Exercise 1) $X \in \mathbb{R}^{d\times n}$ and $W \in \mathbb{R}^{d\times r}$

Find $\hat{H} \in \mathbb{R}^{n\times n}_{\geq 0}$ where

$$\hat{H} = \underset{H \in \mathbb{R}^{r\times n}_{\geq 0}}{\operatorname{argmin}} \left(\ell(H) := \|X - WH\|_F^2 + \lambda_1 \|H\|_1 + \lambda_2 \|H\|_F^2 \right)$$

i) Let $\operatorname{sgn}(H) = 1(H \geq 0) - 1(H < 0)$

$$L = \sum_{i,j} \Big(X_{ij} - \sum_k W_{ik} H_{kj}\Big)^2 + \lambda_1 \sum_{i,j} |H_{ij}| + \lambda_2 \sum_{i,j} (H_{ij})^2$$

$$\frac{\partial L}{\partial H_{ab}} = \sum_{ij} 2\Big(X_{ij} - \sum_k W_{ik} H_{kj}\Big) \frac{\partial H_{kj}}{\partial_{ab}} (-W_{ik}) + \lambda_1 \frac{\partial |H_{ab}|}{\partial |H_{ab}|} + 2\lambda_2 H_{ab}$$

$$= \sum_{ij} 2\Big(X_{ij} - \sum_k W_{ik} H_{kj}\Big) \partial_{ka}\partial_{jb} (-W_{ik}) + \lambda_1 \frac{\partial |H_{ab}|}{\partial |H_{ab}|} + 2\lambda_2 H_{ab}$$

$$= 2\big(X_{ib} - W_{ia} H_{ab}\big)(-W_{ia}) + \lambda_1 \operatorname{sgn}(H_{ab}) + 2\lambda_2 H_{ab}$$

$$= 2\big(-W_{ai} X_{ib} + W_{ai} W_{ia} H_{ab}\big) + \lambda_1 \operatorname{sgn}(H_{ab}) + 2\lambda_2 H_{ab}$$

$$= 2\big(-W^T X + W^T W H\big)_{ab} + \lambda_1 \operatorname{sgn}(H_{ab}) + 2\lambda_2 H_{ab}$$

$$= \boxed{2(W^T W H - W^T X) + \lambda_1 \operatorname{sgn}(H) + 2\lambda_2 H}$$

$2x + y + 2$

$$\frac{\partial L}{\partial H_{ab} H_{cd}} = \frac{\partial H_{ab}}{\partial cd} (2W^T W H - W^T X)_{ab} + \frac{\partial H_{ab}}{\partial cd} (2\lambda_2 H)_{ab}$$

$$= \partial_{ac}\partial_{bd} (2W^T W H - W^T X)_{ab} + \partial_{ac}\partial_{bd} (2\lambda_2 H)_{ab}$$

$$= (2W^T W)_{cd} + 2\lambda_2 I_{cd}$$

$$= 2(W^T W + \lambda_2 I)_{cd}$$

$$= \boxed{2(W^T W + \lambda_2 I)}$$

ii) If we can show that our Hessian(L) is a positive semidefinite matrix, than we know $\ell(H)$ is a convex function and thus every critical over the convex constraint $\mathbb{R}^{r\times n}_{\geq 0}$ is a global minimum.

$\lambda_2 \geq 0$ is given, now $W^T W$ is a PSD:

where $x^T W^T W x = (Wx)^T W x = \|Wx\|^2 \geq 0 \quad \forall\, x \in \mathbb{R}^{r\times n}$

Exercise 2

Input $X \in \mathbb{R}^{d \times n}_{\geq 0}$, $r \geq 1$, $[W_0, H_0] \in \mathbb{R}^{d \times r}_{\geq 0} \times \mathbb{R}^{r \times n}_{\geq 0}$ (initial estimate), N number of iterations

for $n = 1, \dots, N$ do

optimizing for H_n using previous W_{n-1}

while $i <$ sub_iterations

for $k = 1, \dots, r$

grave of individual coordinats

$$\nabla_{\ell(H)} = W^T W[k,:](H[:,idx] - W^T X[:,idx] + \lambda_1 I) + 2\lambda_2 \operatorname{sgn}(H)$$

