EE-556 Homework 1

Edoardo Debenedetti

October 30, 2019

Problem 1 - Geometric properties of the objective function f

Assuming $\mu = 0$, the smooth Hinge loss function f becomes:

$$f(x) = \ell_{sh}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|^2 \tag{1}$$

where

$$\ell_{sh} = \frac{1}{n}\sum_{i=1}^{n} g_i(\mathbf{x}) \tag{2}$$

and

$$g_i(\mathbf{x}) = \begin{cases} \frac{1}{2} - b_i(\mathbf{a}_i^T\mathbf{x}) & b_i(\mathbf{a}_i^T\mathbf{x}) < 0 \\ \frac{1}{2}(1 - b_i(\mathbf{a}_i^T\mathbf{x}))^2 & 0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1 \\ 0 & 1 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \end{cases} \tag{3}$$

(a) Gradient of f

Computation of the gradient

Proof. Since the gradient is a linear operator:

$$\nabla f(\mathbf{x}) = \nabla \ell_{sh} + \nabla \frac{\lambda}{2}\|\mathbf{x}\|^2 \tag{4}$$

We can first compute $\nabla\frac{\lambda}{2}\|\mathbf{x}\|$:

$$\nabla\frac{\lambda}{2}\|\mathbf{x}\|^2 = \frac{\lambda}{2}\nabla\|\mathbf{x}\|^2 = \frac{\lambda}{2}\nabla\sum_{i=1}^{n}|x_i|^2 = \frac{\lambda}{2}\sum_{i=1}^{n}\nabla x_i^2 = \\ = \frac{\lambda}{2}2\mathbf{x} = \lambda\mathbf{x} \tag{5}$$

Now, let us compute $\nabla\ell_{sh}$:

$$\nabla\ell_{sh} = \nabla\frac{1}{n}\sum_{i=1}^{n} g_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\nabla g_i(\mathbf{x}) \tag{6}$$

Where $\nabla g_i(\mathbf{x})$ is the gradient of (3)

$$\nabla g_i(\mathbf{x}) = \begin{cases} \nabla\left[\frac{1}{2} - b_i(\mathbf{a}_i^T\mathbf{x})\right] & b_i(\mathbf{a}_i^T\mathbf{x}) < 0 \\ \nabla\left[\frac{1}{2}(1 - b_i(\mathbf{a}_i^T\mathbf{x}))^2\right] & 0 \le b_i(\mathbf{a}_i^T\mathbf{x}) \le 1 \\ \nabla 0 & 1 \le b_i(\mathbf{a}_i^T\mathbf{x}) \end{cases} \tag{7}$$

The case where $1 \le b_i(\mathbf{a}_i^T\mathbf{x})$ is trivial, since

$$\nabla 0 = 0 \tag{8}$$

In the case where $b_i(\mathbf{a}_i^T\mathbf{x}) < 0$:

$$\nabla\left[\frac{1}{2} - b_i(\mathbf{a}_i^T\mathbf{x})\right] = \nabla(-b_i(\mathbf{a}_i^T\mathbf{x})) = -b_i\mathbf{a}_i \tag{9}$$

Next, in the case where $0 \le b_i(\mathbf{a}_i^T\mathbf{x}) \le 1$:

$$\nabla\left[\frac{1}{2}(1 - b_i(\mathbf{a}_i^T\mathbf{x}))^2\right] = -\frac{1}{2}2b_i\mathbf{a}_i(1 - b_i(\mathbf{a}_i^T\mathbf{x})) = \\ -b_i\mathbf{a}_i(1 - b_i(\mathbf{a}_i^T\mathbf{x})) = b_i\mathbf{a}_i(b_i(\mathbf{a}_i^T\mathbf{x}) - 1) \tag{10}$$

Finally, combining (8), (9) and (10), we get:

$$\nabla g_i(\mathbf{x}) = \begin{cases} -b_i\mathbf{a}_i & b_i(\mathbf{a}_i^T\mathbf{x}) < 0 \\ b_i\mathbf{a}_i(b_i(\mathbf{a}_i^T\mathbf{x}) - 1) & 0 \le b_i(\mathbf{a}_i^T\mathbf{x}) \le 1 \\ 0 & 1 \le b_i(\mathbf{a}_i^T\mathbf{x}) \end{cases} \tag{11}$$

