matrix	$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}$
Minimizers	$Hx^\star = \frac{1}{n} A^T y$

Case $\mu = 0 \Leftrightarrow H$ singular

- **Norm** no convergence
- **Value** $f(x^{(t)}) - f(x^\star) \leq \frac{1}{4t\eta}\|x^{(0)} - x^\star\|_2^2$

Case $\mu > 0 \Leftrightarrow H$ invertible

- **Norm** $\|x^{(t)} - x^\star\|^2 \leq c^{2t}\|x^{(0)} - x^\star\|^2$
- **Value** $f(x^{(t)}) - f(x^\star) \leq \left(1 - \frac{1}{\kappa}\right)^{2t} \left(f(x^{(0)}) - f(x^\star)\right)$

GD on OLS: Puzzle

Ordinary Least Squares (OLS)

$$\arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2n} \|Ax - y\|_2^2$$

Gaussian measurements

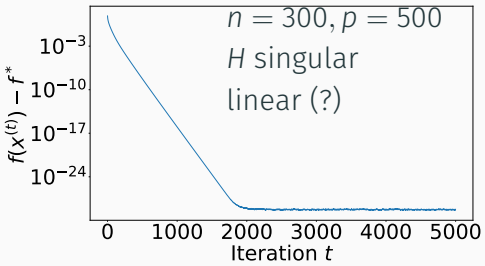
$$A_{ij} \sim \mathcal{N}(0, 1), b = 1$$

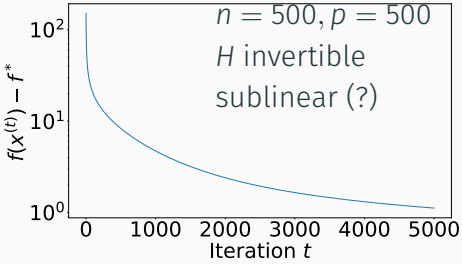
Gradient descent (constant LR)

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

Learning rate

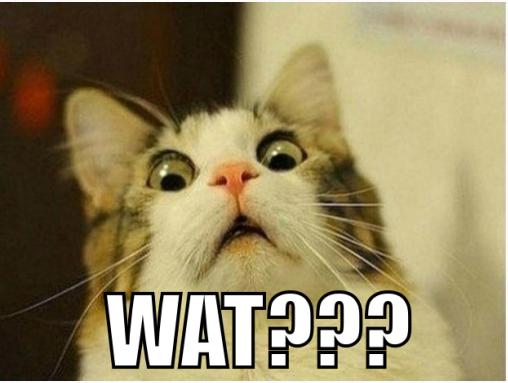
$$\eta = \frac{1}{L}$$





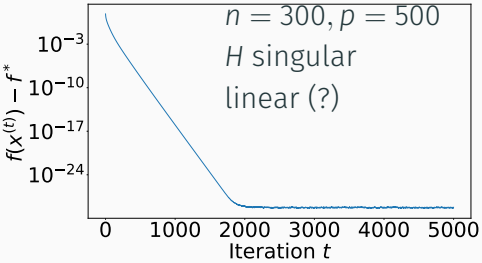
GD on OLS: Puzzle

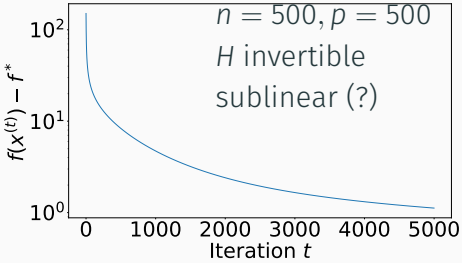
Ordinary Least Squares (OLS)



Learning rate

$$\eta = \frac{1}{L}$$





GD on OLS

- Can we do better than linear convergence for strongly convex problems? better than sublinear for convex problems?

  Yes. Conjugate gradient and Nesterov accceleration (later in this course)

- Can we say something about nonquadratic problems? nonconvex problems?

  Yes. Convex problems enjoy a similar theory. Nonconvex problems have guarantees with respect to local minima.

- Can we prove lower bounds in contrast to upper bound?

  Yes. We are going to show that $O(1/t^2)$ is a lower bound for the convex world.