

---

# Supplemental equations for Regularized estimation of image statistics by Score Matching

---

**Diederik P. Kingma**

Department of Information and Computing Sciences  
Universiteit Utrecht  
d.p.kingma@students.uu.nl

**Yann LeCun**

Courant Institute of Mathematical Sciences  
New York University  
yann@cs.nyu.edu

## 1 Differentiation equations

Recall that we split the SM loss into two terms:  $J(\mathbf{x}; \mathbf{w}) = K + L$  with  $K = \frac{1}{2} \sum_{j=1}^D \delta_j^2$  and  $L = \sum_{j=1}^D -\delta'_j + \lambda(\delta'_j)^2$ .

### 1.1 First partial derivatives

The derivate of the energy with respect to the network's output is trivial:  $\delta_{out} = \frac{\partial E}{\partial g_{out}} = 1$ . The calculation of the derivative of the energy w.r.t. vertex  $j$  is a simple application of the chain rule:

$$\delta_j = \frac{\partial E}{\partial g_j} = \sum_{k \in V_j} \frac{\partial E}{\partial g_k} \frac{\partial g_k}{\partial g_j} = \sum_{k \in V_j} \delta_k g'_{j,k} \quad (1)$$

### 1.2 Backpropagation of full Hessian

For any network output vertex,  $\delta'_{out} = 0$  since  $\frac{\partial \delta_{out}}{\partial g_{out}} = \frac{\partial(1)}{\partial g_{out}} = 0$ . The general formula for backpropagating the full Hessian through the layer is:

$$\begin{aligned} \frac{\partial^2 E}{\partial f_i \partial f_j} &= \frac{\partial \left( \frac{\partial E}{\partial f_i} \right)}{\partial f_j} = \frac{\partial \left( \sum_{k \in V_j} \frac{\partial E}{\partial g_k} \frac{\partial g_k}{\partial f_i} \right)}{\partial f_j} \\ &= \sum_{k \in V_j} \left( \sum_{l \in V_j} \frac{\partial^2 E}{\partial g_k \partial g_l} \frac{\partial g_l}{\partial f_j} \frac{\partial g_k}{\partial f_i} \right) + \frac{\partial E}{\partial g_k} \frac{\partial^2 g_k}{\partial f_i \partial f_j} \end{aligned} \quad (2)$$

### 1.3 Backpropagation of trace of the Hessian

To derivation equation for computing the trace of the Hessian for a node is:

$$\begin{aligned}
\delta'_j &= \frac{\partial \delta_j}{\partial g_j} \\
&= \frac{\partial \left( \sum_{k \in V_j} \delta_k \frac{\partial g_k}{\partial g_j} \right)}{\partial g_j} \\
&= \sum_{k \in V_j} \delta_k \frac{\partial^2 g_k}{(\partial g_j)^2} + \frac{\partial \delta_k}{\partial g_j} \frac{\partial g_k}{\partial g_j} \\
&= \sum_{k \in V_j} \delta_k \frac{\partial^2 g_k}{(\partial g_j)^2} + \frac{\partial \delta_k}{\partial g_k} \frac{\partial g_k}{\partial g_j} \frac{\partial g_k}{\partial g_j} \\
&= \sum_{k \in V_j} \delta_k g''_{j,k} + \delta'_k [g'_{j,k}]^2
\end{aligned}$$

## 2 Differentiating the SM loss

### 2.1 Forward propagation of $\frac{\partial K}{\partial \delta_j}$

For the input nodes:

$$\frac{\partial K}{\partial \delta_j} = \frac{\partial \sum_i \frac{1}{2} \delta_i^2}{\partial \delta_j} = \delta_j \quad (3)$$

For the other nodes:

$$\begin{aligned}
\frac{\partial K}{\partial \delta_j} &= \sum_{i \in U_j} \frac{\partial K}{\partial \delta_i} \frac{\partial \delta_i}{\partial \delta_j} \\
&= \sum_{i \in U_j} \frac{\partial K}{\partial \delta_i} \frac{\partial \left[ \delta_j \frac{\partial g_i}{\partial g_i} \right]}{\partial \delta_j} \\
&= \sum_{i \in U_j} \frac{\partial K}{\partial \delta_i} \frac{\partial g_j}{\partial g_i}
\end{aligned} \quad (4)$$