Now, let us define, as in the problem statement, $\tilde{\mathbf{A}} := [b_1\mathbf{a}_1, ..., b_n\mathbf{a}_n]^T$, and $\mathbf{I}_L, \mathbf{I}_Q$ as the diagonal $n \times n$ matrices such that $\mathbf{I}_L(i,i) = 1$ if $b_i(\mathbf{a}_i^T\mathbf{x}) < 0$ and $\mathbf{I}_Q(i,i) = 1$ if $0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1$, and 0 otherwise.

We can observe that $\tilde{\mathbf{A}}^T\mathbf{I}$ is the matrix whose i-th column is $b_i\mathbf{a}_i$. Instead, $\tilde{\mathbf{A}}^T\mathbf{I}_L$'s i-th columns will be non-zero only in the case where $b_i(\mathbf{a}_i^T\mathbf{x}) < 0$. Then it is possible to represent this case of $\nabla g_i(\mathbf{x})$ where $b_i(\mathbf{a}_i^T\mathbf{x}) < 0$ as

$$-\frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_L\mathbf{1} \tag{12}$$

since multiplying $\tilde{\mathbf{A}}^T\mathbf{I}_L$ by $\mathbf{1}$ will give as result the vector containing the sum of the elements of each column, which means the element-wise sum of the different j-th components of the i-th gradients relative to each $g_i(\mathbf{x})$. Each j-th component can be written as

$$\sum_{i\ \in\ \{i\ |\ b_i(\mathbf{a}_i^T\mathbf{x})\ <\ 0\}}^{n} a_{i,j}b_j$$

In a similar fashion, $\tilde{\mathbf{A}}^T\mathbf{I}_Q$ is the matrix whose i-th column is $\mathbf{a}_i b_i$ only if i is such that $0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1$. Moreover, $\tilde{\mathbf{A}}\mathbf{x}$ is the vector such that $[\tilde{\mathbf{A}}\mathbf{x}]_n = \sum_{i=1}^n b_i\mathbf{a}_i\mathbf{x}$. Consequently, $\tilde{\mathbf{A}}\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}]$ is the vector whose j-th component is

$$[\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}]]_j = \sum_{i\ \in\ \{i\ |\ 0\ \leq\ b_i(\mathbf{a}_i^T\mathbf{x})\ \leq\ 1\}}^{n} b_j a_{i,j}(b_j(\mathbf{a}_i^T\mathbf{x}) - 1)$$

if $0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1$. Then, with

$$\frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}] \tag{13}$$

we can represent the components of $\nabla g_i(\mathbf{x})$ in the aforementioned case.

Combining (12) and (13), it is proven that

$$\nabla \ell_{sh} = \frac{1}{n}(\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}] - \tilde{\mathbf{A}}^T\mathbf{I}_L\mathbf{1}) \tag{14}$$

Finally, combining (5) and (14) we get the final result

$$\nabla f(\mathbf{x}) = \lambda\mathbf{x} + \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}] - \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_L\mathbf{1} \tag{15}$$

$\square$

L-Lipschitz continuity of the gradient

Proof. By definition, a function f has L-Lipschitz continuous gradient if $\exists L < \infty$ such that:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \tag{16}$$

So, let us compute the left term of the inequality for our objective function f:

$$\begin{gathered}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| = \\ \left\|\lambda\mathbf{x} + \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}] - \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_L\mathbf{1} - \left(\lambda\mathbf{y} + \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{y} - \mathbf{1}] - \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_L\mathbf{1}\right)\right\|\end{gathered} \tag{17}$$

We can then observe that the linear parts cancel, that we can take out lambda and expand the expressions in the quadratic region. Eq. 17 becomes:

$$\left\|\lambda(\mathbf{x} - \mathbf{y}) + \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q\tilde{\mathbf{A}}\mathbf{x} - \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q\tilde{\mathbf{A}}\mathbf{y} - \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q + \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q\right\| \tag{18}$$

