Set2Model Networks: Learning Discriminatively To Learn Generative Models Supplemental Materials

Alexander Vakhitov
a.vakhitov@skoltech.ru

Andrey Kuzmin
a.kuzmin@skolkovotech.ru

Victor Lempitsky
lempitsky@skoltech.ru
Skolkovo Institute of Science and Technology
Moscow, Russia

1. A1. Typical structure of a system matrix for implicit function differentiation

In the following appendix, we derive formulas for the required derivatives of the likelihood function for the Gaussian Mixture models with diagonal covariance matrices, and show that very often the second derivative matrices become sparse during meta-learning, which accelerates the process significantly.

The log-likelihood function $l(\theta|D)$ is a sum of one-observation likelihood functions as given in the main text, eq. (1).

Theorem. If log-likelihood function $\log h(\theta|d)$ corresponding to a mixture model can be represented as a logarithm of a sum of multiplied coordinate-wise likelihood functions with non-intersecting parameter sets:

$$\log h(\theta|d) = \log\{\sum_{i=1}^{k} v_i \prod_{j=1}^{n} g(\theta_{i,j}|d)\}, \tag{1}$$

where v_i are mixture weights, $v_i \geq 0$, $\sum_{i=1}^{k} v_i = 1$, $\theta = [\theta_1^T, \theta_2^T, \dots, \theta_k^T, v_1, v_2, \dots, v_k]^T$ is the vector of the model parameters, $\theta_i = [\theta_{i,1}^T, \theta_{i,2}^T, \dots, \theta_{i,q}^T]^T$ is the vector of one mixture component's parameters, $\theta_{ij} = [\theta_{i,j,1}, \theta_{i,j,2}, \dots, \theta_{i,j,n_c}]^T$ is the vector of the coordinate-wise likelihood function parameters of length n_c, $g(\theta_{i,j}|d)$ is a coordinate-wise likelihood function for the coordinate j, *then* the second derivatives of $\log h(\theta_d)$ w.r.t. the parameters of the model (except the weights) are

$$\frac{\partial^2 \log h(\theta|d)}{\partial\theta_{i,j,k}\partial\theta_{u,s,t}} =$$

$$\begin{cases} (r_i - r_i^2)\frac{g'_k(\theta_{i,j}|d_{(j)})g'_t(\theta_{i,s}|d_{(s)})}{g(\theta_{i,j}|d_{(j)})g(\theta_{i,s}|d_{(s)})}, & i = u, j \neq s, \\ r_i\frac{g''_{kt}(\theta_{i,j}|d_{(j)})}{g(\theta_{i,j}|d_{(j)})} - r_i^2\frac{g'_k(\theta_{i,j}|d_{(j)})g'_t(\theta_{i,j}|d_{(j)})}{(g(\theta_{i,j}|d_{(j)}))^2}, & i = u, j = s, \\ -\frac{1}{h^2(\theta|d)}r_i r_u \frac{g'_k(\theta_{i,j}|d_{(j)})}{g(\theta_{i,j}|d_{(j)})}\frac{g'_t(\theta_{u,s}|d_{(s)})}{g(\theta_{u,s}|d_{(s)})}, & i \neq u, \end{cases} \tag{2}$$

where responsibility of the i-th component is defined as a fraction $r_i = \frac{v_i \prod_{j=1}^{q} g(\theta_{i,j}|d_{(j)})}{h(\theta|d)}$.

Proof.

The first derivative is:

$$\frac{\partial}{\partial\theta_{i,j,k}} \log h(\theta|d) = \frac{1}{h(\theta|d)} v_i \prod_{m=1}^{q} g(\theta_{i,m}|d_{(m)}) \frac{g'_k(\theta_{i,j}|d_j)}{g(\theta_{i,j}|d_{(j)})}. \tag{3}$$

In case when both differentiation variables are the parameters of the same mixture component, the second derivative is:

$$\frac{\partial^2}{\partial\theta_{i,j,k}\partial\theta_{i,s,t}} \log h(\theta|d) = \frac{1}{h(\theta|d)} v_i \prod_{m=1}^{q} g(\theta_{i,m}|d_{(m)}) \frac{g'_k(\theta_{i,j}|d_{(j)})g'_t(\mu_{i,s}|d_{(s)})}{g(\theta_{i,j}|d_{(j)})g(\theta_{i,s}|d_{(s)})} - \tag{4}$$

$$-\frac{1}{h^2(\theta|d)} v_i^2 \left(\prod_{m=1}^{q} g(\theta_{i,m}|d_{(m)})\right)^2 \frac{g'_k(\theta_{i,j}|d_{(j)})g'_t(\theta_{i,s}|d_{(s)})}{g(\theta_{i,j}|d_{(j)})g(\theta_{i,s}|d_{(s)})} = \tag{5}$$

$$= (r_i(d) - r_i^2(d))\frac{g'_k(\theta_{i,j}|d_{(j)})g'_t(\theta_{i,s}|d_{(s)})}{g(\theta_{i,j}|d_j)g(\theta_{i,s}|d_{(s)})}. \tag{6}$$

