EE-556 Homework 1

Edoardo Debenedetti

November 1, 2019

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 $\nabla \frac{\lambda}{2}\|\mathbf{x}\|^2$:

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

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

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

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

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

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

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

□

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:

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

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:

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

We can now use Cauchy-Schwartz and triangle inequalities:

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

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

□

(b) Hessian of f

Proof. Assuming that $\mathbf{I}_L = \mathbb{I}$, we can deduce that $\mathbf{I}_Q = \mathbb{0}$, 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

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

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 $\mathbb{R}^p$ (as a matter of fact, it is constant w.r.t. $\mathbf{x}$).

□

(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 $\mu = \lambda$.

□

2 First order methods for linear SVM

Methods implementations

(Accelerated) Gradient Descents

From figure 1 we can see that:

- Assuming strong convexity for both Gradient Descent (GDstr) and Accelerated Gradient Descent (AGDstr) gives a significant advantage in the long run (i.e. after 10^3 iterations).
- Accelerated (AGD and AGDstr) methods are quite unstable and have several jumps, which partially cancel the advantage of acceleration, especially in the case with strong convexity assumptions. We can observe that, indeed, $f(\mathbf{x}^k) - f^\star$ increases between the 2^{nd} and the $\sim 100^{th}$ iteration. This might be due to the fact that the $\mathbf{y}^{k+1}$ *stepsize* is constant with k assuming strong convexity, while in the other case it increases. This feature, combined with the local geometry of the objective function f (e.g. *narrowness*) could lead to the aforementioned increase.

Line Search Methods

From figure 4 we can see that line-search to adapt the step-size to the local geometry makes the loss functions converge with a higher rate, both with Gradient Descent (LSGD) and Accelerated Gradient Descent (LSAGD). However, we should keep in mind that line-search is computationally expensive and then makes each iteration slower.

Restart methods

From figures 3 and 4 it is evident that, in case of AGD and LSAGD, restart (with AGDR and LSAGDR) gives a huge advantage (especially without line search, which is less computationally expensive) at no computational cost.

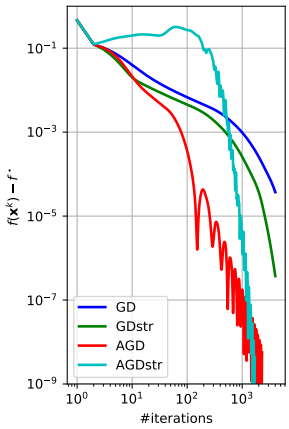

Figure 1: (Accelerated) Gradient Descent

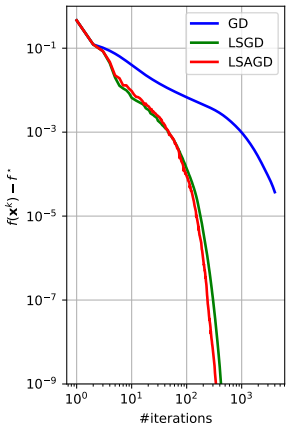

Figure 2: Line Search methods

Adaptive Gradient methods

Figure 5 shows that adaptive methods such as AdaGrad and ADAM have slower convergence rates than line search adaptive gradient descent with restart (LSAGDR). However, it is worth noting that LSAGDR makes use of the Lipschitz-smoothness constant L, which can computationally expensive (or not possibile at all) to retrieve. Thus, in case L is hard to compute, or is not available, ADAM and AdaGrad can provide significant improvements in convergence rates with respect to regular Gradient Descent.