$H = H - (\text{step\_size}) \nabla_{\ell(H)}$

optimizing for W_n using H_n from before

while $i <$ sub_iterations

for $k = 1, \dots, r$

grave of individual coordinats

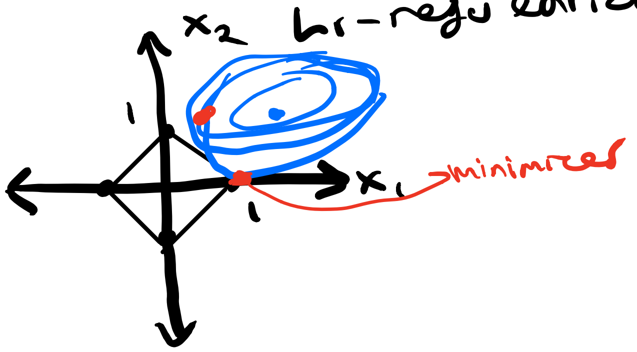
$$\nabla_{\ell(W)} = H^T H[k,:](W[:,idx] - H^T X[:,idx] + \lambda_1 I) + 2\lambda_2 \operatorname{sgn}(W)$$

$W = - - (\text{step\_size}) \nabla_{\ell(W)}$

Return H, W.

Excercise 3

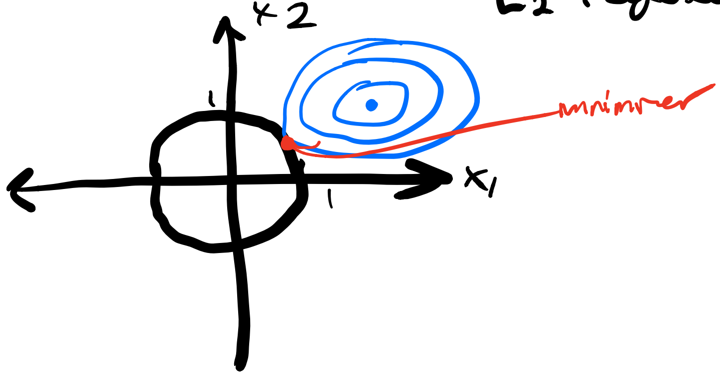
L_1-regularization $\|X\|_1$



$\begin{pmatrix} x_1 \\ 0 \end{pmatrix}$

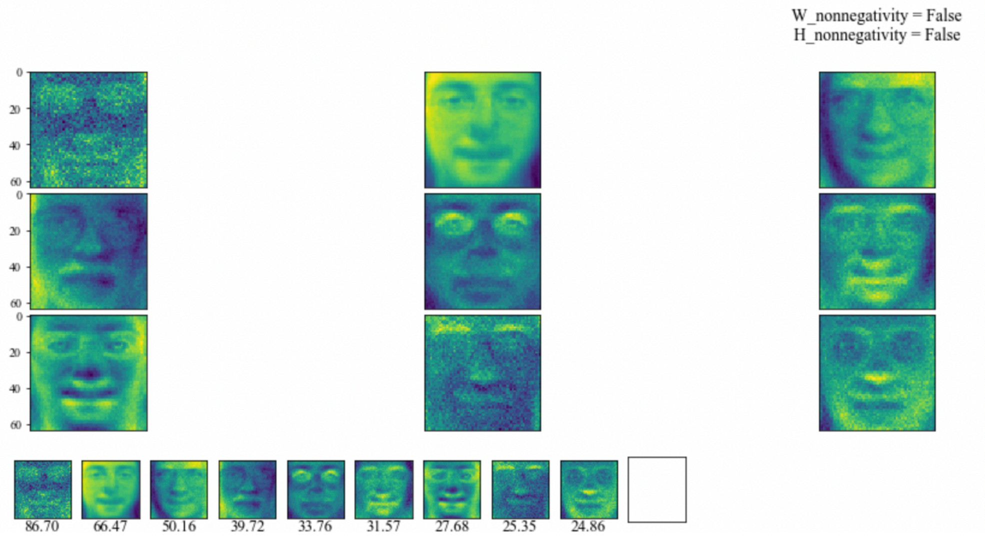
We can see that the regression contour for L_1 looks more like a diamond shape. We can see that L_1 tends to zero out one of our features. We use this when we know we have certain features that are irrelevant in our data set, as it causes that feature's coefficient to be zeroed out, reducing the complexity. This can lead to our model having sparse coefficients.

