[SD-TSIA 211] Optimization for Machine Learning
"TD-03 Corrections"

Olivier FERCOQ — Iyad WALWIL

December 16, 2022

Recalls

Definition 1. *(Sub-differential)*
Let $f\colon \mathcal{X} \to [-\infty, +\infty]$ and $x \in \operatorname{dom} f$. A vector $\phi \in \mathcal{X}$ is called a ***sub-gradient*** *of f at x if:*

$$\forall y \in \mathcal{X}, \quad f(y) - f(x) \geq \langle \phi, y - x \rangle \tag{1}$$

Theorem 1. *(Fermat's rule)*

$$x \in \arg\min f \iff 0 \in \partial f(x) \tag{2}$$

Definition 2. *Operator norm*
Let $B\colon V \to W$ be a linear operator between two normed spaces, the operator norm of B, denoted $\|B\|_{op}$, is defined as:

$$\|B\|_{op} = \sup \left\{ \frac{\|Bv\|}{\|v\|} \ : \ v \neq 0, v \in V \right\} \tag{3}$$

The following inequality is an immediate consequence of the definition:

$$\|Bv\| \leq \|B\|_{op} \|v\| \ \ \forall v \in V \tag{4}$$

Definition 3. *Separable function*
We say that a function $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is separable if there exists n functions $\varphi_i\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that $\forall x \in \mathbb{R}^n$, $\varphi(x) + \sum_{i=1}^n \varphi_i(x_i)$

Proposition 1. *Property of separable functions*
If $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a separable function, then

$$\mathrm{Prox}_{\gamma\varphi}(x) = (\mathrm{Prox}_{\gamma\varphi_1}(x_1), \ldots\ldots, \mathrm{Prox}_{\gamma\varphi_n}(x_n)) \tag{5}$$

Sub-Exercise 1. *Find the sub-differential of the absolute value.* $f(x) = |x|$.

Solution.

$$f(x) = |x| = \begin{cases} -x & , x < 0 \\ x & , x \geq 0 \end{cases}$$

For any $x < 0, f(x) = -x$ *which is differentiable with* $\partial f(x) = f'(x) = -1$. *Similarly, for any* $x > 0, f(x) = x \Rightarrow \partial f(x) = f'(x) = 1$. *The only issue is at* $x = 0$ *where the function is non-differentiable. By the definition of the sub-differential:*

$$\begin{aligned} q \in \partial f(0) &\iff \forall u \in \mathbb{R}, \quad f(u) \geq f(0) + q(u - 0) \\ &\iff |u| \geq qu \\ &\iff -1 \leq q \leq 1 \\ &\iff |q| \leq 1 \end{aligned}$$

Thus,

$$\partial |x| = \begin{cases} -1 & , x < 0 \\ [-1, 1] & , x = 0 \\ 1 & , x > 0 \end{cases} \tag{6}$$

Exercise 1. *(LASSO).*

We consider the problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

1. Prove that the solution is $\{0\}$ for large λ.
2. For an arbitrary λ, provide the expression of the proximal gradient algorithm, using the step size $\gamma_k = \gamma = \frac{1}{L}$ where L is the Lipschitz constant of the gradient of the differentiable function in the problem.
3. Assume that the initial point is at distance D from a minimizer. How many iterations are needed (at most) to achieve an ε-minimizer?

Answers

1. The objective function $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$ is strongly convex and coercive, so it has a unique minimizer, say x^*. Thus,

$$\forall x \in \mathbb{R}^n,\ f(x^*) \le f(x) \Rightarrow f(x^*) \le f(0) \tag{7}$$

$$\lambda\|x\|_1 \le \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \tag{8}$$

Putting everything together:

$$\begin{aligned}\lambda\|x\|_1 &\overset{(7)}{\le} \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1\\ &\overset{(8)}{\le} f(0) = \frac{1}{2}\|b\|_2^2\\ &\Rightarrow \|x\|_1 \le \frac{1}{2\lambda}\|b\|_2^2\\ &\text{As } \lambda \to +\infty, \|x\|_1 \le 0 \Rightarrow x = 0 \quad \square\end{aligned}$$

2. Recall that the Proximal Gradient Algorithm (PGA) solves optimization problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) + g(x) \tag{9}$$

where $f(x)$ is differentiable and has an L-Lipschitz gradient, and $g(x)$ has an easy computable proximal. Hence,

$$f(x) = \frac{1}{2}\|Ax-b\|_2^2 \qquad\qquad g(x) = \lambda\|x\|_1$$

- $f(x)$ is differentiable with $\nabla f(x) = A^T(Ax-b)$ and has an $L = \lambda_{\max}(A^TA)$-Lipschitz gradient where $\lambda_{\max}(A^TA)$ is the largest eigenvalue of A^TA.

