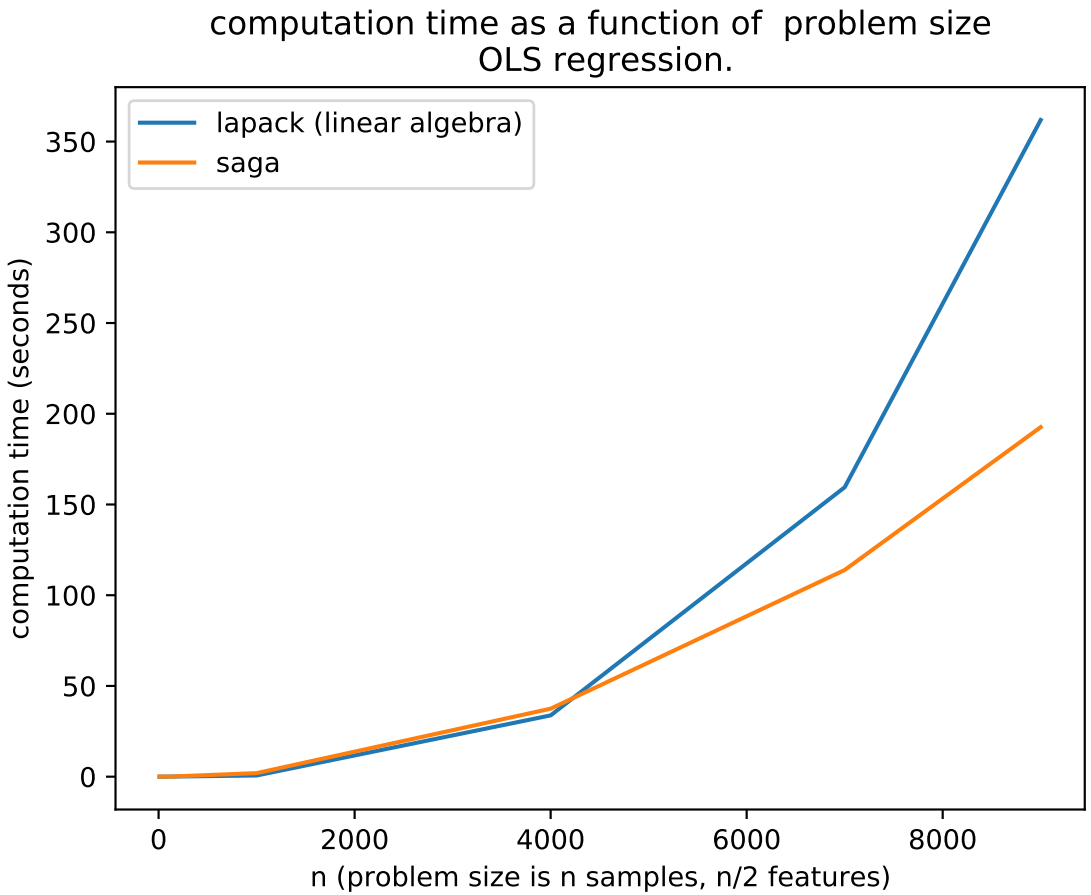
Fondamentaux théoriques du machine learning



Overview of lecture 6

- Density estimation
  - Motivation
  - Kernel density estimation
- Gradient algorithms
- Gradient descent
  - Convergence for least squares
  - General convergence results
  - Variations on gradient descent
- Stochastic gradient descent
- Variance reduction

Density estimation

Applications of density estimation

Kernel density estimation

Context

In machine learning, we often encounter problems in high dimension, where closed-form solutions are not avaible, or where even if they are available, the necessary computation time is too large.

Context

Instead, we use **iterative** algorithm such as Gradient descent (GD) or Stochastic gradient descent (SGD). SGD is the standard optimization algorithm for large-scale machine learning.
In this lecture we will study some theoretical convergence results on GD and SGD. The key properties will be convexity, strong convexity, and smoothness of the functions being optimized.

Least-squares problem

We will study ERM (Empirical risk minimization) of the OLS problem with a gradient algorithm.

- $X \in \mathbb{R}^{n,d}$
- $y \in \mathbb{R}^n$.

$$f(\theta) = \frac{1}{2n}||X\theta - y||_2^2 \tag{1}$$

$$\theta \leftarrow \theta - \gamma \nabla_f(\theta) \tag{2}$$

Gradient

$$f(\theta) = \frac{1}{2n}||X\theta - y||_2^2 \tag{3}$$

The gradient and the Hessian write :

$$\nabla_\theta f = \frac{1}{n}X^T(X\theta - y) \tag{4}$$

$$H = \frac{1}{n}X^T X \tag{5}$$

Minimizers

We note η^* the minimizers of f. If H is not invertible, they might be not unique, but all have the same function value $f(\eta^*)$.
All minimizers verify that

$$\nabla_{\eta^*} f = 0 \tag{6}$$

This means that

$$H\eta* = \frac{1}{n} X^T y \tag{7}$$

Minimizers

With a taylor expansion, we have that

$$f(\theta) - f(\eta^*) = \frac{1}{2}(\theta - \eta^*)H(\theta - \eta^*) \tag{8}$$

