Elia Fantini - Sciper n. 336006

Math of Data: Homework 1

Premise: to the person who is reading and correcting this report, I want to apologize for not having used LaTeX, but I've never used it before and I couldn't find the time to get practical with it.

Exercise 1

Point a):

From slide 29/41 of Recitation 2 we learn that, being Q a convex set in $\mathbb{R}^p$, a function $\sigma \in C^2(Q)$ is called convex on Q if for any $u \in Q$:

$$\nabla^2 \sigma(u) \succeq 0$$

In this case $\sigma : u \mapsto \log(1 + e^{-u})$.

The gradient of σ is:

$$\nabla\sigma(u) = \frac{1}{1+e^{-u}} \cdot e^{-u} \cdot (-1) = -\frac{e^{-u}}{e^{-u}\left(1 + \frac{1}{e^{-u}}\right)} = -\frac{1}{1+e^{u}}$$

The domain of σ is $\mathbb{R}^p$ and the domain of $\nabla\sigma$ is still $\mathbb{R}^p$.

We can notice that $\nabla\sigma(u) < 0 \quad \forall u \in \mathbb{R}^p$.

$$\nabla^2_\sigma(u) = (-1) \cdot \left(-\frac{1}{(1+e^u)^2} \cdot e^u\right) = \frac{e^u}{(1+e^u)^2} > 0 \quad \forall u \in \mathbb{R}^p$$

Since $\nabla^2\sigma$ is strictly positive, with domain $\mathbb{R}^p$, we can say that $\sigma \in C^2(\mathbb{R}^p)$ and that it is convex due to the Theorem mentioned above.

Since the sum of convex functions is again convex, it is convex the function $f(u) = \sum_{i=1}^{n} \log(1 + e^{-u_i})$ with $u_i = -b_i a_i^T x$

For simplicity let's suppose the case $p = 1 \Rightarrow \mathbb{R}^p = \mathbb{R}$. Then, the limit of $u \to +\infty$ is:

$$\lim_{u \to +\infty} \log(1 + e^{-u}) = 0^+$$

Hence, the infima of this function is 0, but it doesn't have a minima. With $f(u) = \sum_{i=1}^{n} \sigma(u_i)$ it's then the same case.

Point b)

The infimum of a subset S of a partially ordered set P is a greatest element in P that is less than or equal to all elements of S. If the infimum is contained in S, than it's also a minimum. In this case, S is the codomain of a function.
An example of a convex function that does not attain its infimum is $f(x) = e^x$, since $\lim_{x \to -\infty} e^x = 0^+$ and $e^x > 0 \ \forall x \in \mathbb{R}$.

Point c)

If $b_i a_i^T x_0 > 0$ this means that $0 < e^{-b_i a_i^T x_0} < 1$ and that $0 < \log(1 + e^{-b_i a_i^T x_0}) < \log(2)$. If $b_i a_i^T x_0 > 0 \ \forall i \in \{1, \dots, n\}$ then $0 < f(x_0) = \sum_{i=1}^{n} \log(1 + e^{-b_i a_i^T x_0}) < n \log(2)$.
We notice that $f(2x_0) < f(x_0)$, in particular, given $f(\alpha x_0)$ such that $f(\alpha x_0) := \sum_{i=1}^{n} \log(1 + e^{-b_i a_i^T x_0 \cdot \alpha})$ we have that

$\lim_{\alpha \to +\infty} f(\alpha x_0) = 0^+$ as $\log(1 + e^{(-\infty)}) \to \log(1^-) \to 0^+$, so the infimum is 0 but the function doesn't attain it.

