Let consider a general mixture model. We observe N random variables $x_1, x_2, ..., x_N$ which are independently and identically distributed with $x_i \sim f_{\boldsymbol{\pi}}(x_i)$ where $f_{\boldsymbol{\pi}}$ is given by:

$$f_{\boldsymbol{\pi}}(x) = \sum_{j=1}^{K} \boldsymbol{\pi}_j f_j(x) \tag{1}$$

Let suppose that each component density f_i is known, but not necessarily Gaussian. We will focus on the estimation of the weights vector $\boldsymbol{\pi} \in \mathbb{R}^K$ and assume that this vector is sparse. We will focus our study on the performance of the Maximum Likelihood Estimator (et excess risk ?). From a rewriting of the loglikelihood we can define $\Phi_N(\boldsymbol{\pi})$ as following:

$$\Phi_N(\boldsymbol{\pi}) = -\frac{1}{N}\sum_{i=1}^{N} \log f_{\boldsymbol{\pi}}(x_i) \tag{2}$$

We can rewrite the minimization problem:

$$\widehat{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi}\in\Pi} \left\{\Phi_N(\boldsymbol{\pi})\right\}, \quad \Pi = \left\{\boldsymbol{\pi} \in [0,1]^K : \sum_{j=1}^{K} \pi_j = 1\right\} \tag{3}$$

For theoretical objectives, we will make the following asumption:

Hypothesis 1. *All realizations are not probably unlikely to be observed. Therefore* $\exists m > 0$ *such that* $f_{\boldsymbol{\pi}}(x) \geq m$ *for all* $x \in \{x_1, \ldots, x_N\}$.

Let denote $M = \max_{x\in\{x_1,\ldots,x_N\}, j\in[K]}\{f_j(x)\}$, since $\boldsymbol{\pi}^T\mathbf{1} = 1$ then $f_{\boldsymbol{\pi}} \leq M$. Therefore, $\forall \boldsymbol{\pi} \in \Pi$, $f_{\boldsymbol{\pi}}(x_1, \ldots, x_N) \in [m, M]$. We have the following lemma:

Lemma 1. *Under hypothesis 1,* Φ_N *is Lipschitz-smooth and strongly convex.*

Proof. For each $i \in [K]$, $g_{x_i}(\boldsymbol{\pi}) = f_{\boldsymbol{\pi}}(x_i)$ is a linear function defined on the convex compact set Π and it's image is the interval $[m, M]$ where $m > 0$. We will prove that $-\log$ is strongly convex on $[m, M]$.

$$\forall x \in [m, M], \frac{1}{M^2} \leq \frac{d^2(-\log)}{dx^2}(x) = \frac{1}{x^2} \leq \frac{1}{m^2}. \tag{4}$$

The first inequality proves the $1/M^2$-strong convexity of $-\log$, the second proves that it is $1/m^2$-Lipschitz smooth. The sum of strongly convex functions is strongly convex. Therefore, Φ_N is strongly convex. □

With these nice property under assumption 1, the minimization problem can be rewritten as follows:

$$\widehat{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi}\in\Pi} \left\{\Phi_N(\boldsymbol{\pi})\right\}, \quad \Pi = \left\{\boldsymbol{\pi} \in [0,1]^K : \boldsymbol{\pi}^T\mathbf{1} = 1, \forall i \in [N], \sum_{j=1}^{K} \pi_j f_j(x_i) \geq m\right\} \tag{5}$$

In this work, we will study different loss function: $||\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*||_1$, $||\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*||_2$ and some $dist(f_{\widehat{\boldsymbol{\pi}}}, f_{\boldsymbol{\pi}^*})$ (donner un example). It turns out that this problem is close to the regression with random design in the context of transductive learning Bellec et al. (2016) since we do not observe the true cluster labels in our problem. We can consider Φ_N as a function of two random variable X_i and $\boldsymbol{\pi}$:

$$\Phi_N(\boldsymbol{\pi}) = \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i, \boldsymbol{\pi}) \tag{6}$$

In this setting $\varphi(.,.)$ (in our problem it is $-log(f_.(.))$) is strongly convex and Lipschitz smooth. We will recall some interesting results for our work on regression with random design.

Let consider the following trace regression model :

$$Y_i = tr(X_i^T\mathbf{B}^*) + \xi_i \quad i = 1, \dots, N \tag{7}$$

with $B^* \in \mathbb{R}^{pxq}$ and let assume that $rank(B^*)$ is small. Let denote $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_p]$ the singular values of B^*. The rank of this matrix is given by $||\sigma||_0$. Unfortunately, the L_0 norm is not convex, we tackle this problem by considering the convex L_1 norm $||\boldsymbol{\sigma}||_1$. Assume the constraint $\boldsymbol{\sigma}^T\mathbf{1} = 1$, then according to Koltchinskii et al. (2016) (quel theoreme ?) an empirical risk minimization method or a Maximum Likelihood Estimator with this constraint leads to a sparse estimator $\widehat{\mathbf{B}}$.

Therefore, it might be interesting to compare this result with our problem 5

References

P. C. Bellec, A. S. Dalalyan, E. Grappin, and Q. Paris. On the prediction loss of the lasso in the partially labeled setting. 2016.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 2016.