Gradient update

Exercice 1 : We perform a gradient update with step size γ. t denotes the iteration number. Show that

$$\theta_t = \theta_{t-1} - \gamma H(\theta_{t-1} - \eta^*) \tag{9}$$

Gradient update

Exercice 2 : Deduce that :

$$\theta_t - \eta^* = (I - \gamma H)(\theta_{t-1} - \eta^*) \tag{10}$$

and that

$$\theta_t - \eta^* = (I - \gamma H)^t(\theta_0 - \eta^*) \tag{11}$$

Measure of performance

We can use two measures of performance of the gradient algorithm :

- Distance to minimizer :

$$||\theta_t - \eta^*||_2^2 = (\theta_0 - \eta^*)^T (I - \gamma H)^{2t} (\theta_0 - \eta^*) \qquad (12)$$

- Convergence in function values :

$$f(\theta_t) - f(\eta^*) = \frac{1}{2} (\theta_0 - \eta^*)^T (I - \gamma H)^{2t} H (\theta_0 - \eta^*) \qquad (13)$$

Distance to minimizer

If we can bound the eigenvalues of $(I - \gamma H)^{2t}$, we can bound $||\theta_t - \eta^*||_2^2$.

Exercice 3 : We note λ_i the eigenvalues of H. What are the eigenvalues of $(I - \gamma H)^{2t}$?

Bounding eigenvalues

We introduce the **condition number** $\kappa = \frac{K}{\mu}$.

- L is the largest eigenvalue of H.
- μ is the largest eigenvalue of H.

All eigenvalues of $(I - \gamma H)^{2t}$ have a magnitude that is smaller than

$$\left(\max_{\lambda \in [\mu, L]} |1 - \gamma \lambda| \right)^{2t} \tag{14}$$

Bounding eigenvalues

We introduce the **condition number** $\kappa = \frac{K}{\mu}$.

- L is the largest eigenvalue of H.
- μ is the largest eigenvalue of H.

All eigenvalues of $(I - \gamma H)^{2t}$ have a magnitude that is smaller than

$$\left(\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda| \right)^{2t} \tag{15}$$

Hence, we want to find γ such that $\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda|$ is **minimum** (or at least small).

Bounding eigenvalues

Exercice 4 : Find γ such that

$$\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda| \leq (1 - \frac{\mu}{L}) = (1 - \frac{1}{\kappa}) \tag{16}$$

Exponential convergence

With $\gamma = \frac{1}{L}$, we obtain an exponential convergence

$$||\theta_t - \eta^*||_2^2 \leq \left(1 - \frac{1}{\kappa}\right)^{2t} ||\theta_0 - \eta^*||_2^2 \tag{17}$$

Approximation error

We have that

$$(1 - \frac{1}{\kappa})^{2t} \leq \exp(-\frac{1}{\kappa})^{2t} = \exp(-\frac{2t}{\kappa}) \tag{18}$$

Exercice 5 : What number of iterations is suffcient in order to have a relative reduction of $||\theta_t - \eta^*||_2^2$ of ϵ ?

Approximation error

$$\begin{aligned} \epsilon &\geq \exp(-\frac{2t}{\kappa}) \\ &\Leftrightarrow -\log(\epsilon) \leq \frac{2t}{\kappa} \\ &\Leftrightarrow \frac{\kappa}{2}\log(\frac{1}{\epsilon}) \leq t \end{aligned} \tag{19}$$

Large condition number

If $\kappa = +\infty$ ($\mu = 0$), we do not have a convergence guarantee.

Convergence in function values

If H is not invertible, we only have a convergence rate in $\mathcal{O}(\frac{1}{t})$.

Generalization

We will now study more general functions.

Smoothness

Definition

Smoothness

A differentiable function f with real values is said L-smooth if and only if

$$\forall x, y \in \mathbb{R}^d, |f(y) - f(x) - \nabla_x f(y - x)| \leq \frac{L}{2}||y - x||^2$$

Smoothness

Lemma

f is L-smooth if and only if it has L-Lipshitz continuous gradients.

Smoothness of least-squares

Exercice 6 : **Smoothness** : Consider

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$
$$= \frac{1}{n}||Y - X\theta||_2^2$$

Is $R_n(\theta)$ smooth ?

Non smooth optimization

Lipshit

Non smooth optimization

Exercice 7 : **Smoothness :** Is f has L-lipshitz continuous gradients, is f necessary Lipshitz-continuous ?

Smooth, strongly convex functions

Theorem

Convergence of GD for a strongly convex function
Let $f : \mathbb{R}^d \Rightarrow \mathbb{R}$ be a μ-strongly convex unction with L-Lipshitz continuous gradients. Let x^* be the global minimum of f (which we know exists since f is strongly convex), $x_0 \in \mathbb{R}$, $T \in \mathbb{N}$.
With constant step size $\gamma_t = \frac{1}{L}$, we have

$$f(x_t) - f(x^*) \leq (1 - \frac{\mu}{L})^t \big(f(x_0) - f(x^*)\big)$$

Optimamity

Nesterov acceleration

page 113

Line search

Conjugate gradient

Algorithmic complexity

Stochastic gradient descent

Convergence result

With averaging

Algorithmic complexity

SAGA algorithm