Point d)

Previously I called σ the function $u \mapsto \log(1 + e^{-u})$ due to a distraction mistake, let's now call σ the function $u \mapsto \frac{1}{1 + e^{-u}}$

The function $h(x) = \log(1 + e^{-b_i a_i^T x})$ is the composition of two functions. $h(x) = g(x) \circ m(x)$ where $m(x) = -b_i a_i^T x$ whose Jacobian in $J_m(x) = -b_i a_i^T$. $g(x) = \log(1 + e^{-x})$ whose Jacobian is

$$J_g(x) = -\frac{1}{e^x + 1}$$

Using the chain rule we have: ~~$J_h(x) = J_g(x) \cdot J_m(x) = -b_i$~~

$$J_h(x) = J_g(m(x)) \cdot J_m(x)$$

$$\Rightarrow J_h(x) = -b_i \cdot \frac{1}{1+e^{-b_i a_i^T x}} \cdot a_i = -b_i \cdot \sigma(-b_i a_i^T x)\, a_i^T$$

Since the gradient of a sum is the sum of the gradients and

$\nabla \|x\|^2 = 2 \cdot x$ and $\nabla h(x) = J_h^T(x) = -b_i\, \sigma(-b_i a_i^T x) \cdot a_i$,

then $\nabla f_n(x) = \sum_{i=1}^{n} -b_i \cdot \sigma(-b_i a_i^T x) \cdot a_i + \mu x$ QED

Point e)

As before $h(x) = -b_i \sigma(-b_i a_i^T x) \cdot a_i = g \circ m$ where

$g(x) = -b_i \sigma(x) \cdot a_i$ and $m(x) = -b_i a_i^T x$

Then, we can apply the chain rule $J_h(x) = J_g(m(x)) \cdot J_m(x)$

$$J_m(x) = J_m(-b_i a_i^T x) = -b_i a_i^T$$

$$J_g(m(x)) = J_g\left(-b_i \cdot \frac{1}{1+e^{-x}} \cdot a_i\right) = -b_i \cdot \frac{e^{-x}}{(1+e^{-x})^2} \cdot a_i = -b_i \cdot \sigma(x) \cdot (1-\sigma(x)) \cdot a_i$$

Therefore:

$$J_h(x) = -b_i \sigma(-b_i a_i^T x) \cdot \left(1 - \sigma(-b_i a_i^T x)\right) \cdot a_i \cdot (-b_i a_i^T)$$

Since $b_i = [-1, 1]$, $(-b_i) \cdot (-b_i) = 1$

$$\Rightarrow J_h(x) = \sigma(-b_i a_i^T x) \cdot (1 - \sigma(-b_i a_i^T x)) \cdot a_i \cdot a_i^T = \nabla^2 h(x)$$

Given $r(x) = \mu x \Rightarrow \nabla r(x) = \mu I$

Hence:

$$\nabla^2 f_\mu(x) = \sum_{i=1}^{N} \sigma(-b_i a_i^T x) \cdot (1 - \sigma(-b_i a_i^T x)) \cdot a_i a_i^T + \mu I$$ QED

Point f)

Given the definition on slide 35/41 from recitation 2, we have that:

A convex function $f: Q \to \mathbb{R}$ is said to be μ-strongly convex if

$$h(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$$

is convex.

We've already proven that $f(x)_n$ is convex $\Rightarrow h(x) = f_\mu(x) - \frac{\mu}{2}\|x\|_2^2$

$= f(x) + \frac{\mu}{2}\|x\|_2^2 - \frac{\mu}{2}\|x\|_2^2 = f(x)$, which is convex. Hence, $f_\mu(x)$ is μ-strongly convex.

Since strong convexity implies strict convexity, it has at least and at most one minimum. In fact, if $\lim_{x \to 0^+} f(x) = 0^+$ now $\lim_{x \to +\infty} f_x(x) = +\infty$.