L_2-regularization

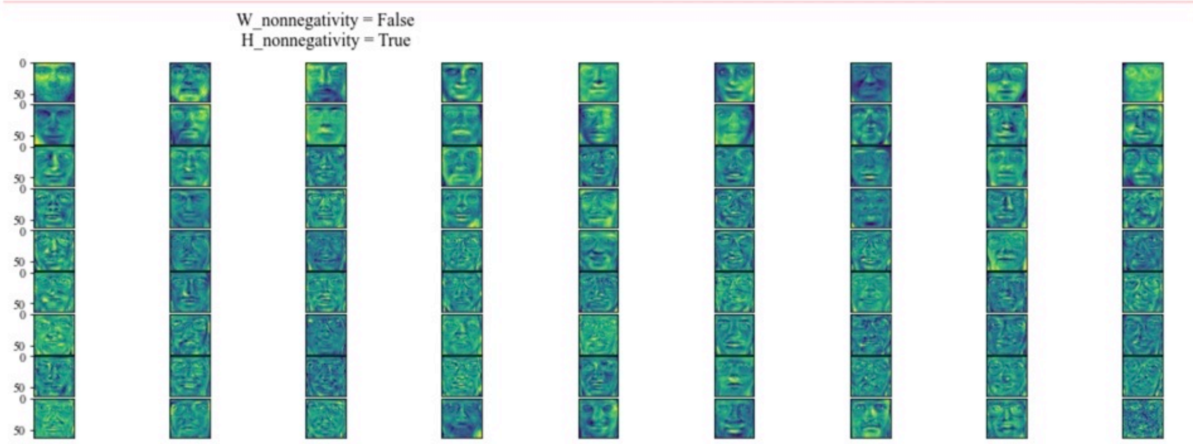


Above we have the contour plot for L_2-regularization, we can see that it looks more like a circle compared the diamond shape in L_1. Here, regularization does not necessarily zero out our feature, but instead we penalize all coefficients to keep any 1 feature from becoming to dominant. We can see in the contour plot that it is not as plausible for our L_2-regulizer to zero out coefficients.

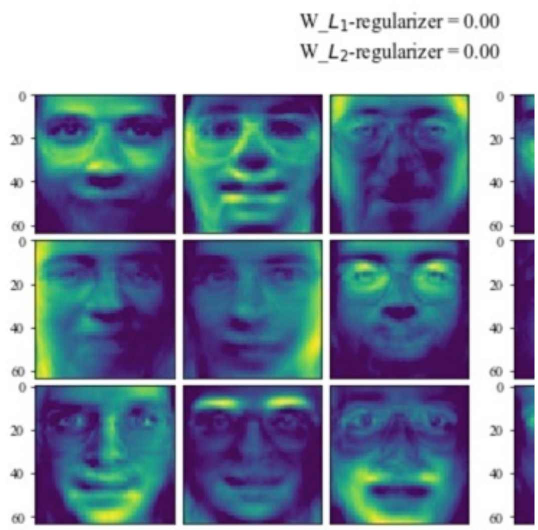
Excercise 4



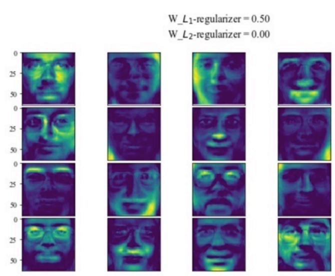
r=9

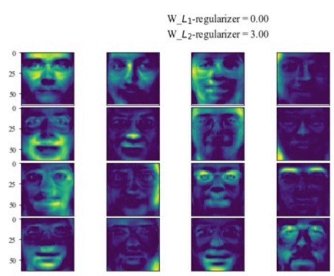


r=81



r=9





r=16

r=16