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 favorite 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

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 8 are differentiable and 7 is strongly convex assuming that $\mathrm{H}^\top\mathrm{W}$ is full rank. The proposed algorithm is now to solve each of the convex optimization problems separately in an alternating fashion. Pseudo code is given in algorithm 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**

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

  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 $\left\|\mathrm{W}\right\|_F$ 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. A similar argument holds for Θ and H.

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

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

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

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

where $\odot$ denotes the Hadamard product and "sign" 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 $\epsilon \in [10^{-4}, 10^{-5}]$). 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 $\left\|\left(\mathrm{W}_1\right)_{i,1:j}-\left(\mathrm{W}_2\right)_{i,1:j}\right\|_F \leq \left\|\mathrm{W}_1-\mathrm{W}_2\right\|_F$. □

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. Sadly we can not guarantee strong convexity or strict convexity for the objectives in 6 and 8 so the theorem which characterizes the best convergence

rate is Theorem 1.9 in lecture slides 6. From our experiments we observe that our initial points W_0 and H_0 are roughly in the order of 10^3 and 10^5 from what we consider an optimal point and with $\alpha \sim 0.005$ we should have convergence roughly as $|f(\mathrm{W}_k) - f^*| \leq \frac{10^3}{0.005*k}$ and $|f(\mathrm{H}_k) - f^*| \leq \frac{10^5}{0.005*k}$. Here f denotes the respective objective function and f^* denotes the optimum objective value.

Surprisingly we can get linear convergence for 7 under mild assumptions that the matrix $\mathrm{H}^\top\mathrm{W}$ is full rank. Such a rate will follow from showing the next lemma.

Lemma 3.2. *Assume* $\mathrm{H}^\top\mathrm{W}$ *is full rank. Then the objective in 7 is strongly convex with strong convexity parameter* $\gamma < \frac{1}{n}\sigma_{min}(\mathrm{H})^2\sigma_{min}(\mathrm{W})^2$.

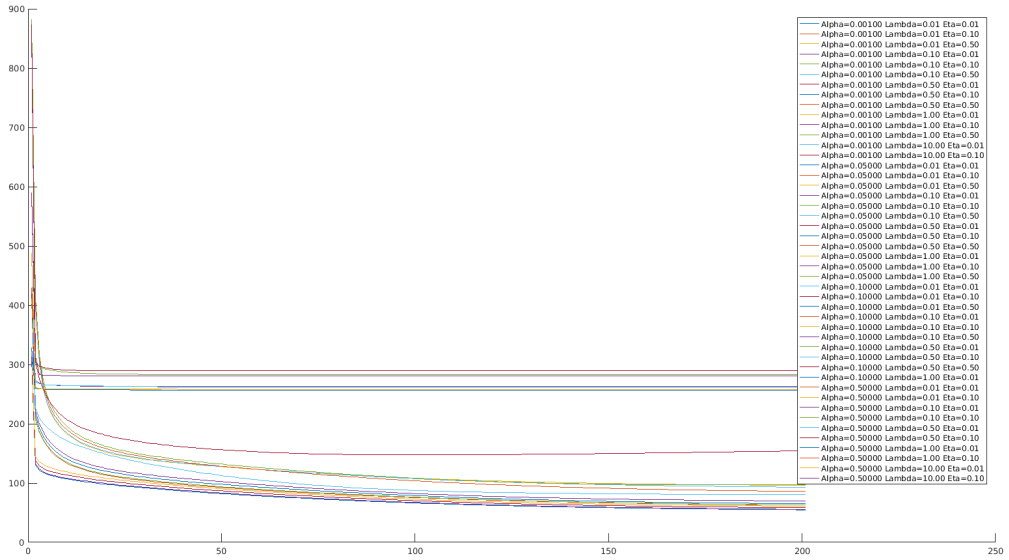
Proof. To show the objective in 7 is strongly convex we are going to show equivalently that $\frac{1}{n}\left\|\mathrm{W}\Theta\mathrm{H}\right\|_F^2$ is strongly convex under the given assumption. To do this we are going to use a second order condition for strong convexity i.e. the fact that the Hessian of the above function should be a positive definite form. Since the Hessian of $\frac{1}{n}\left\|\mathrm{W}\Theta\mathrm{H}\right\|_F^2$ is an order 4 tensor and we would not like to compute it we are going to use a little trick and vectorize $\mathrm{W}\Theta\mathrm{H}$. Let $vec(\mathrm{A})$ denote the vectorization of a matrix A by stacking its columns on top of each other. We use a famous equality $vec\left(\mathrm{W}\Theta\mathrm{H}\right) = \left(\mathrm{H}^\top \otimes \mathrm{W}\right) vec(\Theta)$ where $\otimes$ denotes the Kronecker product. If we denote $\mathrm{A} = \mathrm{H}^\top \otimes \mathrm{W}$ and $\mathrm{x} = vec(\Theta)$ then $\frac{1}{n}\left\|\mathrm{W}\Theta\mathrm{H}\right\|_F^2 = \frac{1}{n}\left\|\mathrm{Ax}\right\|_2^2$. The Hessian of $\frac{1}{n}\left\|\mathrm{Ax}\right\|_2^2$ equals $\frac{2}{n}\mathrm{A}^\top\mathrm{A}$. Now the strong convexity parameter of $\frac{1}{n}\left\|\mathrm{Ax}\right\|_2^2$ is characterized by the smallest singular value of the of $\mathrm{A}^\top\mathrm{A}$ which equals the smallest singular value squared of $\mathrm{A} = \mathrm{H}^\top \otimes \mathrm{W}$. From theorem 13.12 in **?** we know that the smallest singular value of $\mathrm{H}^\top \otimes \mathrm{W}$ is given by $\sigma_{min}(\mathrm{H})\sigma_{min}(\mathrm{W})$ which concludes the proof. □

Theorem 1.8 in lecture slides 4 now characterizes the linear convergence rate to a local neighborhood 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.