Exercise 2.2

Point a) If $i_k \in \{1, \dots, n\}$ is picked uniformly at random then

$$\mathbb{E}\left[\nabla f_{i_k}(w)\right] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i_k}(w) = \nabla f(w)$$

As seen on exercise 1.g, $L = \|A\|_F^2 + \mu$ for the function f, f_{i_k} is a part of f and from the matrix A we consider only the element a_{i_k}, hence now is $L = \|a_{i_k}\|^2 + \mu$

Exercise 33.1

a)

- $fe_1(\alpha) = \frac{1}{2}\|b - P_\Omega W^T \alpha\|_2^2$

We can see fe_1 as a composition of two functions:

$fe_1 = g \circ h$ where $g(\alpha) = \frac{1}{2}\|\alpha\|_2^2$ and $h(\alpha) = b - P_\Omega W^T \alpha$

We calculate their Jacobians and apply the chain rule:

$$J_h(\alpha) = -P_\Omega W^T \qquad J_g(\alpha) = \alpha^T$$

Hence:

$$J_{fe_1}(\alpha) = J_g(h(\alpha)) \cdot J_h(\alpha) = (b - P_\Omega W^T \alpha)^T \cdot (-P_\Omega W^T)$$

The gradient $\nabla fe_1(\alpha) = J_{fe_1}(\alpha)^T$:

$$\nabla fe_1(\alpha) = W P_\Omega^T \cdot (P_\Omega W^T \alpha - b)$$

- We follow the same principles for $f_{TV}(x) = \frac{1}{2}\|b - P_\Omega x\|_2^2$

$$g(\alpha) = \frac{1}{2}\|\alpha\|_2^2 \Rightarrow J_g(\alpha) = \alpha^T$$

$$h(\alpha) = b - P_\Omega \alpha \Rightarrow J_h(\alpha) = -P_\Omega$$

$$J_{f_{TV}}(x) = J_g(h(x)) \cdot J_h(x) = (b - P_\Omega x)^T \cdot (-P_\Omega)$$

$$\Rightarrow \nabla f_{TV}(x) = J_{f_{TV}}(x)^T = P_\Omega^T \cdot (P_\Omega x - b)$$

b)

From slide 32/41 of recitation 2 we get that:

Let $f: \Omega \to \mathbb{R}$ be differentiable and convex, then f has a Lipschitz gradient if there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \forall x, y \in \Omega$$

Hence: $\|\nabla fe_1(x) - \nabla fe_1(y)\|_2 = \|W P_\Omega^T (P_\Omega W^T x - b) - W P_\Omega^T (P_\Omega W^T y - b)\|_2 =$

$\|W P_\Omega^T P_\Omega W^T x - W P_\Omega^T b - W P_\Omega^T P_\Omega W^T y + W P_\Omega^T b\|_2 =$

$\|W P_\Omega^T P_\Omega W^T (\alpha - \beta)\|_2 \leq \|W P_\Omega^T P_\Omega W^T\|_{2\to 2} \|\alpha - \beta\|_2$ (Cauchy-Schwarz)

This brings us to the conclusion that $L_{e_1} = \| W P_\Omega^T P_\Omega W^T \|_{2\to 2} = \|I\|_{2\to 2} = 1$

since W is an orthonormal basis ($W^T W = I$) and $P_\Omega^T P_\Omega = I$

- Similarly, for $f_{TV}(x)$ we have:

$$\|\nabla f_{TV}(x) - \nabla f_{TV}(y)\|_2 = \| P_\Omega^T (P_\Omega x - b) - P_\Omega^T (P_\Omega y - b) \|_2$$

$$= \| P_\Omega^T P_\Omega x - P_\Omega^T b - P_\Omega^T P_\Omega y + P_\Omega^T b \|_2 = \| P_\Omega^T P_\Omega (x-y) \|$$

$$\leq \| P_\Omega^T P_\Omega \|_{2\to 2} \| x - y \|_2 \quad \text{(Cauchy-Schwarz)}$$

Hence $L_{e_{TV}} = \| P_\Omega^T P_\Omega \|_{2\to 2} = \|I\|_{2\to 2} = 1$

Exercise 2

Point 2.1-2.2)