  Proof. $\forall (x,y) \in \mathbb{R}^{2n}$,

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| &= \|A^T(Ax-b) - A^T(Ay-b)\|\\ &= \|A^TAx - A^TAy\|\\ &= \|A^TA(x-y)\|\end{aligned}$$

$$\overset{(4)}{\leq} \|A^TA\|_{op}\|x-y\| \text{ with } B = A^TA \ \& \ v = x-y$$
$$= \lambda_{\max}(A^TA)\|x-y\| \quad \square$$

- Now, we want to find the proximal of g. Firstly, note that $g(x)$ is a separable function (Definition 3).

$$g(x) = \lambda\|x\|_1 = \sum_{i=1}^{n} \lambda|x_i| = \sum_{i=1}^{n} \varphi(x_i)$$
$$\text{with} \quad \varphi(y) = \lambda|y|$$

Thus, we can use Proposition 1 to find its proximal. That's it:

$$p = \text{Prox}_{\gamma g}(x) = (\text{Prox}_{\gamma\varphi}(x_i))_{1\leq i\leq n} = (p_i)_{1\leq i\leq n}$$

So, all what we have to do is to find $p_i = \text{Prox}_{\gamma\varphi}(x_i)$. For $i \in \{1, \ldots, n\}$

$$\begin{aligned}
p_i = \text{Prox}_{\gamma\varphi}(x_i) &\iff 0 \in \partial\Big(\gamma\lambda|.|_1 + \frac{1}{2}(. - x_i)^2\Big)(p_i) \\
&\iff 0 \in \gamma\lambda\partial|p_i|_1 + (p_i - x_i) \\
&\iff p_i \in x_i - \gamma\lambda\partial|p_i|_1 \\
&\overset{(6)}{\iff} p_i \in x_i - \gamma\lambda \begin{cases} -1 & , p_i < 0 \\ [-1, 1] & , p_i = 0 \\ 1 & , p_i > 0 \end{cases} \\
&\iff p_i \in \begin{cases} x_i + \gamma\lambda & , p_i < 0 \\ x_i + \gamma\lambda[-1, 1] & , p_i = 0 \\ x_i - \gamma\lambda & , p_i > 0 \end{cases} \qquad (10) \\
&\iff p_i \in \begin{cases} x_i + \gamma\lambda & , x_i < -\gamma\lambda \\ 0 & , |x_i| \leq \gamma\lambda \\ x_i - \gamma\lambda & , x_i > \gamma\lambda \end{cases} \qquad (11) \\
&\iff p_i = \Big[|x_i| - \gamma\lambda\Big]_+ \text{sgn}(x_i)
\end{aligned}$$

Where we have moved from (10) to (11) as follows:

In (10), we have:

- $p_i = x_i + \gamma\lambda$ whenever $p_i < 0$. Thus, $x_i + \gamma\lambda < 0 \iff x_i < -\gamma\lambda$.
- Similarly, $p_i = x_i - \gamma\lambda$ whenever $p_i > 0$. Thus, $x_i > \gamma\lambda$.
- $p_i \in x_i + \gamma\lambda[-1, 1]$ whenever $p_i = 0$. Thus, $0 \in x_i + \gamma\lambda[-1, 1] \iff x_i \in [-\gamma\lambda, \gamma\lambda] \iff |x_i| \leq \gamma\lambda$

Hence,

$$\text{Prox}_{\gamma g}(x) = ([|x_i| - \gamma\lambda]_+ \operatorname{sgn}(x_i))_{1\leq i\leq n} \tag{12}$$

Therefore, the expression of the PGA is:

$$\begin{aligned}
x_{k+1} &= \text{Prox}_{\gamma g}(x_k - \gamma\nabla f(x_k)) \\
&= \text{Prox}_{\frac{g}{L}}(x_k - \frac{1}{L}A^T(Ax_k - b)) \\
&= \left(\left[|x_k^i - \frac{1}{L}A_i^T(Ax_k - b)| - \frac{\lambda}{L}\right]_+ \operatorname{sgn}(x_k^i - \frac{1}{L}A_i^T(Ax_k - b))\right)_{1\leq i\leq n}
\end{aligned}$$

where A_i^T is the i^{th} row of A^T. □

3. From Theorem 3.4.1 (lecture notes), we know that the PGA with $\gamma = \frac{1}{L}$ satisfies:

$$f(x_k) + g(x_k) - f(x^*) - g(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}$$

We are assuming that the initial point x_0 is at distance D from a minimizer x^*, i.e., $\|x_0 - x^*\| \leq D$. Thus,

$$(x_k) + g(x_k) - f(x^*) - g(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k} \leq \frac{LD^2}{2k}$$

To find ε-minimizer, we need number of iterations, k, such that:

$$\frac{LD^2}{2k} \leq \varepsilon \iff 2k\varepsilon \geq LD^2 \iff k \geq \frac{LD^2}{2\varepsilon}$$

