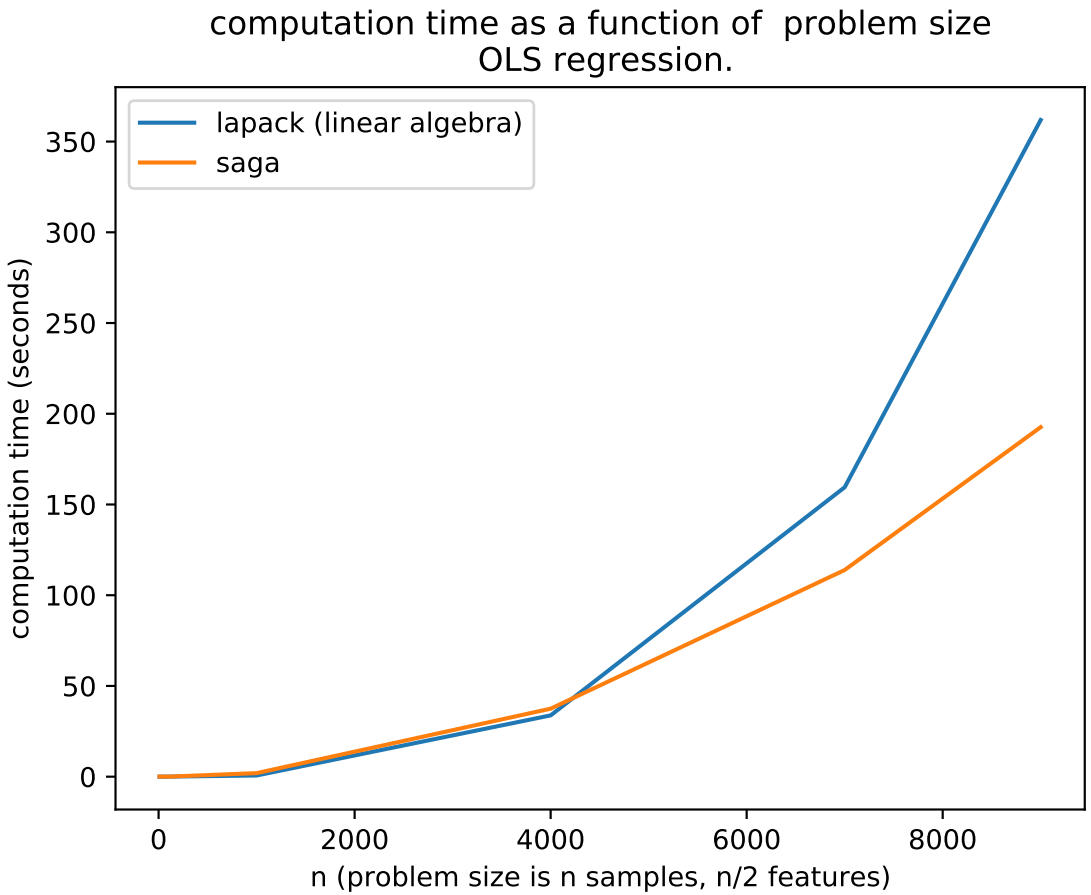
Fondamentaux théoriques du machine learning



Overview of lecture 6

Gradient descent

- Motivation
- Convergence for least squares
- General convergence results

Stochastic gradient descent

- Motivation
- Gradient estimates
- Convergence results
- Comparison between GD and SGD

Probabilistic modelling

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

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.

Example 1 : Computing the OLS estimator requires a matrix inversion, which is $\mathcal{O}(d^3)$.