Again, we can cancel the last two factors, and take out the factor $\frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q\tilde{\mathbf{A}}$. We can also note that since we are now dealing only with elements in the quadratic region and there is no contribution from elements in the linear region, we can consider $\mathbf{I}_Q$ as $\mathbb{I}$ and then we can cancel it. As a consequence, eq. 18 becomes:

$$\begin{gathered}\left\|\lambda(\mathbf{x} - \mathbf{y}) + \frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}(\mathbf{x} - \mathbf{y})\right\| = \\ \left\|\left(\lambda + \frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)(\mathbf{x} - \mathbf{y})\right\|\end{gathered} \tag{19}$$

We can now use Cauchy-Schwartz and triangle inequalities:

$$\begin{gathered}\left\|\left(\lambda + \frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)(\mathbf{x} - \mathbf{y})\right\| \leq \left\|\lambda + \frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right\|\|\mathbf{x} - \mathbf{y}\| \leq \\ \leq \left(\|\lambda\| + \left\|\frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right\|\right)\|\mathbf{x} - \mathbf{y}\| = \\ \left(\lambda + \frac{1}{n}\|\tilde{\mathbf{A}}^T\|\|\tilde{\mathbf{A}}\|\right)\|\mathbf{x} - \mathbf{y}\|\end{gathered} \tag{20}$$

Since λ is a scalar, its norm is the number itself. Moreover, since $\frac{1}{n}$ is a scalar as well, we can take it out of the norm. We can now combine equations 16, 19 and 20 and get the following result:

$$\left(\lambda + \frac{1}{n}\|\tilde{\mathbf{A}}^T\|\|\tilde{\mathbf{A}}\|\right)\|\mathbf{x} - \mathbf{y}\| \leq L\|\mathbf{x} - \mathbf{y}\| \tag{21}$$

if $L = \lambda + \frac{1}{n}\|\tilde{\mathbf{A}}^T\|\|\tilde{\mathbf{A}}\|$. Finally, recalling that $\tilde{\mathbf{A}} := [b_1\mathbf{a}_1, ..., b_n\mathbf{a}_n]^T$ where $b_n \in \{-1, 1\}$, we can note that $\|\tilde{\mathbf{A}}\| = \|\mathbf{A}\|$, since the norm is computed taking in account the absolute value of each entry of a matrix. Hence, as a final result,

$$f(\mathbf{x}) \in \mathcal{F}_L^{1,1} \tag{22}$$

with $L = \lambda + \frac{1}{n}\|\mathbf{A}^T\|\|\mathbf{A}\|$.

$\square$

(b) Hessian of f

Proof. Assuming that $\mathbf{I}_L = \mathbb{I}$, we can deduce that $\mathbf{I}_L = \mathbb{0}$, since it would mean that $\forall i \in [1,\ n],\ b_i(\mathbf{a}_i^T\mathbf{x}) < 0$. Then some simple computations can show that

$$\nabla f(\mathbf{x}) = \lambda\mathbf{x} + \frac{1}{n}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\mathbf{x}) - \tilde{\mathbf{A}}^T \tag{23}$$

We can then compute the Hessian $\nabla^2 f(\mathbf{x})$ as $\nabla \cdot \nabla f(\mathbf{x})$, that is

$$\begin{aligned}\nabla^2 f(\mathbf{x}) = \nabla \cdot \nabla f(\mathbf{x}) = \nabla \cdot \lambda\mathbf{x} + \nabla \cdot \left[\frac{1}{n}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\mathbf{x}) - \tilde{\mathbf{A}}^T\right] &= \\ = \lambda\nabla \cdot \mathbf{x} + \frac{1}{n}\tilde{\mathbf{A}}^T(\tilde{\mathbf{A}}\nabla \cdot \mathbf{x}) &= \\ = \lambda\mathbb{I} + \frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} & \qquad (24)\end{aligned}$$

Hence, $\nabla^2 f(\mathbf{x}) = \lambda\mathbb{I} + \frac{1}{n}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$. Moreover, $f(\mathbf{x})$ is twice differentiable because $\nabla^2 f(\mathbf{x})$ is continuous over $\mathbb{R}^p$