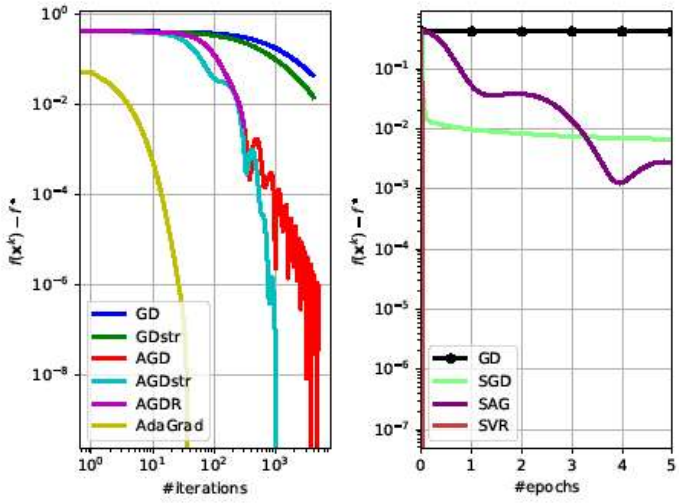

Figure 1

Gradient descent - Time: 32 s, Loss: 0.1386861313868613

Gradient descent with strong convexity - Time: 32 s, Loss: 0.0948905109489051

Accelerated gradient descent – Time: 40 s, Loss: 0.058394160583941604

Accelerated Gradient with strong convexity - Time: 40 s, Loss: 0.058394160583941604

Accelerated Gradient with restart - Time: 4 s, Loss: 0.072992700729927

Adaptive Gradient method – Time:32 s, Loss: 0.058394160583941604

Stochastic Gradient Descent – Time: 7 s, Loss: 0.06569343065693431

Stochastic Gradient Descent with averaging – Time: 7 s, Loss: 0.051094890510948905

Stochastic Gradient Descent with variance reduction – Time: 31 s, Loss: 0.058394160583941604

Among non-stochastic methods it is possible to observe how the method which converges faster is AdaGrad, but stochastic methods do also return models with optimal loss values, and they are also much faster to compute.

Point 2.3)

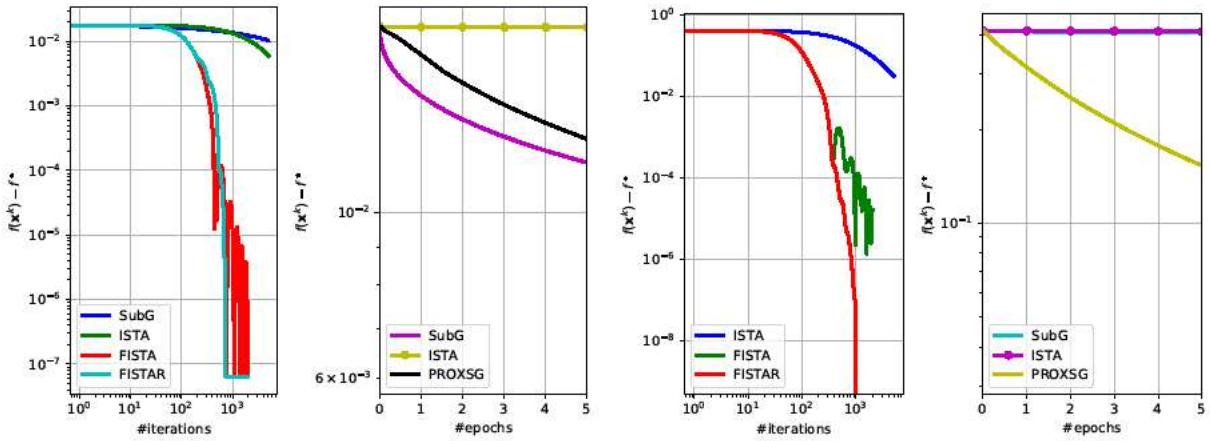

Figure 2 L1-prox (left) L2-prox (right)