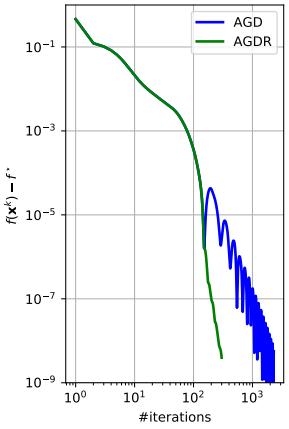

Figure 3: Accelerated GD with restart

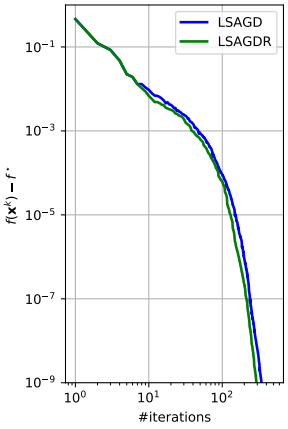

Figure 4: Line Search AGD with restart

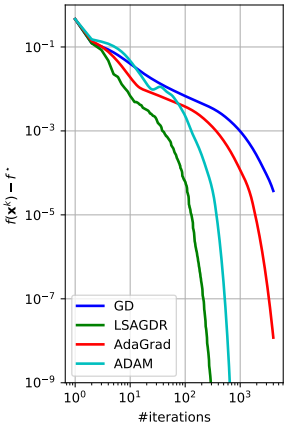

Figure 5: Adaptive Gradient methods

3 Stochastic gradient methods for SVM

Stochastic Gradient properties

Unbiased estimation of stochastic gradient

Proof. In order to prove that $\nabla f_{ik}(\mathbf{x})$ is an unbiased estimate of $\nabla f(\mathbf{x})$, we can take the expectation of $\nabla f_{ik}(\mathbf{x})$. Since the i-th gradients are chosen uniformly at random, each $\nabla f_{ik}(\mathbf{x})$ has the same probability to be drawn, then $P\{\nabla f_{ik}(\mathbf{x})\} = \frac{1}{n}\ \forall i$. Thus,

$$\mathbb{E}[\nabla f_{ik}(\mathbf{x})] = \sum_{i=1}^{n} \frac{1}{n} \nabla f_{ik}(\mathbf{x}) =$$

$$= \frac{1}{n}\lambda\mathbf{x} + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1\}} \mathbf{a}_i(\mathbf{a}_i^T\mathbf{x} - b_i) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{b_i(\mathbf{a}_i^T\mathbf{x}) < 0\}} b_i\mathbf{a}_i \tag{27}$$

We can now work on the central term of eq. (27),

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1\}} \mathbf{a}_i(\mathbf{a}_i^T\mathbf{x} - b_i) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1\}} \mathbf{a}_i(b_i^2\mathbf{a}_i^T\mathbf{x} - b_i) = \tag{28}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1\}} \mathbf{a}_i b_i(\mathbf{a}_i^T\mathbf{x} - 1) \tag{29}$$

Note that in (28), we multiplied $\mathbf{a}_i^T\mathbf{x}$ by b_i^2 since $b_i \in \{-1, 1\}$, and then $b_i^2 = 1\ \forall i$. (27) then, becomes

$$\frac{1}{n}\lambda\mathbf{x} + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1\}} \mathbf{a}_i b_i(\mathbf{a}_i^T\mathbf{x} - 1) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{b_i(\mathbf{a}_i^T\mathbf{x}) < 0\}} b_i\mathbf{a}_i = \tag{30}$$

$$= \lambda\mathbf{x} + \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_Q[\tilde{\mathbf{A}}\mathbf{x} - \mathbf{1}] - \frac{1}{n}\tilde{\mathbf{A}}^T\mathbf{I}_L\mathbf{1} = \nabla f(\mathbf{x}) \tag{31}$$

To go from (30) to (31), we can use the same intuitions we used in the proof of $\nabla f(\mathbf{x})$, in the Computation of the gradient section.

□

L-Lipschitz continuity of the stochastic gradient

Again, in order to prove L-Lipschitz continuity of the stochastic gradient, we use the definition of L-Lipschitz continuity of the gradient of a function, that can be found in eq. 16. We then start computing $\|\nabla f_{ik}(\mathbf{x}) - \nabla f_{ik}(\mathbf{y})\|$

$$\begin{aligned} &\|\nabla f_{ik}(\mathbf{x}) - \nabla f_{ik}(\mathbf{y})\| = \\ &= \|\lambda(\mathbf{x}-\mathbf{y}) + \mathbf{1}_{\{0\leq b_i(\mathbf{a}_i^T\mathbf{x})\leq 1\}}\mathbf{a}_i(\mathbf{a}_i^T\mathbf{x} - b_i) - \mathbf{1}_{\{0\leq b_i(\mathbf{a}_i^T\mathbf{x})\leq 1\}}\mathbf{a}_i(\mathbf{a}_i^T\mathbf{y} - b_i) + \\ &\quad + \mathbf{1}_{\{b_i(\mathbf{a}_i^T\mathbf{x})<0\}}b_i\mathbf{a}_i - \mathbf{1}_{\{b_i(\mathbf{a}_i^T\mathbf{x})<0\}}b_i\mathbf{a}_i\| \end{aligned} \tag{32}$$

