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) [3]. 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 [3] the authors suggest simple alternating multiplicative updates and claim that the proposed algorithm has a fixed point. In [2], 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 [4] 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 [1] 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**
    $\mathrm{W}_{t+1}:=\underset{\mathrm{W}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \quad \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}\Theta_t\mathrm{H}_t\right\|_F^2+\eta\sum_{i,j}\left(\mathrm{W}_{i,j}-\mathrm{W}_{i+1,j}\right)^2$
    subject to $\quad \mathrm{W}_{i,j}\geq 0,\mathrm{H}_{i,j}\geq 0$
    $\mathrm{H}_{t+1}:=\underset{\mathrm{H}\in\mathbb{R}^{d\times n}}{\operatorname{argmin}} \quad \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}_{t+1}\Theta_t\mathrm{H}\right\|_F^2$
    subject to $\quad \mathrm{H}_{i,j}\geq 0$
    $\Theta_{t+1}:=\underset{\Theta\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \quad \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}_{t+1}\Theta\mathrm{H}_{t+1}\right\|_F^2+\lambda\left\|\Theta\right\|_1$
    subject to $\quad \Theta_{i,i}\geq 0,\Theta_{i\neq j}=0$
  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 $\frac{1}{\sqrt{t}}$.

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

$$\left\|\nabla f\left(\mathrm{W}_1 - \mathrm{W}_2\right)\right\|_F = \left\|\frac{2}{n}\left(\left(\mathrm{W}_1 - \mathrm{W}_2\right)\Theta\mathrm{H}\right)(\Theta\mathrm{H})^\top + \eta\left(\tilde{\mathrm{W}}_1 - \tilde{\mathrm{W}}_2\right)\right\|_F \leq \left(\frac{2}{n}\left\|\Theta\mathrm{H}\right\|_F^2 + 12\eta\right)\left\|\mathrm{W}_1 - \mathrm{W}_2\right\|_F \tag{9}$$

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

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. If we now assume that neither of Θ or H are identically zero we have the next lemma.

Lemma 3.2. *The objective in problem 6 is strongly convex*

Proof. We use the following equivalent definition to the one given in lecture slides 2 of strong convexity: $f(x)$ is strongly convex if there exists $\alpha > 0$ s.t. $f(x) - \gamma\left\|x\right\|^2$ is convex. First note that $\eta\sum_{i,j}\left(\mathrm{W}_{i,j} - \mathrm{W}_{i+1,j}\right)^2$ is a convex function of W. We now show that there exists an $\gamma > 0$ s.t. $\frac{1}{n}\left\|\mathrm{X} - \mathrm{W}\Theta\mathrm{H}\right\|_F^2 - \gamma\left\|\mathrm{W}\right\|_F^2$ is a convex function. This is equivalent to $\left(\frac{1}{n}\left\|\Theta\mathrm{H}\right\|_F^2 - \gamma\right)\left\|\mathrm{W}\right\|_F^2 - 2\langle\mathrm{X}, \mathrm{W}\Theta\mathrm{H}\rangle$ is a convex function for some γ. It is easy to see that any $\gamma < \frac{1}{n}\left\|\Theta\mathrm{H}\right\|_F^2$ is going to work by just plugging in the definition of convex function and expanding the Frobenius norm squared term. $f(x) - \gamma\left\|x\right\|^2$ is now a convex function as the sum of two convex functions. □

Combining the above two lemmas with theorem 1.11 in lecture slides 6 tells us we should have a linear rate of convergence with constant $\frac{L/\gamma - 1}{L/\gamma + 1}$. From the lemmas we can have a rough estimate of $\gamma \sim \frac{1}{2n}\left\|\Theta\mathrm{H}\right\|_F^2$ and $L \sim \frac{3}{n}\left\|\Theta\mathrm{H}\right\|_F^2$ (assuming that the $\left\|\Theta\mathrm{H}\right\|_F^2$ is the dominant term for the Lipschitz constant). This rough estimates imply that $\frac{L/\gamma - 1}{L/\gamma + 1} \sim \frac{1}{5}$ (modulo my calculation errors) so we are guaranteed very quick convergence provided we manage to "choose" or estimate the optimal step size. As it is going to become apparent from our experiments we do a very poor job in the selection of our step size.

Similar lemmas and rate of convergence can be shown for the objective in 8. We still need to address the convergence for the projected sub-gradient descent procedure for problem 7. In the same way as in lemma 3.2 one can show that the objective in 7 is strongly convex with parameter $\gamma < \frac{1}{n}\left\|\mathrm{WH}\right\|_F^2$ (TODO: check that this is actually true). Theorem 1.8 in lecture slides 4 now characterizes the linear convergence rate to a local neighbourhood of the solution. To address our choice of stopping criteria, from Theorem 1.8 we know that $d(\Theta_{k+1}, \Theta^*)^2 < \alpha\frac{\kappa_g^2}{\gamma} + c^k d(\Theta_0, \Theta^*)^2$, where Θ^* is the optimal solution to 7, $c < 1$ depends

on α and γ and κ_g is a bound on the norm of the elements in the sub-differential of the objective. For k large enough this implies that all of the Θ_k's are going to be contained in an open ball of fixed radius which is approximately $c^k d(\Theta_0, \Theta^*)^2$ – this implies that the distance between any two consecutive iterates $d(\Theta_k, \Theta_{k+1})^2$ is also going to be less than $\alpha\frac{\kappa_g^2}{\gamma}$. Since the convergence theory does not guarantee anything more stopping our algorithm when $d(\Theta_k, \Theta_{k+1})^2$ becomes small enough seems acceptable. To be absolutely fair $d(\Theta_k, \Theta_{k+1})^2$ being small is only a necessary condition for convergence but not sufficient – it might happen that two consecutive iterates are close to each other, however, they are still not close to the optimal Θ^*. To alleviate this problem one might check that all the pair-wise distances between $\Theta_k, \Theta_{k+1}, ..., \Theta_{k+\tau}$ are small.

3.2 Decreasing step size

TODO:The theory for fixed step size does not hold anymore thus we can only use the sub-linear convergence results – write these up and for stopping criteria for 7 comment on using lemma 1.6 in lecture 4 together with convergence in objective and the fact convergent sequences are Cauchy. Include experiments and comment in same way as in previous section

4 Simple Dual Averaging

5 Augmented Lagrangian

References

[1] David M Blei, Perry R Cook, and Matthew Hoffman. Bayesian nonparametric matrix factorization for recorded music. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 439–446, 2010.

[2] Edward F Gonzalez and Yin Zhang. Accelerating the lee-seung algorithm for non-negative matrix factorization. Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02, 2005.

[3] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pages 556–562, 2001.

[4] JOEL A Tropp. An alternating minimization algorithm for non-negative matrix approximation, 2003.