$$\hat{\theta} = (X^T X)^{-1} X^T Y \tag{1}$$

Context

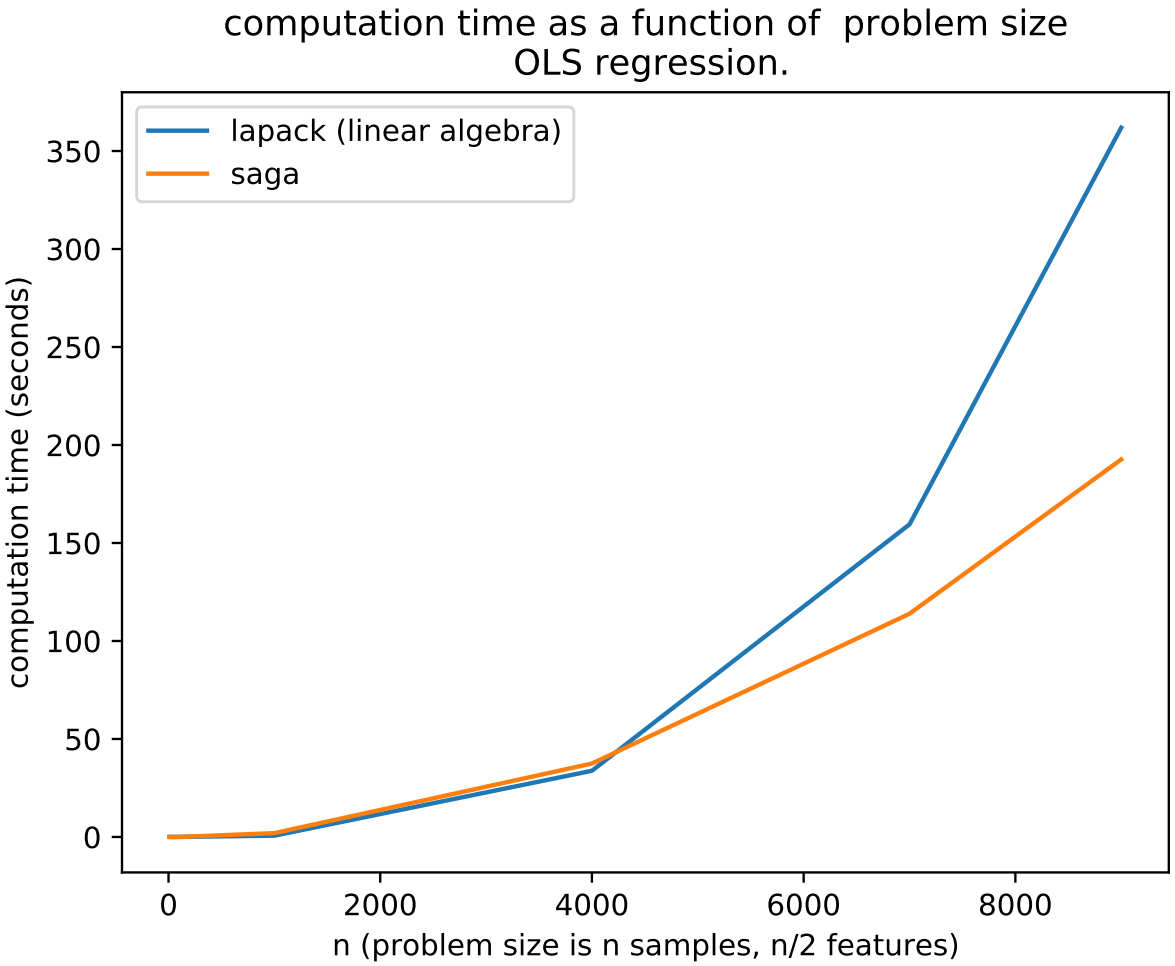
In machine learning, we often encounter problems in high dimension, where closed-form solutions are not avaible, or where even if they are available, the necessary computation time is too large.

Example 2 : The cancellation of the gradient of the objective function with logistic loss has no closed-form solution.

Context

Instead, we often 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.

SGD vs Lapack



Gradient descent

We want to minimize a function f defined over $\mathbb{R}^d$.

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

Least-squares problem

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

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

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

$$\theta \leftarrow \theta - \gamma \nabla_f(\theta) \tag{4}$$

Gradient

$$f(\theta) = \frac{1}{2n}||X\theta - y||_2^2 \tag{5}$$

The gradient and the Hessian write :

$$\nabla_\theta f = \frac{1}{n}X^T(X\theta - y) \tag{6}$$

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

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

$$\nabla_{\eta^*} f = 0 \tag{8}$$

This means that

$$H\eta* = \frac{1}{n} X^T y \tag{9}$$

Minimizers

With a taylor expansion, we have that

$$f(\theta) - f(\eta^*) = \frac{1}{2}(\theta - \eta^*)H(\theta - \eta^*) \tag{10}$$

