



Exercise Sheet 4

Machine Learning 2, SS16

May 19, 2016

Mario Tambos, 380599; Viktor Jeney, 348969; Sascha Huk, 321249; Jan Tinapp, 0380549

Exercise 1 - Sparse Coding

(a)

$$\begin{aligned}\frac{\partial E}{\partial W} &= \frac{\partial}{\partial W} \eta \|W\|_F^2 + \frac{\partial}{\partial W} \sum_{i=1}^N (|x^{(i)} - W s^{(i)}|^2 + \lambda |s^{(i)}|_1) \\ &= \eta \sum_l \sum_k \frac{\partial}{\partial W} (W_{lk})^2 + \sum_{i=1}^N \frac{\partial}{\partial W} (x^{(i)} - W s^{(i)})^\top (x^{(i)} - W s^{(i)}) \\ &= 2\eta W + \sum_{i=1}^N -2(x^{(i)} - W s^{(i)}) s^{(i)\top} = 2\eta W - 2 \sum_{i=1}^N (x^{(i)} - W s^{(i)}) s^{(i)\top}\end{aligned}$$



(b)

$$\begin{aligned}\frac{\partial E}{\partial s^{(i)}} &= \frac{\partial}{\partial s^{(i)}} \eta \|W\|_F^2 + \frac{\partial}{\partial s^{(i)}} \sum_{j=1}^N (|x^{(j)} - W s^{(j)}|^2 + \lambda |s^{(j)}|_1) \\ &= \frac{\partial}{\partial s^{(i)}} (x^{(i)} - W s^{(i)})^\top (x^{(i)} - W s^{(i)}) + \frac{\partial}{\partial s^{(i)}} \lambda |s^{(i)}|_1 \\ &= -2W^\top (x^{(i)} - W s^{(i)}) + \lambda \sum_{k=1}^h \frac{\partial}{\partial s^{(i)}} s^{(i)}_k \quad (s^{(i)}_k \geq 0) \\ &= -2W^\top (x^{(i)} - W s^{(i)}) + \lambda \mathbf{1}_h\end{aligned}$$



Exercise 2 - Sparsifying Non-Linearities

(a) The derivative wrt. to W is already equivalent. Taking the derivative wrt. $r^{(i)}$ we obtain:

$$\begin{aligned}\frac{\partial}{\partial r^{(i)}} \sum_{j=1}^N (|x^{(j)} - W g(r^{(j)})|^2 + \lambda |r^{(j)}|^2) &= \frac{\partial}{\partial r^{(i)}} |x^{(i)} - W g(r^{(i)})|^2 + \frac{\partial}{\partial r^{(i)}} \lambda |r^{(i)}|^2 \\ &= -2W^\top (x^{(i)} - W g(r^{(i)})) \frac{\partial}{\partial r^{(i)}} g(r^{(i)}) + \lambda 2r^{(i)}\end{aligned}$$

Comparing the factors of λ yields componentwise differences by the componentwise factors $2r^{(i)}_k \forall k \in \{1, \dots, h\}$. In order for the two problems to be equal, we choose g such that it's Jacobian is diagonal and such that it has the k -th of these factors on the k -th diagonal element. Therefore we have an equivalent problem for $g(r^{(i)}) = (r^{(i)}_1^2, \dots, r^{(i)}_h^2)$.

(b) Both, the reconstruction error as well as the sparsity penalty, are convex in both approaches. Since the sum of convex functions is again a convex function we find a global, **unique** optimum in both problems. Unfortunately, the derivative of the original problem wrt. s_k for $k \in \{1, \dots, h\}$ does not exist in 0. It is therefore theoretically problematic to use gradient descent on the original problem, even though encountering this problem might be rather unlikely.



Exercise 3 - Autoencoders

(a) Focus on the result of (2a), which is $\frac{\partial E}{\partial r^{(i)}}$. Having $r^{(i)} = V^\top x^{(i)}$, by chainrule we get:

$$\begin{aligned}\frac{\partial E}{\partial V} &= \frac{\partial E}{\partial r^{(i)}} \frac{\partial r^{(i)}}{\partial V} = [-2W^\top(x^{(i)} - Wg(r^{(i)})) \frac{\partial}{\partial r^{(i)}} g(r^{(i)}) + \lambda 2r^{(i)}] x^{(i)\top} \\ &= [-2W^\top(x^{(i)} - Wg(V^\top x^{(i)})) g'(V^\top x^{(i)}) + \lambda 2V^\top x^{(i)}] x^{(i)\top}\end{aligned}$$



(b)

(2) Optimizing wrt. W, s^1, \dots, s^N and V is not a convex optimization problem while the approaches in task 1 and 2 are convex. The problem became harder to optimize.

(1) Inferring the sources $r^{(i)}$ from $x^{(i)}$ is just the product $V^\top x^{(i)}$ (requires $d \times h$ multiplications). This does not endanger the feasibility of this method.

(3) Non-convexity (or non-concavity) generally endangers the feasibility of finding a solution. The computation generally becomes either harder or more approximative. In comparison, the 1-layer network is comparatively quick and exact solvable.

