

Exercise Sheet 4

Exercise 1: Sparse Coding (5+5 P)

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d$ be a dataset of N examples. Let $\mathbf{s}_i \in \mathbb{R}^h$ be the source associated to example \mathbf{x}_i , and $W \in \mathbb{R}^{d \times h}$ be a matrix of size $d \times h$ that linearly projects the source onto the reconstructed example $\hat{\mathbf{x}}_i$. We optimize the following sparse coding objective:

$$\min_{W, \mathbf{s}_1, \dots, \mathbf{s}_N} \eta \|W\|_F^2 + \sum_{i=1}^N \|\mathbf{x}_i - W\mathbf{s}_i\|^2 + \lambda \|\mathbf{s}_i\|_1 \quad \text{where } \forall_{i=1}^N : \mathbf{s}_i \geq 0$$

- Compute the gradient of the objective with respect to the model parameters W . (i.e. compute the matrix $\frac{\partial E}{\partial W}$).
- Compute the gradient of the objective with respect to the sources \mathbf{s}_i for each data point.

Exercise 2: Sparsifying Non-Linearities (10+10 P)

As an alternative to the sparse coding problem above, we would like to minimize the reparameterized objective of the form:

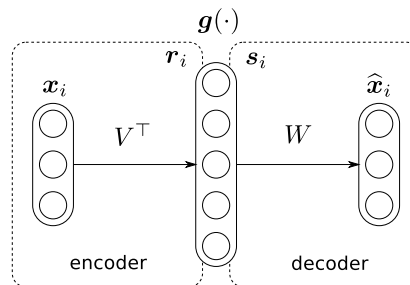
$$\min_{W, \mathbf{r}_1, \dots, \mathbf{r}_N} \eta \|W\|_F^2 + \sum_{i=1}^N \|\mathbf{x}_i - W\mathbf{g}(\mathbf{r}_i)\|^2 + \lambda \|\mathbf{r}_i\|^2 \quad \text{where } \forall_{i=1}^N : \mathbf{r}_i \in \mathbb{R}^h$$

We call \mathbf{r}_i the source parameter, and $\mathbf{s}_i = \mathbf{g}(\mathbf{r}_i)$ the reparameterized source associated to example \mathbf{x}_i . Note that the new objective no longer involves the minimization of an L_1 -norm, and also does not include positivity constraints.

- Find a reparameterization function $\mathbf{g} : \mathbb{R}^h \rightarrow \mathbb{R}^h$ for which the optimization problem is equivalent to the one of Exercise 1.
- Explain what are the advantages and disadvantages of using such formulation of the optimization problem when compared to the original sparse coding problem. Your answer may include: (1) Applicability of gradient descent to find sources \mathbf{s}_i . (2) Ease of using an encoder to initialize the search for optimal sources.

Exercise 3: Auto-Encoders (10+10 P)

We now give an explicit definition of the encoder $\mathbf{r}_i = V^\top \mathbf{x}_i$, where $V \in \mathbb{R}^{d \times h}$ is a matrix of size $d \times h$. A graphical depiction of the resulting auto-encoder for $\mathbf{x}_i \in \mathbb{R}^3$ and $\mathbf{r}_i, \mathbf{s}_i \in \mathbb{R}^5$ is given below:



- Assuming the same error function as in Exercise 2, use the chain rule to express the gradient of the objective with respect to the encoder parameter $\frac{\partial E}{\partial V}$.
- Explain what are the advantages and disadvantages of using an autoencoder instead of directly optimizing \mathbf{s}_i or \mathbf{r}_i . Your answer should include the following aspects: (1) Computational requirements of inferring sources \mathbf{s}_i from observations \mathbf{x}_i . (2) Difficulty of the optimization problem. (3) Computational requirements at training time.

Exercise 4: Programming (50 P)

Download the programming files on ISIS and follow the instructions.