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**

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

$$\text{subject to} \quad \mathrm{W}_{i,j}\geq 0,\mathrm{H}_{i,j}\geq 0$$

$$\mathrm{H}_{t+1}:=\underset{\mathrm{H}\in\mathbb{R}^{d\times n}}{\operatorname{argmin}} \quad \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}_{t+1}\Theta_t\mathrm{H}\right\|_F^2$$

$$\text{subject to} \quad \mathrm{H}_{i,j}\geq 0$$

$$\Theta_{t+1}:=\underset{\Theta\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \quad \frac{1}{n}\left\|\mathrm{X}-\mathrm{W}_{t+1}\Theta\mathrm{H}_{t+1}\right\|_F^2+\lambda\left\|\Theta\right\|_1$$

$$\text{subject to} \quad \Theta_{i,i}\geq 0,\Theta_{i\neq j}=0$$

  end 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 $\|\mathrm{W}\|_F$ 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 $\left\|\left(\mathrm{W}_1\right)_{i,1:j}-\left(\mathrm{W}_2\right)_{i,1:j}\right\|_F \leq \left\|\mathrm{W}_1-\mathrm{W}_2\right\|_F$. □

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 charaterizes 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 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.

3.1.1 Fixed step size experiment results

TODO: clean up

3.2 Decreasing step size

None of the convergence results in lecture slides 6 hold any longer for projected subgradient descent with decreasing step size. However, we can still characterize the convergence in terms of objective and iterates for step size decreasing as $\frac{1}{t}$. From Theorem 1.11 in lecture slides 4 we know that the iterates for projected gradient descent for problems 6,7 and 8 will converge to an optimal point (given such exists). For 6 and 8 we can use Lemma 1.3 in lecture slides 6 to show that this would imply convergence in objective. From Lemma 1.3 in lecture slides 6 we know that if f has an L-Lipschitz continuous gradient we have $f(x_k) - f(x^*) \leq \langle\nabla f(x^*)(x_k - x^*)\rangle + \frac{L}{2}\left\|x_k - x^*\right\|^2$. Using Cauchy-Schwartz inequality we have $\langle\nabla f(x^*)(x_k - x^*)\rangle \leq \left\|\nabla f(x^*)\right\|\left\|x_k - x^*\right\|$ thus $f(x_k) - f(x^*) \leq \left\|x_k - x^*\right\|\left(\left\|\nabla f(x^*)\right\| + \frac{L}{2}\left\|x_k - x^*\right\|\right)$ and since $\left\|\nabla f(x^*)\right\|$ is bounded this shows convergence in objective. By triangle inequality we have $\left\|x_k - x^*\right\| \leq \left\|x_k - x_{k+1}\right\| + \left\|x_{k+1} - x^*\right\|$ and thus $f(x_k) - f(x^*) \leq \left\|x_k - x_{k+1}\right\|\left(\left\|\nabla f(x^*)\right\| + \frac{L}{2}\left\|x_k - x^*\right\|\right) + \left\|x_{k+1} - x^*\right\|\left(\left\|\nabla f(x^*)\right\| + \frac{L}{2}\left\|x_k - x^*\right\|\right) \leq$

$\|x_k - x_{k+1}\| c_1 + \tilde{\epsilon}$ where $c_1 = \|\nabla f(x^*)\| + \frac{L}{2}\|x_k - x^*\|$ and $\tilde{\epsilon} = \|x_{k+1} - x^*\|\left(\|\nabla f(x^*)\| + \frac{L}{2}\|x_k - x^*\|\right)$. Clearly for k large enough $\tilde{\epsilon}$ is as small as we would like and c_1 is bounded thus for k large enough $\|x_{k+1} - x_k\| < \epsilon$ implies $f(x_k) - f(x^*)$ is small. This should somewhat justify our stopping criteria for 6 and 8. None of the above derivations, however, hold for 7 in fact the only thing we can argue is that the iterates Θ_k are going to converge to Θ^*. This is quite disappointing compared to the results we were able to obtain for a constant step size. In practice as we can see from the experiments we still get satisfactory results. TODO:Include experiments and comment in same way as in previous section

3.2.1 Decreasing step size experiment results

4 Simple Dual Averaging

In this section we compare the SDA given in lecture slides 4 as **Algorithm 3**. We also address details in the implementation, convergence theory and stopping criteria used.