### 2.2 Forward propagation of $\frac{\partial L}{\partial \delta_j}$

For the input nodes:

$$\frac{\partial L}{\partial \delta_j} = 0 \quad (5)$$

For the other nodes:

$$\begin{aligned}
\frac{\partial L}{\partial \delta_j} &= \sum_{i \in U_j} \frac{\partial L}{\partial \delta'_i} \frac{\partial \delta'_i}{\partial \delta_j} + \frac{\partial L}{\partial \delta_i} \frac{\partial \delta_i}{\partial \delta_j} \\
&= \sum_{i \in U_j} \frac{\partial L}{\partial \delta'_i} \frac{\partial \left[ \delta_j \frac{\partial^2 g_j}{(\partial g_i)^2} + \delta'_j \left( \frac{\partial g_j}{\partial g_i} \right)^2 \right]}{\partial \delta_j} + \frac{\partial L}{\partial \delta_i} \frac{\partial \left[ \delta_j \frac{\partial g_j}{\partial g_i} \right]}{\partial \delta_j} \\
&= \sum_{i \in U_j} \frac{\partial L}{\partial \delta'_i} \frac{\partial^2 g_j}{(\partial g_i)^2} + \frac{\partial L}{\partial \delta_i} \frac{\partial g_j}{\partial g_i}
\end{aligned} \quad (6)$$

### 2.3 Forward propagation of $\frac{\partial L}{\partial \delta'_j}$

For the input nodes:

$$\frac{\partial L}{\partial \delta'_j} = \frac{\partial (-\sum_i \delta'_i)}{\partial \delta'_j} = -1 \quad (7)$$

Assuming the diagonal Hessian backpropagation method is exact, the equation for the other nodes is:

$$\begin{aligned} \frac{\partial L}{\partial \delta'_j} &= \sum_{i \in U_j} \frac{\partial L}{\partial \delta'_i} \frac{\partial \delta'_i}{\partial \delta'_j} \\ &= \sum_{i \in U_j} \frac{\partial L}{\partial \delta'_i} \frac{\partial \left[ \delta_j \frac{\partial^2 g_j}{(\partial g_i)^2} + \delta'_j \left( \frac{\partial g_j}{\partial g_i} \right)^2 \right]}{\partial \delta'_j} \\ &= \sum_{i \in U_j} \frac{\partial L}{\partial \delta'_i} \left( \frac{\partial g_j}{\partial g_i} \right)^2 \end{aligned} \quad (8)$$

### 2.4 Backward propagation of $\frac{\partial K}{\partial g_j}$

$$\begin{aligned} \frac{\partial K}{\partial g_j} &= \sum_{k \in V_j} \frac{\partial K}{\partial g_k} \frac{\partial g_k}{\partial g_j} + \frac{\partial K}{\partial \delta_j} \frac{\partial \delta_j}{\partial \left( \frac{\partial g_k}{\partial g_j} \right)} \frac{\partial \left( \frac{\partial g_k}{\partial g_j} \right)}{\partial g_j} \\ &= \sum_{k \in V_j} \frac{\partial K}{\partial g_k} \frac{\partial g_k}{\partial g_j} + \frac{\partial K}{\partial \delta_j} \frac{\partial \left( \delta_k \frac{\partial g_k}{\partial g_j} \right)}{\partial \left( \frac{\partial g_k}{\partial g_j} \right)} \frac{\partial^2 g_k}{(\partial g_j)^2} \\ &= \sum_{k \in V_j} \frac{\partial K}{\partial g_k} \frac{\partial g_k}{\partial g_j} + \frac{\partial K}{\partial \delta_j} \delta_k \frac{\partial^2 g_k}{(\partial g_j)^2} \end{aligned} \quad (9)$$

