MATH 156S HOMEWORK 8

Due May. 28

Exercise 1 (Nonnegative Coding with L_1 regularizer)**.** Suppose we have matrices $\mathbf{X} \in \mathbb{R}^{d\times n}$ and $\mathbf{W} \in \mathbb{R}^{d\times r}$. We seek to find a matrix $\hat{\mathbf{H}} \in \mathbb{R}^{r\times n}_{\geq 0}$ where

$$\hat{\mathbf{H}} = \operatorname*{arg\,min}_{\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \left(\ell(\mathbf{H}) := \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 + \lambda_1\|\mathbf{H}\|_1 + \lambda_2\|\mathbf{H}\|_F^2\right). \tag{1}$$

Here $\lambda \geq 0$ is called the L_1-regularization parameter. (This is an instance of *constrained* quadratic optimization problem.) (c.f. The use of L_1 regularizer instead of L_2 as in Exercise 1.2.1 is more effective to increase the 'sparsity' of the solution, meaning that it will have less nonzero components.)

(i) Show that (use Exercise 1.2.2), denoting $\operatorname{sgn}(\mathbf{H}) = \mathbf{1}(\mathbf{H} \geq 0) - \mathbf{1}(\mathbf{H} < 0)$,

$$\frac{\partial}{\partial \mathbf{H}}\ell = 2(\mathbf{W}^T\mathbf{W}\mathbf{H} - \mathbf{W}^T\mathbf{X}) + \lambda_1 \operatorname{sgn}(\mathbf{H}) + 2\lambda_2\mathbf{H}, \qquad \frac{\partial^2}{\partial \mathbf{H}^2}\ell = 2(\mathbf{W}^T\mathbf{W} + \lambda_2\mathbf{I}). \tag{2}$$

(ii) From **(i)**, conclude that the quadratic function $\mathbf{H} \mapsto \ell(\mathbf{H})$ is convex, and and hence its every critical point is a local minimum. It follows that its every critical over the convex constraint set $\mathbb{R}^{r\times n}_{\geq 0}$ is a global minimum (but not necessarily unique). (See Ref)

Exercise 2. Give a psuedocode of a Block Coordinate Descent algorithm for the following matrix factorization problem

$$\inf_{\mathbf{W}\in\mathbb{R}^{d\times r}_{\geq 0},\, \mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 + a_0\|\mathbf{H}\|_1 + a_1\|\mathbf{W}\|_1 + a_2\|\mathbf{W}\|_F \tag{3}$$

The algorithm should be similar to the Alternating Least Squares (ALS) in Algorithm 2 in LN, but the subproblems of finding $\mathbf{H}_n$ and $\mathbf{W}_n$ should be more detailed by using gradient descent algorithms. (Remark: You may use the gradient given in Exercise 1. (Hint: The full algorithm is implemented in Jupyter Notebook))

Exercise 3. Give an explanation on the effect of L_1- and L_2-regularization on the dictionary matrix $\mathbf{W}$ as we see in Figure 51. Drawing contour plots of L_1- and L_2-norm of 2-dimensional vectors would be helpful. (Ref. [Bis06, Sec. 3.1.4]).

Exercise 4. Use the code in Jupyter Notebook that generates Figures 50 and 51 to determine the *largest* $r \in \mathbb{N}$ *such that every dictionary image properly converges to some non-random image*, where the nonnegativity constraints and regularizers are chosen as in the six cases in Figures 50 and 51 (estimate such r in each of the six cases.) (Remark: Use the revised `display_dictionary` function with the option 'grid_shape'. For example, if you want to plot $\mathbf{W}$ with 13 columns, you can use 'grid_shape=[1,13]').

REFERENCES

[Bis06] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.