4.1 Algorithm and implementation

Lower case bold letters denote matrices (contrary to standard convention) and the norm is the Frobenius norm together with the associated standrard inner product for matrices. We follow the pseudo-code given in **Algorithm 3** in the lecture slides as already stated. The most interesting part of the algorithm is implementing the update $\mathrm{x}_{k+1} \leftarrow \arg\min_{\mathrm{x}\in\mathcal{X}} \langle \mathrm{x}, \mathrm{s}_{k+1}\rangle + \frac{\beta_{k+1}}{2}\|\mathrm{x} - \mathrm{x}_0\|^2$ for problems 6,7 and 8. The following lemma shows us that this is equivalent to the projection of $x_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1}$ onto the convex set $\mathcal{X}$.

Lemma 4.1. *Solving* $\arg\min_{\mathrm{x}\in\mathcal{X}} \langle \mathrm{x}, \mathrm{s}_{k+1}\rangle + \frac{\beta_{k+1}}{2}\|\mathrm{x} - \mathrm{x}_0\|^2$ *is equivalent to solving* $\arg\min_{\mathrm{x}\in\mathcal{X}} \left\|\mathrm{x} - (\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})\right\|^2$

Proof.

$$
\begin{aligned}
\arg\min_{\mathrm{x}\in\mathcal{X}} \langle \mathrm{x}, \mathrm{s}_{k+1}\rangle + \frac{\beta_{k+1}}{2}\|\mathrm{x} - \mathrm{x}_0\|^2 = \\
\arg\min_{\mathrm{x}\in\mathcal{X}} \langle \mathrm{x}, \mathrm{s}_{k+1} - \beta_{k+1}\mathrm{x}_0\rangle + \frac{\beta_{k+1}}{2}\|\mathrm{x}\|^2 = \\
\arg\min_{\mathrm{x}\in\mathcal{X}} \langle \mathrm{x}, \mathrm{s}_{k+1} - \beta_{k+1}\mathrm{x}_0\rangle + \frac{\beta_{k+1}}{2}\|\mathrm{x}\|^2 + \frac{1}{2\beta_{k+1}}\|\mathrm{s}_{k+1} - \beta_{k+1}\mathrm{x}_0\|^2 = \\
\arg\min_{\mathrm{x}\in\mathcal{X}} \frac{\beta_{k+1}}{2}\left\|\mathrm{x} - (\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})\right\|^2 = \\
\arg\min_{\mathrm{x}\in\mathcal{X}} \left\|\mathrm{x} - (\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})\right\|^2
\end{aligned}
\tag{13}
$$

□

For the set $\mathcal{X} := \{\mathrm{x}_{i,j} \geq 0\}$ it is easy to verify that the solution to $\arg\min_{\mathrm{x}\in\mathcal{X}} \left\|\mathrm{x} - (\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})\right\|^2$ is exactly given by the operator $\mathcal{P}(\mathrm{x}) = \tilde{\mathrm{x}}$ where

$$
\mathcal{P}(\mathrm{x})_{i,j} = \begin{cases} \mathrm{x}_{i,j} & \mathrm{x}_{i,j} \geq 0 \\ 0 & \text{otherwise} \end{cases}
$$

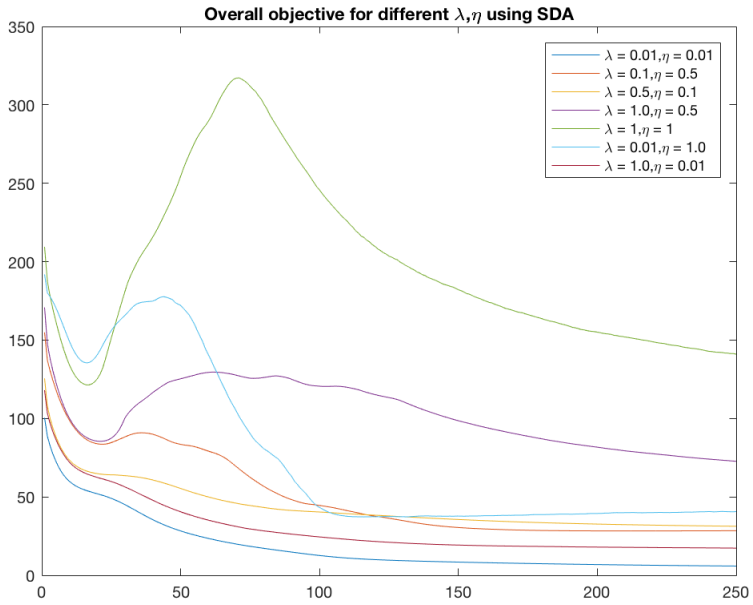
. A simple way to see this is to form the Lagrangian $\mathcal{L}(\mathrm{x}, \mathrm{y})$ for $\min_{\mathrm{x}\in\mathcal{X}} \frac{1}{2}\left\|\mathrm{x} - (\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})\right\|^2$ and notice that the pair $(\mathcal{P}(\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1}), \mathrm{y}^*)$ satisfies KKT conditions. Here y^* is given by