Hence, $k = \lceil \frac{LD^2}{2\varepsilon} \rceil$ □

Exercise 2. *(Proximal gradient for logistic regression)*

We consider a classification problem defined by observations $(x_i, y_i)_{1\leq i\leq n}$ where for all i, $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. We propose the following linear model for the generation of the data. Each observation is supposed to be independent and there exists a vector $w \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$ such that for all i, (y_i, x_i) is a realization of the random variable $(\mathbf{Y}, \mathbf{X})$ whose low satisfies:

$$\mathbb{P}_{w,w_0}(\mathbf{Y} = 1|\mathbf{X}) = \frac{\exp(\mathbf{X}^T w + w_0)}{1 + \exp(\mathbf{X}^T w + w_0)}$$

1. Show that $\forall i \in \{1, \dots, n\}$,

$$\mathbb{P}(\mathbf{Y}_i = y_i|x_i) = \frac{1}{1 + \exp(-y_i(x_i^T w + w_0))}$$

2. Show that the maximum likelihood estimator is

$$(\hat{w}, \hat{w}_0) = \arg\min_{w,w_0} \sum_{i=1}^{n} \log(1 + \exp(-y_i(x_i^T w + w_0)))$$

3. Denote $f(w, w_0) = \sum_{i=1}^{n} \log(1+\exp(-y_i(x_i^T w+w_0)))$. Compute $\nabla f(w, w_0)$

4. Compute the proximal operator of $(x \mapsto \frac{\lambda}{2}\|x\|^2)$

5. Write the proximal gradient method for the logistic regression problem with ridge regularizer:

$$(\hat{w}^{(\lambda)}, \hat{w}_0^{(\lambda)}) = \arg\min_{w,w_0} \sum_{i=1}^{n} \log(1 + \exp(-y_i(x_i^T w + w_0))) + \frac{\lambda}{2}\|w\|^2$$

Answers:

1. y_i takes values in $\{-1, 1\}$ only, as $\mathbb{P}_{w,w_0}(\mathbf{Y} = 1|\mathbf{X})$ is given, all what we have to find is $\mathbb{P}_{w,w_0}(\mathbf{Y} = -1|\mathbf{X})$.

$$
\begin{aligned}
\mathbb{P}(\mathbf{Y} = -1|\mathbf{X}) &= 1 - \mathbb{P}(\mathbf{Y} = 1|\mathbf{X}) \\
&= 1 - \frac{\exp(\mathbf{X}^T w + w_0)}{1 + \exp(\mathbf{X}^T w + w_0)} \\
&= \frac{1}{\exp(\mathbf{X}^T w + w_0)}
\end{aligned}
$$

Also, note that:

$$\mathbb{P}(\mathbf{Y} = 1|X) = \frac{\exp(\mathbf{X}^T w + w_0)}{1 + \exp(\mathbf{X}^T w + w_0)} = \frac{1}{1 + \exp(-(\mathbf{X}^T w + w_0))}$$

Thus,

$$\mathbb{P}(\mathbf{Y}_i = y_i|x_i) = \frac{1}{1 + \exp(-y_i(\mathbf{X}^T w + w_0))}$$

2. Recall the formula of the likelihood functions:

$$\ell_{w,w_0}(x, y) = \mathbb{P}_{w,w_0}(\mathbf{Y} = y|\mathbf{X} = x) \tag{13}$$

Thus,

$$
\begin{aligned}
\ell_{w,w_0}(x, y) &= \mathbb{P}_{w,w_0}(\mathbf{Y} = y|\mathbf{X} = x) \\
&= \prod_{i=1}^{n} \mathbb{P}_{w,w_0}(\mathbf{Y} = y_i|\mathbf{X} = x_i) \quad \text{In-dependant observations} \\
&= \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i(x_i^T w + w_0))}
\end{aligned}
$$

To simplify the calculations, it's convenient to work with:

$$\tilde{\ell}_{w,w_0}(x, y) = \log \ell_{w,w_0}(x, y)$$

As the log is a monotonic function, both functions ℓ & $\tilde{\ell}$ will share the same maximizer. Hence, the maximum likelihood estimator (MLE) is:

$$
\begin{aligned}
(\hat{w}, \hat{w}_0) &= \arg\max_{w,w_0} \tilde{\ell}_{w,w_0}(x, y) \\
&= \arg\max_{w,w_0} \log \left(\prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i(x_i^T w + w_0))} \right)
\end{aligned}
$$

$$= \arg\max_{w,w_0} \sum_{i=1}^{n} \log\left(\frac{1}{1+\exp(-y_i(x_i^T w + w_0))}\right)$$