Gradient update

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

$$\theta_t = \theta_{t-1} - \gamma H(\theta_{t-1} - \eta^*) \tag{11}$$

Gradient update

Exercice 2 : Deduce that :

$$\theta_t - \eta^* = (I - \gamma H)(\theta_{t-1} - \eta^*) \tag{12}$$

and that

$$\theta_t - \eta^* = (I - \gamma H)^t(\theta_0 - \eta^*) \tag{13}$$

Measure of performance

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

- Distance to minimizer :

$$||\theta_t - \eta^*||_2^2 = (\theta_0 - \eta^*)^T (I - \gamma H)^{2t} (\theta_0 - \eta^*) \qquad (14)$$

- Convergence in function values :

$$f(\theta_t) - f(\eta^*) = \frac{1}{2} (\theta_0 - \eta^*)^T (I - \gamma H)^{2t} H (\theta_0 - \eta^*) \qquad (15)$$

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.
- As H is positive semidefinite (matrice positive), $\mu \geq 0$.
- By convention, if $\mu = 0$, $L = +\infty$.

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

$$\left(\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda| \right)^{2t} \tag{16}$$

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

$$\left(\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda| \right)^{2t} \tag{17}$$

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

$$\max_{\lambda \in [\mu, L]} |1 - \gamma\lambda| \leq (1 - \frac{\mu}{L}) = (1 - \frac{1}{\kappa}) \tag{18}$$

Exponential convergence

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

$$||\theta_t - \eta^*||_2^2 \leq \left(1 - \frac{1}{\kappa}\right)^{2t} ||\theta_0 - \eta^*||_2^2 \tag{19}$$

Approximation error

We have that

$$(1 - \frac{1}{\kappa})^{2t} \leq \exp(-\frac{1}{\kappa})^{2t} = \exp(-\frac{2t}{\kappa}) \tag{20}$$

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

$$\begin{aligned} & \exp(-\frac{2t}{\kappa}) \leq \epsilon \\ \Leftrightarrow & -\log(\epsilon) \leq \frac{2t}{\kappa} \\ \Leftrightarrow & \frac{\kappa}{2}\log(\frac{1}{\epsilon}) \leq t \end{aligned} \tag{21}$$

Large condition number

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

Large condition number

If $\kappa = +\infty$ ($\mu = 0$), we do not have a convergence guarantee. However, we can still obtain a convergence result by studying the function values.

Convergence in function values

We recall that

$$f(\theta_t) - f(\eta^*) = \frac{1}{2}(\theta_0 - \eta^*)^T (I - \gamma H)^{2t} H(\theta_0 - \eta^*) \qquad (22)$$