$$
\mathrm{y}^*_{i,j} = \begin{cases} 0 & (\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})_{i,j} \geq 0 \\ -(\mathrm{x}_0 - \frac{1}{\beta_{k+1}}\mathrm{s}_{k+1})_{i,j} & \text{otherwise} \end{cases}
$$

All other steps of the algorithm are as given in the lecture slides and the β_k's are chosen according to Theorem 1.15 in lecture slides 4.

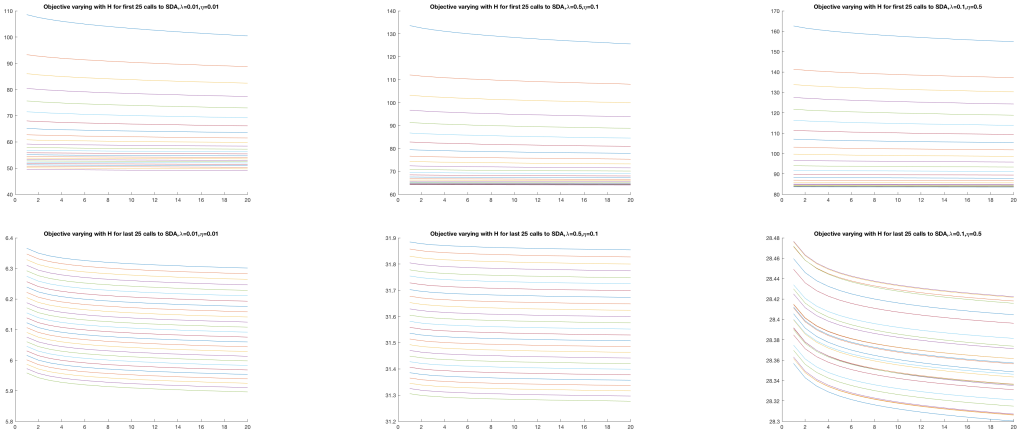
4.1.1 SDA experiment results

Even though we are given a reliable stopping criteria for this algorithm we are not able to implement it as it requires computation of the conjugate function of the objective which we do not know how to do except for problem 8. One can argue that an iterative method for approximately computing the conjugate functions and therefore the stopping criteria might give sufficient results, however, the overall run-time to obtain a solution to optimization problem 5 is long enough without having to solve an additional optimization problem at each step of the sub-problem optimization routines. In the light of these concerns we choose to stop each SDA routine either after a fixed number of iterations or when consecutive iterates become ϵ close to each other. From Theorem 1.17 in lecture slides 4 we know that to get ϵ-suboptimality for the optimization problem we would need the order of $\frac{1}{\epsilon^2}$ iterations, where the constant depends on an upper bound on the gradients and the radius of the ball centered at an optimal solution containing the initialization. From our experiments the norm of the gradients are bounded by roughly ~ 100 and the norm of the difference between initializations and the best solution we manage to recover is roughly ~ 600. The number of iterations for which we run our algorithms are 200, 1000 and 5000 with additional 500 iterations for optimization problem 8. The initializations are the same as for the experiments in projected gradient descent. As before we choose to run our λ and η experiments for only 200 iterations per alternating minimization step for problems 6 and 7 and for 500 iterations for problem 8 in the interest of time. As before we run the alternating minimization algorithm 1 for 200 iterations.



We see that the smallest objective is obtained at $\eta = 0.01, \lambda = 0.01$ – this is no surprise as already discussed since the original matrix X was not obtained as a product of matrices which are sparse or smooth. We see that the smoothness constraint is what seems to hurt the overall objective the most and might be the reason why we see a peak of the objective around the 100 iteration of the alternating minimization. We also see that the sparsity constraint does not hurt the objective too much. To compare the rank of Θ we recover for

$\lambda = 1, \eta = 0.01$ is 32 while when $\lambda = 0.01, \eta = 0.01$ it is 37. To remind the reader the rank of X is 18. The next figures show how the objective decreases at each alternating minimization step of 200/500 iterations for different values of λ, η.



5 Augmented Lagrangian

TODO