## 2.5 Backward propagation of $\frac{\partial L}{\partial g_j}$

$$\begin{aligned}
\frac{\partial L}{\partial g_j} &= \sum_{k \in V_j} \frac{\partial L}{\partial g_k} \frac{\partial g_k}{\partial g_j} + \frac{\partial L}{\partial \delta_j} \frac{\partial \delta_j}{\partial \left( \frac{\partial g_k}{\partial g_j} \right)} \frac{\partial \left( \frac{\partial g_k}{\partial g_j} \right)}{\partial g_j} \\
&\quad + \frac{\partial L}{\partial \delta'_j} \frac{\partial \delta'_j}{\partial \left( \frac{\partial g_k}{\partial g_j} \right)} \frac{\partial \left( \frac{\partial g_k}{\partial g_j} \right)}{\partial g_j} \\
&\quad + \frac{\partial L}{\partial \delta'_j} \frac{\partial \delta'_j}{\partial \left( \frac{\partial^2 g_k}{(\partial g_j)^2} \right)} \frac{\partial \left( \frac{\partial^2 g_k}{(\partial g_j)^2} \right)}{\partial g_j} \\
&= \sum_{k \in V_j} \frac{\partial L}{\partial g_k} \frac{\partial g_k}{\partial g_j} + \frac{\partial L}{\partial \delta_j} \frac{\partial \left( \delta_k \frac{\partial g_k}{\partial g_j} \right)}{\partial \left( \frac{\partial g_k}{\partial g_j} \right)} \frac{\partial^2 g_k}{(\partial g_j)^2} \\
&\quad + \frac{\partial L}{\partial \delta'_j} \frac{\partial \left( \delta_k \frac{\partial^2 g_k}{(\partial g_j)^2} + \delta'_k \left( \frac{\partial g_k}{\partial g_j} \right)^2 \right)}{\partial \left( \frac{\partial g_k}{\partial g_j} \right)} \frac{\partial^2 g_k}{(\partial g_j)^2} \\
&\quad + \frac{\partial L}{\partial \delta'_j} \frac{\partial \left( \delta_k \frac{\partial^2 g_k}{(\partial g_j)^2} + \delta'_k \left( \frac{\partial g_k}{\partial g_j} \right)^2 \right)}{\partial \left( \frac{\partial^2 g_k}{(\partial g_j)^2} \right)} \frac{\partial^3 g_k}{(\partial g_j)^3} \\
&= \sum_{k \in V_j} \frac{\partial L}{\partial g_k} \frac{\partial g_k}{\partial g_j} + \frac{\partial L}{\partial \delta_j} \delta_k \frac{\partial^2 g_k}{(\partial g_j)^2} \\
&\quad + 2 \frac{\partial L}{\partial \delta'_j} \delta'_k \frac{\partial g_k}{\partial g_j} \frac{\partial^2 g_k}{(\partial g_j)^2} + \frac{\partial L}{\partial \delta'_j} \delta_k \frac{\partial^3 g_k}{(\partial g_j)^3}
\end{aligned} \tag{10}$$

## 3 Equations for specific neural-network layers

In many neural network implementations, the datastructure of a hidden state is a vector, and component functions are implemented by so-called 'layers', which map from one or more input states to an output state (being 'input' or 'output' w.r.t. the layer, not network). A neural network consists of an ordered set of layers. Forward- and backward propagation consists of walking through these layers in their ordering. We will now derive the the gradient propagation functions for common layers, which were only given in general in the previous section and in summary in Algorithm 1. There are two common layer types: linear layers that perform linear transformations, and non-linear layers that perform point-wise nonlinear transformations.

### 3.1 Fully connected linear layer

A so-called 'linear layer' perform a linear transformation on its input, usually in the form of a matrix multiplication. Consider a linear layer with  $N$ -dimensional input vector  $\mathbf{g}$  and  $M$ -dimensional output vector  $\mathbf{f}$ . With  $w_{ji}$  we denote the element at the  $j$ -th row and  $i$ -th column of an  $M \times N$ -dimensional transformation/weight matrix  $\mathbf{W}$ . Forward propagation is performed by  $g_j = \sum_i w_{ji} f_i$  or in matrix notation:  $\mathbf{g} = \mathbf{W}\mathbf{f}$ .

For this linear transformation,  $\frac{\partial g_j}{\partial f_i} = w_{ji}$ ,  $\frac{\partial^2 g_j}{(\partial f_i)^2} = 0$  and  $\frac{\partial^3 g_j}{(\partial f_i)^3} = 0$ . This greatly simplifies the propagation equations.

Backpropagation consists of:  $\frac{\partial E}{\partial f_i} = \sum_j \frac{\partial E}{\partial g_j} \frac{\partial g_j}{\partial f_i} = \sum_j w_{ji} \frac{\partial E}{\partial g_j}$  or in matrix notation:  $\delta_f = \mathbf{W}^T \delta_g$ , where  $\delta_f = \left[ \frac{\partial E}{\partial f_1} \dots \frac{\partial E}{\partial f_n} \right]^T$  and  $\delta_g = \left[ \frac{\partial E}{\partial g_1} \dots \frac{\partial E}{\partial g_n} \right]^T$ . From this point we will not use matrix

notation, since we would have to introduce so many new vector symbols for different derivatives, that it is probably easier to read and understand the equations in non-matrix form. It is important to keep in mind that most computations can be done in matrix form though. Consider  $\delta_i = \frac{\partial E}{\partial f_i}$  and  $\delta_j = \frac{\partial E}{\partial g_j}$ .