$\square$

(c) Strong convexity of f

Proof. First, let use recall that $f(\mathbf{x}) = \ell_{sh}(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|^2$ and that a function $f(\mathbf{x})$ is μ-strongly convex iff, given $h(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$, $h(\mathbf{x})$ is convex. In the case of the smooth Hinge loss function,

$$h(\mathbf{x}) = \ell_{sh} + \frac{\lambda}{2}\|\mathbf{x}\|^2 - \frac{\mu}{2}\|\mathbf{x}\|^2 \tag{25}$$

Now, setting $\mu = \lambda$, we get that $h(\mathbf{x}) = \ell_{sh}(x)$. We know that ℓ_{sh} is convex, and then $h(\mathbf{x})$ us convex as well. Thus,

$$f(\mathbf{x}) \in \mathcal{F}^{2,1}_{L,\mu} \tag{26}$$

with $L = \lambda + \frac{1}{n}\|\mathbf{A}^T\|\|\mathbf{A}\|$ and $\mu = \lambda$.

$\square$

First order methods for linear SVM

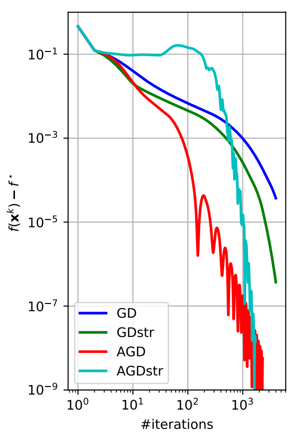

Figure 1: (Accelerated) Gradient Descent

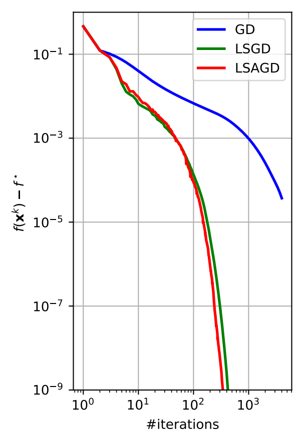

Figure 2: Line Search methods

Stochastic gradient methods for SVM

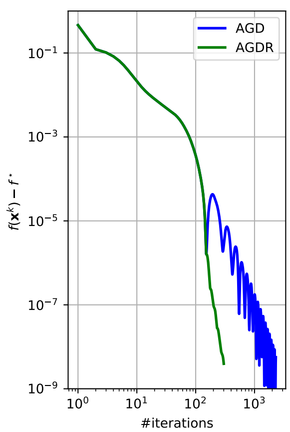

Figure 3: Accelerated GD with restart

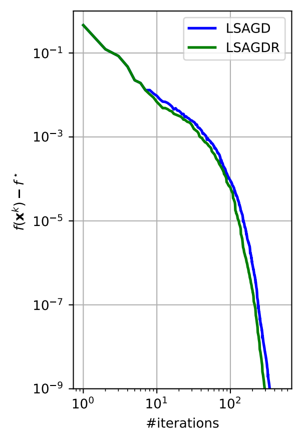

Figure 4: Line Search AGD with restart

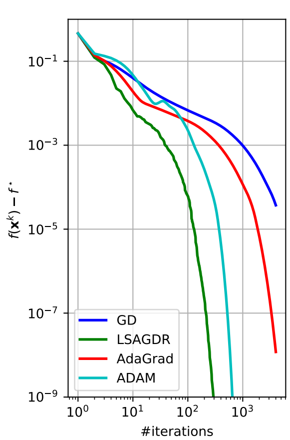

Figure 5: Adaptive methods

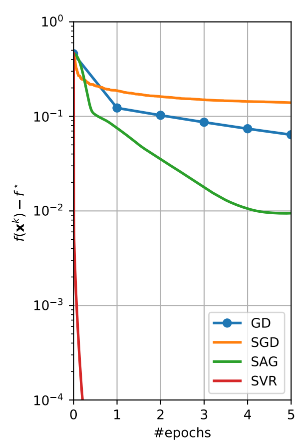

Figure 6: Stochastic Gradient methods