Exercise A.

1. (a) $E = \eta \|W\|_F^2 + \sum_{i=1}^{N} \|x_i - Ws_i\|^2 + \lambda \|s_i\|_1 \quad \forall_{i=1}^{N} \; s_i \geq 0$

$$\frac{\partial E}{\partial W} = 2\eta W + \sum_{i=1}^{N} 2(x_i - Ws_i)(-s_i) = 0.$$

1. (b) $$\frac{\partial E}{\partial s_i} = 2(x_i - Ws_i)(-W) + \lambda \, \mathrm{sign}(s_i) = 0.$$

$\mathrm{sign}(s_i) = 1$ as $s_i \geq 0$.

$$\frac{\partial E}{\partial s_i} = 2(x_i - Ws_i)(-W) + \lambda \mathbb{1} = 0.$$

ML2 ES4 T2a) $\|r_i\|^2 \overset{!}{=} \|s_i\|_1$ with $s_i = g(r_i)$

$\Rightarrow \|r_i\|^2 = \|g(r_i)\|_1$ with $\|g(r_i)\|_1 = \sum_{i=1}^{N} |g(r_i)|$

$\Rightarrow \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} |g(r_i)|$ and $\|r_i\|^2 = \left(\sqrt{\sum_{i=1}^{N} r_i^2}\right)^2 = \sum_{i=1}^{N} r_i^2$

$\rightarrow g: r_i \rightarrow r_i^2$

$(g: \mathbb{R}^1 \rightarrow \mathbb{R}^1)$

b) (1) l1 norm is not applicable to gradient descent as it is not differentiable at 0. We can use a sparsity (or smoothing) parameter ϵ: $\sqrt{x^2+\epsilon}$ to be able to perform the gradient descent. The l2 norm is differentiable at each point.

(2) The l2 norm penalties discourage sparsity because of the diminishing returns when elements move closer to 0 but this is what we wish for in using the encoder.

T3a) $\frac{\partial E}{\partial V} = \sum_{i=1} \frac{\partial E}{\partial r_i} \cdot \underbrace{\frac{\partial r_i}{\partial V}}_{x_i} = \sum_{i=1} \frac{\partial E}{\partial g(\cdot)} \cdot \underbrace{\frac{\partial g(\cdot)}{\partial r_i}}_{g'(\cdot)} \cdot x_i = \sum \frac{\partial E}{\partial s_i} \cdot \underbrace{\frac{\partial s_i}{\partial g(\cdot)}}_{\Rightarrow \frac{\partial g(\cdot)}{\partial g(\cdot)} = 1} \cdot g'(\cdot) \cdot x_i$

$= \frac{\partial E}{\partial \hat{x}_i} \cdot \underbrace{\frac{\partial \hat{x}_i}{\partial s_i}}_{W} \cdot g'(\cdot) \cdot x_i$

with $E = \eta \|W\|_F^2 + \sum_{i=1}^{N} \|x_i - W g(r_i)\|^2 + \lambda \|r_i\|^2$

and $\frac{\partial E}{\partial \hat{x}_i} = x_i - \hat{x}_i$ where $\hat{x}_i = W g(\cdot)$

$\rightarrow \frac{\partial E}{\partial V} = \sum_{i=1} (x_i - \hat{x}_i) \cdot W \cdot g'(\cdot) \cdot x_i$

Exercise 3b
Advantages:

- it is often faster and simpler to obtain sparse representations via autoencoders
- huge reduction in parameters(example: in case of natural images)
- sparsifying non-linearity
- the estimate of the expectation E [h_j(x; W, b)] is very noisy in direct optimization but autoencoder denoises the data
- easier to implement
- faster to optimize (no need to keep track of source codes)
- good initial guess for si, optimize from there --> save iterations of source optimization
- can be trained by backpropagation

Disadvantages:

- 2 layers to train
- (more parameters)
- no control of regularization
- bad encoder --> bad decoder