Non-negative matrix factorization with smoothness and sparse penalties

Teodor Marinov, Matthew Francis-Landau, Ryan Cotterell

1 Problem formulation

In this project we consider a variant of the non-negative matrix factorization problem (NMF) **?**. The basic NMF problem is posed as follows

$$\begin{aligned} \underset{\mathrm{W}\in\mathbb{R}^{d\times k},\mathrm{H}\in\mathbb{R}^{k\times n}}{\text{minimize:}} \quad & \|\mathrm{X}-\mathrm{WH}\|_F^2 \\ \text{subject to:} \quad & \mathrm{W}_{i,j}\geq 0, \mathrm{H}_{i,j}\geq 0 \end{aligned} \tag{1}$$

where $\mathrm{X}\in\mathbb{R}^{d\times n}$ is some data matrix and k is given and fixed. This is a non-convex optimization problem. In **?** the authors suggest simple alternating multiplicative updates and claim that the proposed algorithm has a fixed point. In **?**, however, it is indicated that the claim is wrong. Another approach to solving problem 1 is the following algorithm – initialize $\mathrm{W}_0, \mathrm{H}_0$ randomly, at step t set W_t to be the minimizer of

$$\begin{aligned} \underset{\mathrm{W}\in\mathbb{R}^{d\times k}}{\text{minimize:}} \quad & \|\mathrm{X}-\mathrm{W}_{t-1}\mathrm{H}_{t-1}\|_F^2 \\ \text{subject to:} \quad & \mathrm{W}_{i,j}\geq 0 \end{aligned} \tag{2}$$

and H_t to be the minimizer of

$$\begin{aligned} \underset{\mathrm{H}\in\mathbb{R}^{k\times n}}{\text{minimize:}} \quad & \|\mathrm{X}-\mathrm{W}_t\mathrm{H}_{t-1}\|_F^2 \\ \text{subject to:} \quad & \mathrm{H}_{i,j}\geq 0 \end{aligned}. \tag{3}$$

Proceed to carry out this alternating minimization approach until some stopping criteria is met e.g. $\|\mathrm{W}_t\mathrm{H}_t-\mathrm{W}_{t+1}\mathrm{H}_{t+1}\|_F^2<\epsilon$. In **?** it is shown that this algorithm is going to have a fixed point. Note that 2,3 are now constraint convex-optimization problems so one can choose their favourite method to solve them.

Usually NMF is applied to real-world problems where the W and H term have some interpretation – for example X can be the Fourier power spectogram of an audio signal where the m,n-th entry is the power of signal at time window n and frequency bin m. The assumption is that the observed signal is coming from a mixture of k static sound sources. Now each column of W can be interpreted as the average power spectrum of an audio source and each row of H can be interpreted as time-varying gain of a source. In practice the number of sources k is not known and we would like to also infer it from the data. This can be done by introducing an additional factor in the optimization problem which indicates the weight of a source in the mixture.

$$\begin{aligned} \underset{\mathrm{W}\in\mathbb{R}^{d\times d},\Theta\in\mathbb{R}^{d\times d},\mathrm{H}\in\mathbb{R}^{d\times n}}{\text{minimize:}} \quad & \|\mathrm{X}-\mathrm{W}\Theta\mathrm{H}\|_F^2+\lambda\|\Theta\|_1 \\ \text{subject to:} \quad & \mathrm{W}_{i,j}\geq 0, \mathrm{H}_{i,j}\geq 0, \Theta_{i,i}\geq 0, \Theta_{i\neq j}=0 \end{aligned} \tag{4}$$

In problem 4 Θ is introduced as the weight matrix for the mixture and an l_1 penalty is introduced to keep the number of "active" sources small. Such a NMF problem has been considered in **?** and a Bayesian approach is taken in solving it by specifying distributions over the elements of W, H and Θ. In our project we directly try to solve a problem similar 4 with an additional penalty term which forces the columns of W to vary smoothly. To conclude the section we present the optimization problem:

$$\begin{aligned} \underset{\mathrm{W}\in\mathbb{R}^{d\times d},\Theta\in\mathbb{R}^{d\times d},\mathrm{H}\in\mathbb{R}^{d\times n}}{\text{minimize:}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}\Theta\mathrm{H}\right\|_F^2+\lambda\left\|\Theta\right\|_1+\eta\sum_{i,j}\left(\mathrm{W}_{i,j}-\mathrm{W}_{i+1,j}\right)^2 \\ \text{subject to:} \quad & \mathrm{W}_{i,j}\geq 0, \mathrm{H}_{i,j}\geq 0, \Theta_{i,i}\geq 0, \Theta_{i\neq j}=0 \end{aligned} \tag{5}$$