Subgradient – Time: 40 s for L1-prox. Loss: 0.12408759124087591 L1-prox

ISTA – Time: 40 s both for L1 and L2 prox. Loss: 0.145985401459854 L1-prox, 0.11678832116788321 L2-prox

FISTA – Time: 16 s both for L1 and L2 prox. Loss: 0.145985401459854 L1-prox, 0.06569343065693431 L2-prox

FISTAR – Time: 16 s both for L1 and L2 prox. Loss: 0.145985401459854 L1-prox, 0.058394160583941604 L2-prox

PROXSG- Time: 12s both for L1 and L2 prox. Loss: 0.13138686131386862 L1-prox, 0.1386861313868613 L2-prox

With proximal methods it can be observed that the faster method to compute is PROXSG. On the other hand, FISTA and FISTAR do provide a much better loss score in the case of L2-prox methods, and they are also not too slow to compute (16s).

Point 2.4)

a)

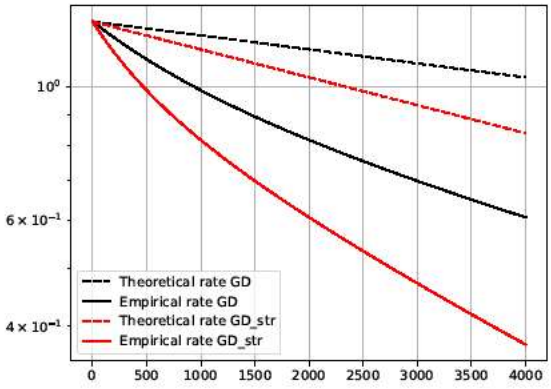

Figure 4 Convergence rates

Considering that the function is L-smooth and μ-strongly convex, we expected a linear convergence rate both for GD and GD strong methods. The empirical results are in fact linear, but they are slightly faster, which makes sense because theoretical rates are just an upper bound. Therefore, observed convergence rates are consistent with the theoretical ones.

b) As stated in the previous point, the convergence rates are linear. This can be concluded by knowing that in plots where the y-axis is logarithmic, as in this case, linear rates have a straight line behavior, due to the following property of linear converging sequences (u_1, u_2, …, u_k, …) :

$$||u_k - u_*|| = O(\alpha^k)$$

When we apply the logarithmic transformation, we obtain the semilogy plot of the upper theoretical bound as the line logy = k*log(α).

c-d)

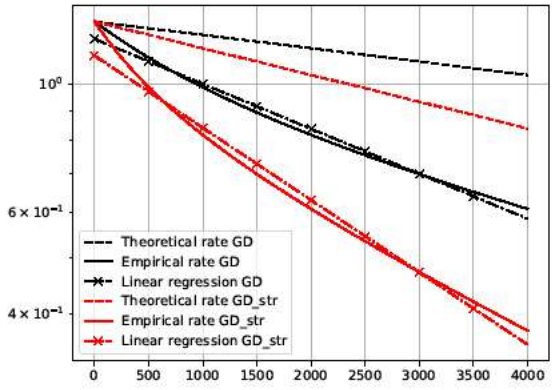

Figure 5 Linear regression

The linear regression on GD returns a slope of -0.00017967907472995833 and a rate of 0.9999465952157935

The linear regression on GD_str returns a slope of -0.0002886563918928778 3and a rate of 0.9998931932836581

Converting those results we obtain a and b coefficients:

GD: a_GD = 1.196692, b_GD = 0.999820

GD theoretical rate: a_GD = 1.282909, b_GD = 0.999947

GD rate diff = -0.00012625814920530498

GD_str: a_GD =1.119548, b_GD = 0.999711

GD_str theoretical rate: a_GD = 1.282909, b_GD = 0.9998931932836581

GD_str rate diff = -0.000182

Since the number of iterations is not high enough to observe the convergence at the empirical asymptotic rate, regression's result converges slightly faster. In both cases, empirical convergence rate is faster than the theoretical upper bound, as expected.