Problem 1.a.

Solution: We compute the gradient wrt W elementwise, that is we compute $\frac{\partial E}{\partial w_{jk}}$. First lets compute the gradient for the first term.

$$\frac{\partial}{\partial w_{jk}}\eta||W||_F^2 = \eta\frac{\partial}{\partial w_{jk}}\sum(w_{mn})^2 = 2\eta w_{jk}\,.$$

And for the other terms, where we can compute the derivative for each term in the sum separately. The $\lambda||s_i||_1$ term vanishes since it does not depend on the model parameters

$$\frac{\partial}{\partial w_{jk}}||x_i - Ws_i||^2 = 2(x_{i,j} - \sum_{l=1}^{h} w_{jl}s_{i,l})(-w_{jk}s_{i,k})\,.$$

Thus the gradient is

$$\frac{\partial}{\partial W} = 2\cdot\left(\eta w_{jk} + \Big(\sum_{i=1}^{N}(x_{i,j} - \sum_{l=1}^{h} w_{jl}s_{i,l})(-w_{jk}s_{i,k})\Big)\right)$$

□

Problem 1.b.

Solution: Now we want to compute the gradients wrt to the s_i. The first term obviously vanishes. The gradients of the λ terms are

$$\begin{aligned}\frac{\partial}{\partial s_i}(\lambda||s_i||_1) &= \lambda\left(\frac{\partial}{\partial s_{i,1}}||s_i||_1, ..., \frac{\partial}{\partial s_{i,h}}||s_i||_1\right)\\ &= \lambda\left(\frac{s_{i,1}}{|s_{i,1}|}, ..., \frac{s_{i,h}}{|s_{i,h}|}\right)\,,\end{aligned}$$

and zero when the we take the gradient wrt to another data point s_j. The other terms have the following gradients

$$\frac{\partial}{\partial s_i}||x_i + Ws_i||^2 = 2\left(\sum_{k=1}^{d}(x_{i,k} + \sum_{j=1}^{h} w_{kj}s_{i,j})\cdot(w_{k1}s_{i,1}), ..., \sum_{k=1}^{d}(x_{i,k} + \sum_{j=1}^{h} w_{kj}s_{i,j})\cdot(w_{kh}s_{i,h})\right)\,,$$

and zero when the we take the gradient wrt to another data point s_j. Thus the total gradient becomes

$$\frac{\partial E}{\partial s_i} = 2\lambda\left(\frac{s_{i,1}}{|s_{i,1}|} + \sum_{k=1}^{d}(x_{i,k} + \sum_{j=1}^{h} w_{kj}s_{i,j})\cdot(w_{k1}s_{i,1}), ..., \frac{s_{i,h}}{|s_{i,h}|} + \sum_{k=1}^{d}(x_{i,k} + \sum_{j=1}^{h} w_{kj}s_{i,j})\cdot(w_{kh}s_{i,h})\right)$$

□

Problem 2.a..

Solution: It can be seen that the original and the reparameterized objective function are generally equivalent except for the source norm penalization terms. Therefore, it has to be made sure that

$$||s_i||_1 = ||r_i||_2^2.$$

With the reparameterized source $s_i = g(r_i)$, this can be written as

$$||g(r_i)||_1 = ||r_i||_2^2.$$
$$\implies \sum_i \sum_j g_j(r_i) = \sum_i \sum_j r_{ij}^2$$

If we choose the reparameterization function to be $g_j(r_i) = r_{ij}^2$ this can be achieved. Furthermore, the positivity constraints from the original formulation are also met as $r_{ij} \in \mathbb{R}$ and therefore $r_{ij}^2 > 0 \quad \forall i, j$. □

Problem 2.b.

Solution: One of the disadvantages of the original formulation is that the L_0 "norm" is not differentiable and not convex, it is therefore not possible to apply gradient descent for optimization. However, only the L_0 "norm" exactly captures the idea of using the least amount of basis vectors possible.
Having the reparameterized objective of Ex. 2 enables us to use an encoder, which can be used to find a good initial set of sources and is faster to optimize. The downside of this is that in addition to the model parameters of the decoder we need to optimize the parameters of the encoder at the same time. □

Problem 3.a.

Solution: we have the objective of the form:

$$\min_{W, r_1, r_2, \ldots, r_N} \eta ||W||_F^2 + \sum_{i=1}^{N} ||x_i - W.g(r_i)||^2 + \lambda ||r_i||^2$$

with $r_i = V^\top x_i$, Now we use the chain rule to express the gradient of the objective:

$$\frac{\partial E}{\partial V} = \sum_i \frac{\partial E}{\partial r_i} . \frac{\partial r_i}{\partial V}$$
$$\frac{\partial E}{\partial V} = \sum_i \left(2.W.g(r_i)'(x_i - W.g(r_i)) + 2\lambda r_i \right) .x_i$$

□

Problem 3.b.

Solution: When using an autoencoder we are giving up direct control of the sources s_i. We also got two layers to train which makes the optimization problem non-convex and there in addition to the parameters of the decoder we need to train the parameters of the encoder. However, the optimization objective is simpler than for sparse coding with an encoder, since we set the inferred sources equal to the sources. □