2 Algorithm

TODO: write down the gradients/subgradients of 6,7 and 8

Problem 5 is not a convex optimization problem, however, if one considers the 3 separate problems

$$\begin{aligned} \underset{\mathrm{W}\in\mathbb{R}^{d\times d}}{\text{minimize:}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}\Theta\mathrm{H}\right\|_F^2+\eta\sum_{i,j}\left(\mathrm{W}_{i,j}-\mathrm{W}_{i+1,j}\right)^2 \\ \text{subject to:} \quad & \mathrm{W}_{i,j}\geq 0, \mathrm{H}_{i,j}\geq 0 \end{aligned} \tag{6}$$

$$\begin{aligned} \underset{\Theta\in\mathbb{R}^{d\times d}}{\text{minimize:}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}\Theta\mathrm{H}\right\|_F^2+\lambda\left\|\Theta\right\|_1 \\ \text{subject to:} \quad & \Theta_{i,i}\geq 0, \Theta_{i\neq j}=0 \end{aligned} \tag{7}$$

$$\begin{aligned} \underset{\mathrm{H}\in\mathbb{R}^{d\times n}}{\text{minimize:}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}\Theta\mathrm{H}\right\|_F^2 \\ \text{subject to:} \quad & \mathrm{H}_{i,j}\geq 0 \end{aligned} \tag{8}$$

each one is a convex optimization problem. What is more the objectives in 6 and 7 are smooth and each of the objectives is also strongly convex. The proposed algorithm is now to solve each of the convex optimization problems separately in an alternating fashion. Pseudo code is given in 1.

Algorithm 1 Alternating minimization meta algorithm for problem 5

Input: $\mathrm{X}, \mathrm{W}_0, \mathrm{H}_0, \Theta_0, \epsilon$
Output: $\mathrm{W}_T, \mathrm{H}_T, \Theta_T$
while $\left\|\mathrm{W}_{t-1}\mathrm{H}_{t-1}\Theta_{t-1}-\mathrm{W}_t\mathrm{H}_t\Theta_t\right\|_F^2>\epsilon$ **do**

$$\begin{aligned} \mathrm{W}_{t+1} := \underset{\mathrm{W}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}\Theta_t\mathrm{H}_t\right\|_F^2+\eta\sum_{i,j}\left(\mathrm{W}_{i,j}-\mathrm{W}_{i+1,j}\right)^2 \\ \text{subject to} \quad & \mathrm{W}_{i,j}\geq 0, \mathrm{H}_{i,j}\geq 0 \end{aligned}$$

$$\begin{aligned} \mathrm{H}_{t+1} := \underset{\mathrm{H}\in\mathbb{R}^{d\times n}}{\operatorname{argmin}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}_{t+1}\Theta_t\mathrm{H}\right\|_F^2 \\ \text{subject to} \quad & \mathrm{H}_{i,j}\geq 0 \end{aligned}$$

$$\begin{aligned} \Theta_{t+1} := \underset{\Theta\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \quad & \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}_{t+1}\Theta\mathrm{H}_{t+1}\right\|_F^2+\lambda\left\|\Theta\right\|_1 \\ \text{subject to} \quad & \Theta_{i,i}\geq 0, \Theta_{i\neq j}=0 \end{aligned}$$

end while

The main focus of our project is now to solve each of the problems 6,7,8 by using different algorithms explored in class, comparing our empirical observations with the derived convergence results. The algorithms we choose to compare are Projected Gradient/Subgradient Descent, Simple Dual Averaging and Augmented Lagrangian. For Projected Gradient/Subgradient Descent we both experiment with fixed step size and decreasing step size as $\frac{1}{t}$. We are also going to assume that all the minimizers of the above problems are in some compact set – it is not hard to imagine that this holds true, for example consider minimizing the objective in 6. If we let $\left\|\mathrm{W}\right\|_F$ go to infinity for fixed Θ, H and X the objective is going to go to infinity and thus $\|\mathrm{W}\|_F$ must be bounded so we can assume that there exists optimal W^* is in some bounded closed ball with respect to the Frobenius norm. Thus we can restrict our attention on solving the optimization problems on the intersection of closed set with a compact set i.e. a compact set. Thus we can assume the existence of at least one minimizer of each of the optimization problems 6,7 and 8