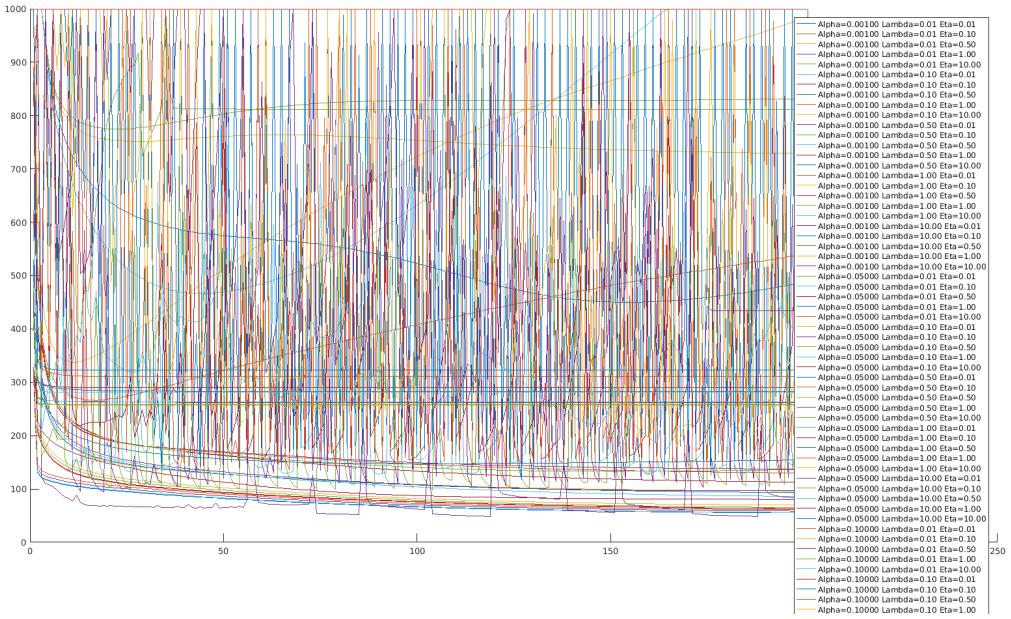
3.1.1 Fixed step size experiment results

All of our experiments are ran on a matrix $\mathrm{X} \in \mathbb{R}^{40\times 5000}$ constructed by sampling $\mathrm{W} \in \mathbb{R}^{40\times 40}, \Theta \in \mathbb{R}^{40\times 40}$ and $\mathrm{H} \in \mathbb{R}^{40\times 5000}$ from standard Gaussian and then thresholding so each of the matrices have non-negative entries. We also only keep the diagonal entries of Θ and finally set $\mathrm{X} := \mathrm{W}\Theta\mathrm{H}$. The number of non-zero entries in Θ is 18. We note that this is not the best construction of X we could have come up since neither Θ is very sparse nore is there any smoothness constraints on the columns of W. Thus as it is about to be shown the optimal solutions the the variants parametrized by η and λ of problem 5 we solve are not going to be close (in the sense of Frobenius norm) to the factors from which X was constructed. In fact for various settings of λ and η we achieve much better performance on the overall objective of 5 than the original W, Θ and H factors from which X was constructed. Our initializations for each of the tests is done in the same way the original factors of X were constructed. While this is by no means a good initialization we still see roughly the same performance of all of the methods we use for solving the sub-problems in 5 so initialization does not seem to play a major role in the final performance. We have performed several different experiments specifically for the projected gradient descent method with both increasing and decreasing step size. We begin by running each of the projected gradient descent algorithms on the separate optimization problems for 200 iterations when solving 6 and 7 and 500 iterations when solving 8. Each of these are ran a total of 250 times in the outer most loop of the alternating minimization algorithm. The parameters which yielded

the smallest overall objective were $\lambda = 0.01$, $\eta = 0.01$ and $\alpha = 0.001$ was 54.1. This is no surprise as because X was not constructed with factors taking into account the smoothness and sparsity penalties when these penalties are enforced the least we should get the smallest loss. We chose to run the internal minimization projected gradient descent algorithm for only 200/500 iterations as the number of different values of λ, η and initial step size α we try together with the 200 iterations of the outer loop take quite a bit of time. The overall objective results can be seen in figure 1. The number of non-zero entrie sof the obtained Θ was 21.



(a) Projected gradient on experiments with well behaved results



(b) Projected gradient on all of our 64 hyperparameter experiments. (Or as we like to call it, modern art)

Figure 1: Experimental results from fixed step sized for projected gradient decent running $\alpha \in \{0.001, 0.05, 0.1, 0.5\}, \eta \in \{0.01, 0.1, 0.5, 1, 10\}, \lambda \in \{0.01, 0.1, 0.5, 1, 10\}$. For our experiments with larger step sizes, we observe that the obective values are not stable when plotting on the range of < 1000. However even looking on a large scale there still tend to be spikes with some of the methods when using larger η, λ as well. As with our experiments using other methods, we observe that we are able to find the lowest value of our objective when setting η and λ to small (0.01) values. This is expected since these behave as reguralarization on the matrices which we are able to find. Additionally, we note again that for $\lambda = \eta = 0$ we kown that there exists an exact solution to this problem.