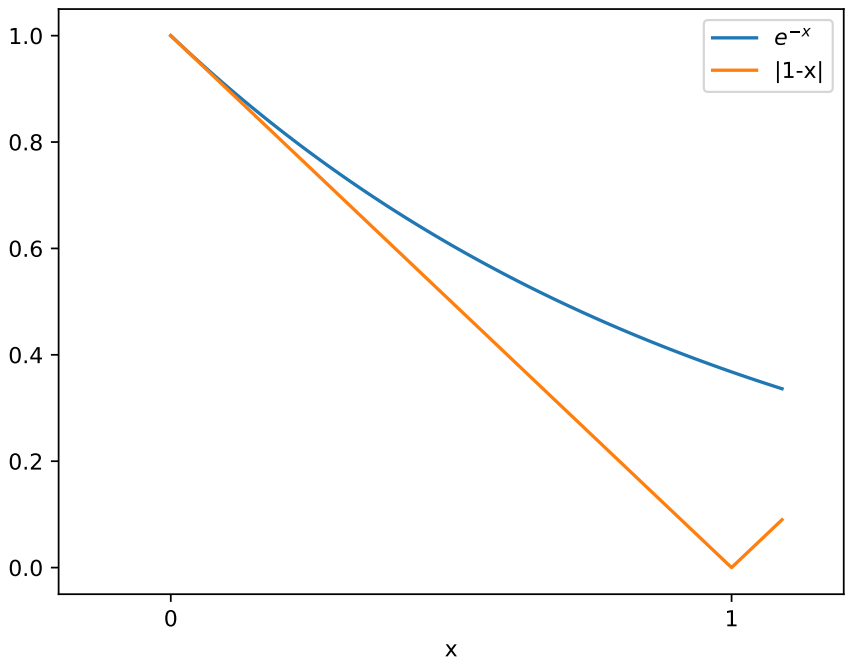
Convergence in function values

As previously, we can show that the eigenvalues of $(I - \gamma H)^{2t} H$ are the $\lambda(1 - \gamma\lambda)^{2t}$, with λ being an eigenvalue of H.

Convergence in function values

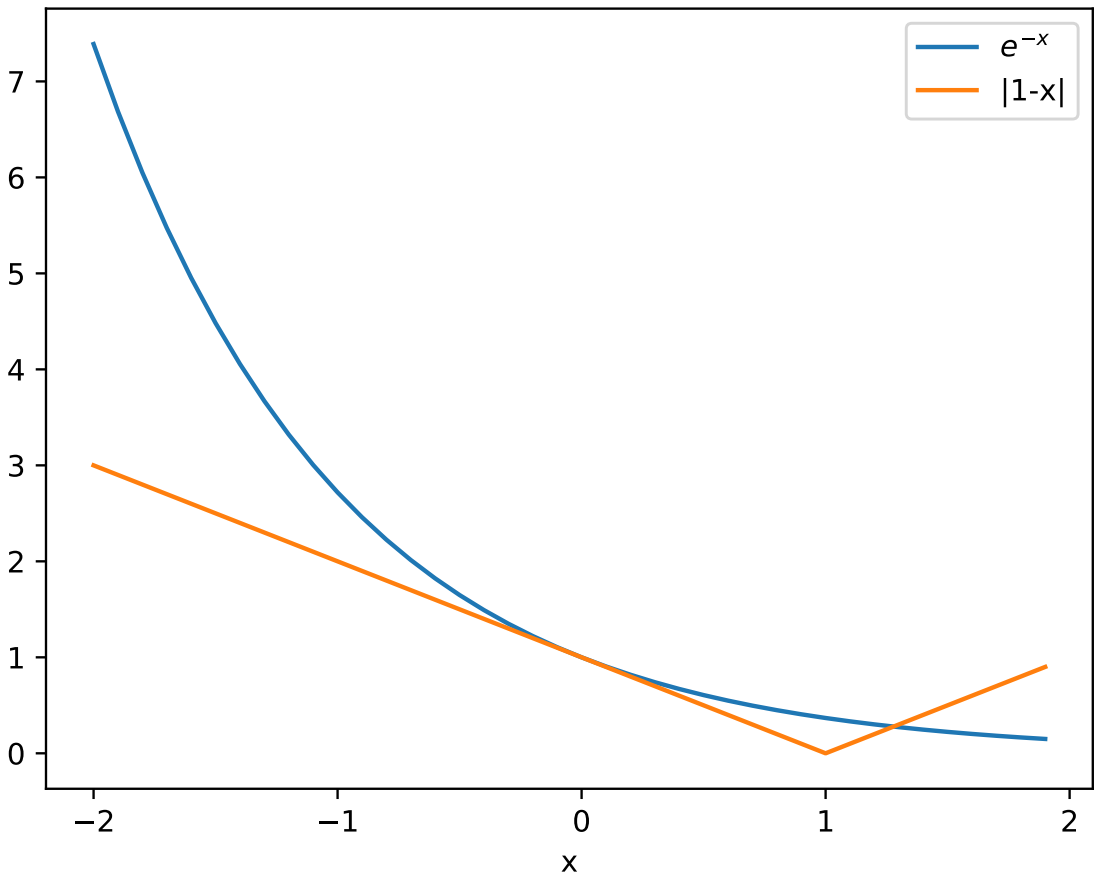
As previously, we can show that the eigenvalues of $(I - \gamma H)^{2t} H$ are the $\lambda(1 - \gamma\lambda)^{2t}$, with λ being an eigenvalue of H.
If $\gamma \leq \frac{1}{L}$, we have $0 \leq \gamma\lambda \leq 1$.