We can then note that the components of the linear region. Consequently, we are only concerned with $\{i \mid 0 \leq b_i(\mathbf{a}_i^T\mathbf{x}) \leq 1\}$. Then, we can consider $\mathbf{1}_{\{0\leq b_i(\mathbf{a}_i^T\mathbf{x})\leq 1\}}$ as $\mathbf{1}$ and cancel it as well. Eq. (32) becomes:

$$\begin{aligned} \|\lambda(\mathbf{x}-\mathbf{y}) + \mathbf{a}_i\mathbf{a}_i^T\mathbf{x} - \mathbf{a}_i\mathbf{a}_i^T\mathbf{y}\| = \|\lambda(\mathbf{x}-\mathbf{y}) + \mathbf{a}_i\mathbf{a}_i^T(\mathbf{x}-\mathbf{y})\| = \\ = \|(\lambda + \mathbf{a}_i\mathbf{a}_i^T)(\mathbf{x}-\mathbf{y})\| \end{aligned} \tag{33}$$

We can now apply Cauchy-Schwartz on (33):

$$\|(\lambda + \mathbf{a}_i\mathbf{a}_i^T)(\mathbf{x}-\mathbf{y})\| \leq \|\lambda + \mathbf{a}_i\mathbf{a}_i^T\|\|\mathbf{x}-\mathbf{y}\|$$

Since $\mathbf{a}_i\mathbf{a}_i^T = \|\mathbf{a}_i\|^2$ is a scalar, as well as λ, the norm of their sum is their sum itself. Hence:

$$\|(\lambda + \mathbf{a}_i\mathbf{a}_i^T)(\mathbf{x}-\mathbf{y})\| \leq (\lambda + \|\mathbf{a}_i\|^2)\|\mathbf{x}-\mathbf{y}\| \tag{34}$$

Which satisfies the definition (16) with $L = \lambda + \|\mathbf{a}_i\|^2$.

Methods implementations

Stochastic Gradient Descent

We can see from figure 6 that SGD converges significantly faster than Gradient Descent in the 1^{st} epoch, but then it slows down. This is due to the fact that the convergence rate of SGD is sublinear ($\frac{1}{\sqrt{k}}$), while that of GD is linear (ρ^k). As a matter of fact, SGD is more powerful with a large n (which corresponds to larger epochs). Moreover, it has been possible to notice a high variance in the convergence doing doing different trainings: in some cases SGD kept faster than GD even after the first epoch, in other cases it slowed down significantly before.

Stochastic Averaged Gradient

SAG can keep fast until about the 4^{th} epoch. This first advantage is given by the averaging performed that makes SAG more stable than SGD. However, after the 4^{th} epoch, it slows down at the same rate as SGD does.

Stochastic Gradient Descent with Variance Reduction

In this case, the plot of SVR in figure 6 is quite misleading. In fact, even though SVR converges in very few *epochs*, it is worth noting that each *epoch* of SVR corresponds to 1 complete pass over the gradient (that correspond to 1 *real* epoch), plus $q \approx 9700$ gradients to be computed in the variance reduction phase (corresponding to an equivalent of 17.75 *true* epochs in the case of a dataset counting $n = 546$ datapoints) with a total of about 18.75 *true* epochs each iteration. However, it can be seen from figure 6, that SVR converges very fast in very few iterations.

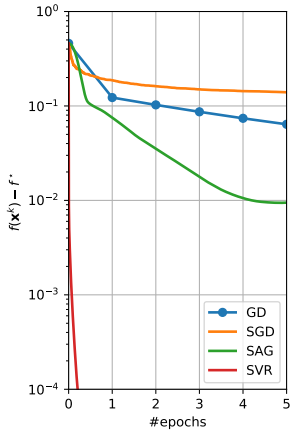

Figure 6: Stochastic Gradient methods

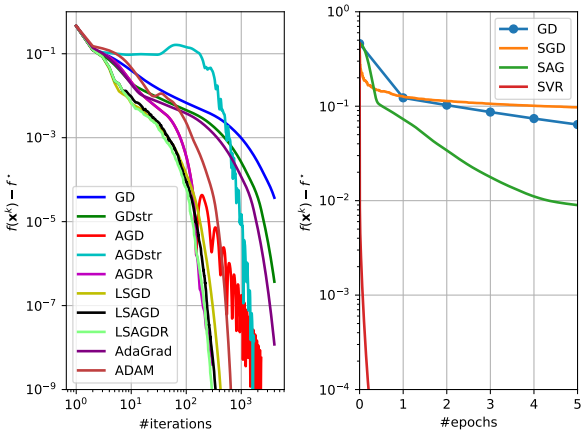

Figure 7: A comprehensive plot with all the methods