$$= \arg\max_{w,w_0} \sum_{i=1}^{n} -\log(1+\exp(-y_i(x_i^T w + w_0)))$$

$$= \arg\min_{w,w_0} \sum_{i=1}^{n} \log(1+\exp(-y_i(x_i^T w + w_0)))$$

3. Note that, $f(w, w_0) = \sum_{i=1}^{n} f_i(w, w_0)$, with

$$f_i(w, w_0) = \log(1+\exp(-y_i(x_i^T w + w_0)))$$

Thus,

$$\nabla_w f(w, w_0) = \sum_{i=1}^{n} \nabla_w f_i(w, w_0) \qquad \nabla_{w_0} f(w, w_0) = \sum_{i=1}^{n} \nabla_{w_0} f_i(w, w_0)$$

$$\begin{aligned}
\nabla_w f_i(w, w_0) &= \nabla_w\Big[\log(1+\exp(-y_i(x_i^T w + w_0)))\Big] \\
&= \frac{-y_i x_i \exp(-y_i(x_i^T w + w_0))}{1+\exp(-y_i(x_i^T w + w_0))} \\
&= \frac{-y_i x_i}{1+\exp(y_i(x_i^T w + w_0))}
\end{aligned}$$

$$\begin{aligned}
\nabla_{w_0} f_i(w, w_0) &= \nabla_{w_0}\Big[\log(1+\exp(-y_i(x_i^T w + w_0)))\Big] \\
&= \frac{-y_i \exp(-y_i(x_i^T w + w_0))}{1+\exp(-y_i(x_i^T w + w_0))} \\
&= \frac{-y_i}{1+\exp(y_i(x_i^T w + w_0))}
\end{aligned}$$

4. Compute $\mathrm{Prox}_{\gamma g}(x)$ where $g(x) = \frac{\lambda}{2}\|x\|^2$

$$\begin{aligned}
p = \mathrm{Prox}_{\gamma g}(x) &\iff 0 \in \partial\left(\frac{\gamma\lambda}{2}\|.\|^2 + \frac{1}{2}\|. - x\|^2\right)(p) \\
&\iff 0 \in \gamma\lambda p + (p - x) \\
&\iff p = \frac{1}{\gamma\lambda + 1} x
\end{aligned}$$

5. Write the proximal gradient method for the logistic regression problem with ridge regularizer:

$$
\begin{aligned}
(\hat{w}^{(\lambda)}, \hat{w}_0^{(\lambda)}) &= \arg\min_{w,w_0} \sum_{i=1}^{n} \log(1+\exp(-y_i(x_i^T w + w_0))) + \frac{\lambda}{2}\|w\|^2 \\
&= \arg\min_{w,w_0} f(w,w_0) + g(w)
\end{aligned}
$$

where $f(w, w_0)$ is a differentiable function (question 3) and has an L-Lipschitz gradient (to be computed), and $g(w)$ has an easy computable proximal (question 4). Thus,

$$
\begin{aligned}
w_{k+1}^{(\lambda)} &= \mathrm{Prox}_{\gamma g}\left(w_k^{(\lambda)} - \gamma \nabla f(w_k^{(\lambda)}, w_{0,k}^{(\lambda)})\right) \\
&= \frac{1}{\gamma\lambda+1}\left(w_k^{(\lambda)} - \gamma \nabla f(w_k^{(\lambda)}, w_{0,k}^{(\lambda)})\right) \\
&= \frac{1}{\frac{1}{L}\lambda+1}\left(w_k^{(\lambda)} - \frac{1}{L}\nabla_w f\left(w_k^{(\lambda)}, w_{0,k}^{(\lambda)}\right)\right) \\
&= \frac{L}{L+\lambda}\left[w_k^{(\lambda)} - \frac{1}{L}\nabla_w f\left(w_k^{(\lambda)}, w_{0,k}^{(\lambda)}\right)\right] \\
&= \frac{L}{L+\lambda}\left[w_k^{(\lambda)} - \frac{1}{L}\sum_{i=1}^{n} \frac{-y_i x_i}{1+\exp(y_i(x_i^T w_k^{(\lambda)} + w_{0,k}^{(\lambda)}))}\right]
\end{aligned}
$$

And,

$$
\begin{aligned}
w_{0,k+1}^{(\lambda)} &= \mathrm{Prox}_{\gamma 0}\left(w_{0,k}^{(\lambda)} - \gamma \nabla f(w_k^{(\lambda)}, w_{0,k}^{(\lambda)})\right) \\
&= w_{0,k}^{(\lambda)} - \frac{1}{L}\nabla_{w_0} f\left(w_k^{(\lambda)}, w_{0,k}^{(\lambda)}\right) \\
&= w_{0,k}^{(\lambda)} - \frac{1}{L}\sum_{i=1}^{n} \frac{-y_i}{1+\exp(y_i(x_i^T w_k^{(\lambda)} + w_{0,k}^{(\lambda)}))}
\end{aligned}
$$