If $\gamma \leq \frac{1}{L}$, we have $0 \leq \gamma\lambda \leq 1$ and $|1 - \gamma\lambda| \leq \exp(-\gamma\lambda)| \leq 1$.



Careful !

If $\gamma\lambda \geq 1$, it is possible that $|1 - \gamma\lambda| \geq \exp(-\gamma\lambda)|$.



Convergence rate

We then have that if $\gamma \leq \frac{1}{L}$, then for each eigenvalue λ of H,

$$\begin{aligned} |\lambda(1-\gamma\lambda)^{2t}| &\leq \lambda \exp(-\gamma\lambda)^{2t} \\ &= \lambda \exp(-2t\gamma\lambda) \\ &= \frac{1}{2t\gamma} 2t\gamma\lambda \exp(-2t\gamma\lambda) \\ &\leq \frac{1}{2t\gamma} \sup_{\alpha \geq 0} \alpha \exp(-\alpha) \end{aligned} \tag{23}$$

Convergence rate

We then have that if $\gamma \leq \frac{1}{L}$, then for each eigenvalue λ of H,

$$\begin{aligned} |\lambda(1-\gamma\lambda)^{2t}| &\leq \lambda \exp(-\gamma\lambda)^{2t} \\ &= \lambda \exp(-2t\gamma\lambda) \\ &= \frac{1}{2t\gamma} 2t\gamma\lambda \exp(-2t\gamma\lambda) \\ &\leq \frac{1}{2t\gamma} \sup_{\alpha \geq 0} \alpha \exp(-\alpha) \end{aligned} \tag{24}$$

Exercice 6 : What is the maximum of $\alpha \mapsto \alpha \exp(-\alpha)$ for $\alpha \in \mathbb{R}_+$?

Convergence rate

Finally, we have that

$$f(\theta_t) - f(\eta^*) \leq \frac{1}{4t\gamma}||\theta_0 - \eta^*||_2^2 \tag{25}$$

Convergence in function values

Conclusion :

- if H is invertible, we have a convergence rate in $\exp(-\frac{2t}{\kappa})$.
- if H is not invertible, we have a convergence rate in $\mathcal{O}(\frac{1}{t})$.

Convergence rates

In the practical sessions we will observe experimentally these convergence rates, or the fact that they are just **bounds**.

Generalization

These results can be extended to a more general setting. The convergence garantees that we can obtain will depend on the following properties of the objective function :

- convexity or strong convexity
- smoothness (Lipshitz-continuous gradients) or non-smoothness

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.

Definition

L-Lipschitz continuous gradients

f has L-Lipschitz continuous gradients if $\forall x, y \in \mathbb{R}^d$,

$$||\nabla_x f - \nabla_y f|| \leq L||x - y||$$

Smoothness of least-squares

Exercice 7 : **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 ?

Link with eigenvalues

If f is two times differentiable, with a Hessian $H(\theta)$.

- f is μ-convex (strongly convex) if and only if

$$\forall \theta, H''(\theta) \geq \mu I_d \tag{26}$$

(with the Loewner order : all eigenvalues are $\geq \mu$)

- f is L-smooth if and only if

$$\forall \theta, -L I_d \leq H''(\theta) \leq L I_d \tag{27}$$

All eigenvalues λ are such that $|\lambda| \leq L$.

`https://en.wikipedia.org/wiki/Loewner_order`
We can then define again the condition number $\kappa = \frac{L}{\mu}$.

$$\kappa \geq 1 \tag{28}$$

Smooth and strongly convex problems

- If the loss is smooth (for instance square loss, logistic loss) and the predictor is linear, then the objective function is smooth.
- If the objective function is convex, then the $L2$-regularized version (obtained by adding $\frac{\mu}{2}||\theta||^2$ to the objective function) is μ-strongly convex.