### 3.1.1 Forward propagation of SM derivatives

$$\frac{\partial K}{\partial \delta_j} = \sum_i w_{ji} \frac{\partial K}{\partial \delta_i} \quad (11)$$

$$\frac{\partial L}{\partial \delta_j} = \sum_i w_{ji} \frac{\partial L}{\partial \delta_i} \quad (12)$$

$$\frac{\partial L}{\partial \delta'_j} = \sum_i (w_{ji})^2 \frac{\partial L}{\partial \delta'_i} \quad (13)$$

### 3.1.2 Backward propagation of SM derivatives

$$\frac{\partial K}{\partial f_i} = \sum_j w_{ji} \frac{\partial K}{\partial g_j} \quad (14)$$

$$\frac{\partial L}{\partial f_i} = \sum_j w_{ji} \frac{\partial L}{\partial g_j} \quad (15)$$

Recall that  $\frac{\partial J}{\partial f_i} = \frac{\partial K}{\partial f_i} + \frac{\partial L}{\partial f_i}$ . Again, the equations above are very simple since the second and third partial derivatives of the elements of  $\mathbf{g}$  w.r.t. the elements of  $\mathbf{f}$  are zero.

### 3.1.3 Derivatives w.r.t. the weights

$$\frac{\partial K}{\partial w_{ji}} \left( = \frac{\partial K}{\partial \delta_i} \frac{\partial \delta_i}{\partial w_{ji}} + \frac{\partial K}{\partial g_j} \frac{\partial g_j}{\partial w_{ji}} \right) = \frac{\partial K}{\partial \delta_i} \delta_j + \frac{\partial K}{\partial g_j} f_i \quad (16)$$

$$\frac{\partial L}{\partial w_{ji}} \left( = \frac{\partial L}{\partial \delta_i} \frac{\partial \delta_i}{\partial w_{ji}} + \frac{\partial L}{\partial \delta'_i} \frac{\partial \delta'_i}{\partial w_{ji}} + \frac{\partial L}{\partial g_j} \frac{\partial g_j}{\partial w_{ji}} \right) = \frac{\partial L}{\partial \delta_i} \delta_j + \frac{\partial L}{\partial \delta'_i} 2w_{ji} \delta'_j + \frac{\partial L}{\partial g_j} f_i \quad (17)$$

## 3.2 Piecewise non-linear layer

In a common type of non-linear layer, the elements of the input vector undergo a piecewise non-linear transformation  $g_i \rightarrow g_j$ , such that  $g_j = f(g_i)$ . Since the operation  $f(\cdot)$  is non-linear, the second and third derivatives of  $f(\cdot)$  are non-zero (with the exception of quadratic  $f(\cdot)$ , e.g.  $g_j = g_i^2$ , for which the third derivative is zero). Consider  $\delta_i = \frac{\partial E}{\partial g_i}$  and  $\delta_j = \frac{\partial E}{\partial g_j}$ .

The following equations are almost equal to the generic equations, with the only difference that input nodes  $g_i$  are connected to only one output node  $g_j$  since the transformation is piecewise.

### 3.2.1 Forward propagation of SM derivatives

$$\frac{\partial K}{\partial \delta_j} = \frac{\partial K}{\partial \delta_i} \frac{\partial g_j}{\partial g_i} \quad (18)$$

$$\frac{\partial L}{\partial \delta_j} = \frac{\partial L}{\partial \delta'_i} \frac{\partial^2 g_j}{(\partial g_i)^2} + \frac{\partial L}{\partial \delta_i} \frac{\partial g_j}{\partial g_i} \quad (19)$$

$$\frac{\partial L}{\partial \delta'_j} = \frac{\partial L}{\partial \delta'_i} \left( \frac{\partial g_j}{\partial g_i} \right)^2 \quad (20)$$

### 3.2.2 Backward propagation of SM derivatives

$$\frac{\partial K}{\partial g_i} = \frac{\partial K}{\partial g_j} \frac{\partial g_j}{\partial g_i} + \frac{\partial K}{\partial \delta_i} \delta_j \frac{\partial^2 g_j}{(\partial g_i)^2} \quad (21)$$

$$\frac{\partial L}{\partial g_i} = \frac{\partial L}{\partial g_j} \frac{\partial g_j}{\partial g_i} + \frac{\partial L}{\partial \delta_i} \delta_j \frac{\partial^2 g_j}{(\partial g_i)^2} + 2 \frac{\partial L}{\partial \delta'_i} \delta'_j \frac{\partial g_j}{\partial g_i} \frac{\partial^2 g_j}{(\partial g_i)^2} + \frac{\partial L}{\partial \delta'_i} \delta_j \frac{\partial^3 g_j}{(\partial g_i)^3} \quad (22)$$