2.1 Subgradients for problems 6,7,8

If f denotes the respective objective of problems 6,7 and 8 then gradients and an element of the subdifferential of 7 is given by

$$\begin{aligned}\nabla f(\mathrm{W}) &= \frac{2}{n}\left(\mathrm{W}\Theta\mathrm{H}-\mathrm{X}\right)\left(\Theta\mathrm{H}\right)^\top+\eta\tilde{\mathrm{W}} \text{ where}\\ \tilde{\mathrm{W}}_{i,j} &= 2\left(2\mathrm{W}_{i,j}-\mathrm{W}_{i+1,j}-\mathrm{W}_{i-1,j}\right),\\ \tilde{\mathrm{W}}_{1,j} &= 2\left(\mathrm{W}_{1,j}-W_{2,j}\right),\\ \tilde{\mathrm{W}}_{d,j} &= 2\left(\mathrm{W}_{d,j}-\mathrm{W}_{d-1,j}\right)\end{aligned} \tag{9}$$

$$\nabla f(\mathrm{H}) = \frac{2}{n}\left(\mathrm{W}\Theta\right)^\top\left(\mathrm{W}\Theta\mathrm{H}-\mathrm{X}\right) \tag{10}$$

$$\left(\frac{2}{n}\mathrm{W}^\top\left(\mathrm{W}\Theta\mathrm{H}-\mathrm{X}\right)\mathrm{H}+\lambda\mathrm{sgn}\left(\Theta\right)\right)\odot\mathrm{I}\in\partial f(\Theta) \tag{11}$$

where $\odot$ denotes the Hadamard product and “sgn” is the sign function applied element wise to Θ. The derivation in 11 holds because Θ is always constraint to be a diagonal matrix.

3 Projected Gradient Descent

3.1 Fixed step size

TODO: include experiments and comment on comparison with the theory

For this part of the project a modified version of **Algorithm 1** from lecture slides 4 is used with different choices of fixed step size α_k. The difference with the algorithm given in lecture 4 is the stopping criteria – as already discussed in class checking if the norm of the gradient is close to 0 will not work well for objectives including l_1 penalty term, instead we choose to stop our procedure either after a fixed number of steps (in our experiments this is 200 when solving problems 6 and 7 and 500 when solving problem 8) or if the distance between consecutive iterates becomes less than ϵ (where ϵ was set to be in the range $[10^{-4}, 10^{-5}]$). As discussed in class this is usually not a good stopping criteria unless the objective is differentiable with L-Lipschitz continuous derivatives. Luckily both the objectives in 6 and 8 are differentiable with Lipschitz continuous gradients which we show now.

Lemma 3.1. *The objective in problem 6 is differentiable with L-Lipschitz continuous gradients.*

Proof. Denote the objective in problem 6 by $f(\mathrm{W})$. Then $\nabla f(\mathrm{W}) = \frac{2}{n}\left(\mathrm{W}\Theta\mathrm{H}-\mathrm{X}\right)\left(\Theta\mathrm{H}\right)^\top+\eta\tilde{\mathrm{W}}$ where $\tilde{\mathrm{W}}_{i,j} = 2\left(2\mathrm{W}_{i,j}-\mathrm{W}_{i+1,j}-\mathrm{W}_{i-1,j}\right), \tilde{\mathrm{W}}_{1,j} = 2\left(\mathrm{W}_{1,j}-W_{2,j}\right), \tilde{\mathrm{W}}_{d,j} = 2\left(\mathrm{W}_{d,j}-\mathrm{W}_{d-1,j}\right)$. With this we have

$$\left\|\nabla f\left(\mathrm{W}_1-\mathrm{W}_2\right)\right\|_F = \left\|\frac{2}{n}\left(\left(\mathrm{W}_1-\mathrm{W}_2\right)\Theta\mathrm{H}\right)\left(\Theta\mathrm{H}\right)^\top+\eta\left(\tilde{\mathrm{W}}_1-\tilde{\mathrm{W}}_2\right)\right\|_F \leq \left(\frac{2}{n}\left\|\Theta\mathrm{H}\right\|_F^2+12\eta\right)\left\|\mathrm{W}_1-\mathrm{W}_2\right\|_F \tag{12}$$

where we used triangle inequality and bounded each of the $\|(\mathrm{W}_1)_{i,1:j}-(\mathrm{W}_2)_{i,1:j}\|_F \leq \|\mathrm{W}_1-\mathrm{W}_2\|_F$. □

The above lemma shows that the Lipschitz constant for the objective can indeed be very large as it depends on the product $\Theta\mathrm{H}$, however, in practice setting fixed step size $\alpha \leq 0.05$ seems to be in the range $(0, \frac{2}{L})$ which is when convergence for the algorithm is guaranteed. Sadly we can not guarantee strong convexity or strict convexity for the objectives in 6 and 8 so the theorem which charaterizes the best convergence rate