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 function 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

$$\begin{aligned} f(x_t) - f(x^*) &\leq (1-\kappa)^t \big(f(x_0) - f(x^*)\big) \\ &\leq \exp(-\kappa t)\big(f(x_0) - f(x^*)\big) \end{aligned} \tag{29}$$

Smooth, convex function

Theorem

Convergence of GD for a smooth convex function
Let $f : \mathbb{R}^d \to \mathbb{R}$, *with global minimiser* x^*. *With constant step-size* $\gamma_t = \frac{1}{L}$, *the iterates* x_t *of GD satisfy :*

$$f(x_t) - f(\eta^*) \leq \frac{L}{2t}||x_0 - \eta^*||^2$$

Extensions

- Line search
- Nesterov accleration (optimal rates among algorithms that linearly combine gradients)

Comparison with Newton method

Newton's method minimizes the second-order Taylor expansion around θ_{t-1} in order to compute θ_t.
The convergence of Newton method is faster in the number of iterations, but each iteration is expensive : $\mathcal{O}(d^3)$ since it requires to solve a linear system.

$$C||\theta_t - \theta_*|| \leq (C||\theta_t - \theta_*||)^2 \tag{30}$$

As in machine learning we often have an estimation error $\mathcal{O}(\frac{1}{\sqrt{n}})$, the tradeoff is not in favor of Newton's method.

Stochastic gradient descent

In machine learning, we often consider an objective function of the form

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f_\theta(x_i)) + \Omega(\theta) \tag{31}$$

Batch gradient

In machine learning, we often consider an objective function of the form

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f_\theta(x_i)) + \Omega(\theta) \tag{32}$$

Computing the gradient of f requires at least n calculations, and each calculation also has a complexity that depends on the dimension d. When n and d are large, this can be quite slow.

Stochastic gradient descent

We consider an objective function of the form

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f_\theta(x_i)) + \Omega(\theta) \tag{33}$$

Instead of computing the **batch gradient** $\nabla_\theta F$, we will compute **unbiased stochastic estimations** of the gradient, $g_t(\theta_{t-1})$. For all t,

$$E[g_t(\theta_{t-1})] = \nabla_{\theta_{t-1}} F \tag{34}$$

SGD update

The SGD update reads

$$\theta_t = \theta_{t-1} - \gamma_t g_t(\theta_{t-1})$$

Empirical risk minimization : at each time step we choose uniformly $i(t) \in \{1, \dots, n\}$ and

$$g_t = \nabla_\theta \Big(l(y_{i(t)}, f_\theta(x_{i(t)})) + \Omega(\theta) \Big) \tag{35}$$

SGD as an estimation of GD

Given θ_{t-1}, we have that

$$\begin{aligned} E\left[\theta_t\right] &= E\left[\theta_{t-1} - \gamma_t g_t(\theta_{t-1})\right] \\ &= \theta_{t-1} - \gamma_t E\left[g_t(\theta_{t-1})\right] \\ &= \theta_{t-1} - \gamma_t \nabla_{\theta_{t-1}} F \end{aligned}$$

SGD as an estimation of GD

Given θ_{t-1}, we have that

$$\begin{aligned} E\left[\theta_t\right] &= E\left[\theta_{t-1} - \gamma_t g_t(\theta_{t-1})\right] \\ &= \theta_{t-1} - \gamma_t E\left[g_t(\theta_{t-1})\right] \\ &= \theta_{t-1} - \gamma_t \nabla_{\theta_{t-1}} F \end{aligned}$$

In expectation, SGD behaves as GD.

Convergence result

For SGD, the convergence results and proofs are more abstract. In the general case :

- Either we have results on expected values, such as that of $||\theta_t - \theta^*||$.
- Or convergence garantees on **averages of the iterates.**

Strongly convex smooth objective

Definition

Variance of the estimator of the gradient

We define the **variance** of the estimator of the gradient as

$$\sigma^2(\theta) = E_z||\nabla_\theta f(\theta) - \nabla_\theta L(\theta, z)||^2$$

We note that σ depends on θ.