In case when both differentiation variables are the parameters of the same coordinate likelihood function for the same mixture component, the second derivative is:

$$\frac{\partial^2}{\partial\theta_{i,j,k}\partial\theta_{i,j,t}} \log h(\theta|d) = r_i(d)\frac{g''_{kt}(\theta_{i,j}|d_j)}{g(\theta_{i,j}|d_j)} - \tag{7}$$

$g(\theta_{i,j}\|d)$	$(\sqrt{2\pi}\sigma_{i,j})^{-1}e^{-\frac{(d_{(j)}-\mu_{i,j})^2}{2\sigma_{i,j}^2}}$
$g'_\mu(\theta_{i,j}\|d_{(j)})$	$\frac{d_{(j)}-\mu_{i,j}}{\sigma_{i,j}^2}g(\theta_{i,j}\|d_{(j)})$
$g'_\sigma(\theta_{i,j}\|d_{(j)})$	$\sigma_{i,j}^{-1}(\frac{(d_{(j)}-\mu_{i,j})^2}{\sigma_{i,j}^2}-1)g(\theta_{i,j}\|d_{(j)})$
$g''_{\mu\mu}(\theta_{i,j}\|d_{(j)})$	$(-\frac{1}{\sigma_{i,j}^2}+\frac{(d_{(j)}-\mu_{i,j})^2}{\sigma_{i,j}^4})g(\theta_{i,j}\|d_{(j)})$
$g''_{\sigma\sigma}(\theta_{i,j}\|d_{(j)})$	$\sigma_{i,j}^{-2}(2-5\frac{(d_{(j)}-\mu_{i,j})^2}{\sigma_{i,j}^2}+\frac{(d_{(j)}-\mu_{i,j})^4}{\sigma_{i,j}^4})g(\theta_{i,j}\|d_{(j)})$
$g''_{\sigma\mu}(\theta_{i,j}\|d_{(j)})$	$(-3\frac{d_{(j)}-\mu_{i,j}}{\sigma_{i,j}^3}+\frac{(d_{(j)}-\mu_{i,j})^3}{\sigma_{i,j}^5})g(\theta_{i,j}\|d_{(j)})$

Table 1. Coordinate-wise likelihood function and its derivatives for the Gaussian mixture models.

$$-r_i(d)^2\frac{g'_k(\theta_{i,j}|d_j)g'_t(\theta_{i,j}|d_j)}{(g(\theta_{i,j}|d_j))^2}.$$

In case when the differentiation variables are the parameters of different mixture components, the second derivative is:

$$\frac{\partial^2}{\partial\theta_{i,j,k}\partial\theta_{u,s,t}}\log h(\theta|d) = -\frac{1}{h^2(\theta|d)}v_i\prod_{m=1}^{q}g(\theta_{i,m}|d_{(m)}) \tag{8}$$

$$\frac{g'_k(\theta_{i,j}|d_j)}{g(\theta_{i,j}|d_{(j)})}v_u\prod_{m=1}^{q}g(\theta_{u,m}|d_{(m)})\frac{g'_t(\theta_{u,s}|d_{(s)})}{g(\theta_{u,s}|d_{(s)})} = \tag{9}$$

$$= -\frac{1}{h^2(\theta|d)}r_i(d)r_u(d)\frac{g'_k(\theta_{i,j}|d_{(j)})}{g(\theta_{i,j}|d_{(j)})}\frac{g'_t(\theta_{u,s}|d_{(s)})}{g(\theta_{u,s}|d_{(s)})}. \tag{10}$$

Corollary. In conditions of the Theorem, if for an observation responsibility of some mixture component i is equal to 1, then among the second derivatives, only the ones corresponding to the same mixture component and coordinate can differ from zero.

Proof. Those derivatives which contain a term $r_i - r_i^2$ become zero, and the fact $r_i = 1$ leads to $r_u = 0$ for $u \neq i$, so only the second case ($i = u, j = s$) of the theorem formulation can lead to a non-zero second order derivative.

The responsibility of an observation is equal to 1 (up to numerical precision) in some 95% of cases in our experiments. Therefore second derivative matrices are very often sparse, having block-diagonal structure, which follows from the Corollary. It makes the meta-learning process for the Gaussian mixture models faster.

For the Gaussian mixture model with diagonal covariance matrices, the coordinate-wise function $g(\theta_{i,j}|d)$ and its derivatives w.r.t. the parameters $\theta_{i,j} = [\mu_{i,j}, \sigma_{i,j}]^T$ is given in the table 1.