Strongly convex smooth objective

Theorem

Convergence of SGD for a strongly convex L-smooth function
We assume that :

- *The gradients estimates are unbiased.*
- *f is μ strongly convex.*
- *f has L-Lipshitz continuous gradients*
- $\exists \sigma^2 > 0, \forall \theta \in \mathbb{R}^d, \sigma^2(\theta) \leq \sigma^2$

Let $\gamma_t = \gamma$, $\forall t \in \mathbb{N}$ and $0 < \gamma < \frac{1}{2L}$. Then

$$E_{z_1,\dots,z_T}||\theta_T - \theta^*|| \leq (1 - \mu\gamma)^T ||\theta_0 - \theta^*|| + \frac{\gamma}{\mu}\sigma^2$$

Convex objective

Theorem

- *The gradients estimates are unbiased.*
- *f is convex*
- *f is L-Lipshitz*
- *The gradient is bounded : $\forall t$, $||g_t(\theta_{t-1})||_2^2 \leq L^2$ almost surely.*
- *f admits a minimiser θ^* such that $||\theta^* - \theta_0||_2 \leq D$.*

Let $\gamma_t = \frac{D}{L\sqrt{t}}$, $\forall t \in \mathbb{N}$. Then, the iterates of SGD satisfy :

$$E[f(\hat{\theta}_t) - f(\theta^*)] \leq DL\frac{2 + \log(t)}{2\sqrt{t}}$$

where $\hat{\theta}_t = \frac{\sum_{s=1}^{t} \gamma_s \theta_{s-1}}{\sum_{s=1}^{t} \gamma_s}$.

Strongly convex objective

We consider the SGD update corresponding to the regularized objective g.

$$g(\theta) = f(\theta) + \frac{\mu}{2}||\theta||_2^2$$

Now, the SGD iteration reads :

$$\theta_t = \theta_{t-1} - \gamma_t \left(g_t(\theta_{t-1}) + \mu\theta_{t-1} \right)$$

Strongly convex objective

Theorem

- *The gradients estimates are unbiased.*
- *f is μ-convex*
- *f is L-Lipshitz*
- *The gradient is bounded : $\forall t$, $||g_t(\theta_{t-1})||_2^2 \leq L^2$ almost surely.*
- *f admits a minimiser θ^* such that $||\theta^* - \theta_0||_2 \leq D$.*

Then, with $\gamma_t = \frac{1}{\mu t}$.

$$E[f(\hat{\theta}_t) - f(\theta^*)] \leq \frac{2L^2(1 + \log t)}{\mu t}$$

where $\hat{\theta}_t = \frac{1}{t}\sum_{s=1}^{t} \theta_{s-1}$.

Algorithmic complexities

Exercice 8 : We consider a least squares problem. Compute the computational complexities of

- an iteration of GD.
- an iteration of SGD

Algorithmic complexities

Exercice 9 : We consider a least squares problem. Compute the computational complexities of

- an iteration of GD : $\mathcal{O}(nd)$.
- an iteration of SGD : $\mathcal{O}(d)$.

Comparison

The ridge regression problem is smooth and strongly convex.

- GD has a convergence rate of $\mathcal{O}(\exp(-\frac{t}{\kappa}))$. To get an error of ϵ, we must have $t = \mathcal{O}(\kappa \log \frac{1}{\epsilon})$. Since each iteration requires $\mathcal{O}(nd)$ computations, the computation time will be $\mathcal{O}(\kappa nd \log \frac{1}{\epsilon})$.
- SGD has a convergence rate of $\mathcal{O}(\frac{\kappa}{t})$. To get an error of ϵ, we must have $t = \mathcal{O}(\frac{\kappa}{\epsilon})$. Since each iteration is $\mathcal{O}(d)$, we have a computation time of $\mathcal{O}(\frac{\kappa d}{\epsilon})$.

Comparison

As a consequence :

- When n is large and ϵ not too small, GD will need more computation time to reach error ϵ. An order of magnitude can be obtained by studying the value ϵ^* such that

$$\kappa n d \log \frac{1}{\epsilon^*} = \frac{\kappa d}{\epsilon^*}$$

Which translates to

$$\epsilon^* \log \epsilon^* = -\frac{1}{n}$$

- When $\epsilon \to 0$, GD becomes faster than SGD to reach this precision.

Conclusion

For low precision and large n, SGD is a preferable.
In machine learning, due to the estimation error that is $\mathcal{O}(\frac{1}{\sqrt{n}})$, a very high precision is often not needed.

Extensions of SGD

See also :

- Variance reduction methods (SAG, SAGA)

Context

We are given a set of observations $\{y_1, \ldots, y_n\} \in \mathcal{Y}$ that we assume are generated i.i.d from an unknown distribution. We look for a **probabilistic model** that explains well the data. We could use this model to predict well new data, that would be statistically similar to the observed ones.

Density estimation

We will consider **parametric models** for density estimation.

Definition

Parametric model

Let $d > 1$ and $\Theta \subset \mathbb{R}^p$ be a set of parameters. A parametric model $\mathcal{P}$ is a set of probability distributions on $\mathcal{Y}$, indexed by Θ.

$$\mathcal{P} = \{p_\theta | \theta \in \Theta\}$$

Examples :

- Bernoulli model (parameter p)
- Gaussian model (parameter (μ, σ))
- Binomial model (parameter θ)

Objective

If we assume that the data were generated from some $p_{\theta^*} \in \mathcal{P}$, with a unknown parameter θ^*, our goal if to find a good estimation of θ. If the data are indeed generated by a distribution in $\mathcal{P}$, the problem is said to be **well specified**. Otherwise, the problem is said to be **misspecified**.

Likelihood

Definition

Likelihood

Let $\mathcal{P} = \{p_\theta, \theta \in \Theta\}$ be a parametric model. Given $y \in \mathcal{Y}$, the **likelihood** of θ is defined as the function $\theta \mapsto p_\theta(y)$.

The likelihood $L(.|D_n)$ of a dataset $D_n = (y_1, \ldots, y_n)$ is defined as

$$L(.|D_n) : \theta \mapsto \prod_{i=1}^{n} p_\theta(y_i)$$

The **maximum likelihood estimator** (MLE) is the parameter θ that maximises the likelihood :

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta}(L(\theta|D_n)$$

Remarks

- Since the samples y_i are assumed to be independent, the likelihood corresponds to the probability of observing the dataset according to p_θ.
- We often maximise the log of the likelihood, as it is easier to differentiate a sum. Since log is an increasing function, the MLE is also the maximiser of the log of L.

Example 1

Exercice 10 : We observe the data $(1, 0)$. We assume that these data come from a random variable that follows a Bernoulli distribution of parameter p. What is the likelihood of these observations as a function of p ?

Example 1

Exercice 10 : We observe the data $(1, 0)$. We assume that these data come from a random variable that follows a Bernoulli distribution of parameter θ. What is the likelihood of these observations as a function of θ ?

$$L = P_\theta(1)P_\theta(0) \tag{36}$$

Example 1

Exercice 10 : We observe the data $(1, 0)$. We assume that these data come from a random variable that follows a Bernoulli distribution of parameter θ. What is the likelihood of these observations as a function of θ ?

$$L = P_\theta(1)P_\theta(0) \tag{37}$$

For which value of p is this likelihood **maximum** ?

Example 2

We observe the data $(2.5, 3.5)$. We assume that these data come from a normal law of parameters μ and σ.

$$\begin{aligned} L &= p(2.5|\mu,\sigma)p(3.5|\mu,\sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{2.5-\mu}{\sigma})^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{3.5-\mu}{\sigma})^2} \end{aligned} \tag{38}$$

We wan show that the likelihood is maximum for :

- $\hat{\mu} = \frac{2.5+3.5}{2}$
- $\hat{\sigma^2} = \frac{(2.5-\hat{\mu})^2+(3.5-\hat{\mu})^2}{2}$

ERM

In the context of density estimation, we can define a loss function as the **negative log-likelihood.**

$$\Theta \times \mathcal{Y} \mapsto -\log(p_\theta(y))$$

Given this loss, the risk writes :

$$R(\theta) = E_Y[-\log\left(p_\theta(y)\right)]$$

and the empirical risk (ER) :

$$R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(p_\theta(y_i)\right)]$$

The MLE is then also the empirical risk minimizer.

KL divergence

The Kullback-Leibler divergence is a quantity used to compare two probability distributions.

Definition

Kullback-Leibler divergence
Given two distributions p and q, the KL divergence from p to q is defined as :

$$KL(p||q) = E_{Y\sim p}\left[\log \frac{p(Y)}{q(Y)}\right]$$

Lemma

If the data are generated by p_{θ^*}, then $KL(p_{\theta^*}||p_\theta)$ is the excess risk of p_θ, with the legative log-likelihood loss.