Supplementary Material for Hessian-Sum-Mixtures

Vassili Korotkine, Mitchell Cohen, and James Richard Forbes
Department of Mechanical Engineering, McGill University
817 Sherbrooke Street West, Montreal QC H3A 0C3

April 2, 2024

---

1 Practical Implementation

The Jacobians of the different mixtures as well as details of multipliers for practical implementation are covered in this section. The negative log likelihoods are propertional to, but not exactly equal to, the squared error terms because of the normalization constants required to make sure the square root argument does not become negative.

1.1 Max-Mixture

The log-likelihood is computed from the max-mixture approximation to the Gaussian mixture such that

$$-\log p_{\max}(\mathbf{y}|\mathbf{x}) = \frac{1}{2}\mathbf{e}_{\max}^{\mathsf{T}}\mathbf{e}_{\max} \tag{1}$$

$$= -\log \alpha_{k^*} + \frac{1}{2}\mathbf{e}_{k^*}^{\mathsf{T}}\mathbf{e}_{k^*}, \quad k^* = \arg\max \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^{\mathsf{T}}\mathbf{e}_k\right) \tag{2}$$

$$\sim \frac{1}{2}\left\| \begin{bmatrix} \sqrt{2}\sqrt{\log c - \log \alpha_{k^*}} \\ \mathbf{e}_{k^*} \end{bmatrix} \right\|_2^2, \tag{3}$$

with the normalizing constant is given by $c = \max_k \alpha_k$ [1] to avoid the square root argument becoming negative, and the corresponding Jacobian given by

$$\frac{\partial \mathbf{e}_{\max}}{\partial \mathbf{x}} = \begin{bmatrix} \mathbf{0}^{1\times n_{\mathrm{e}}} \\ \frac{\partial \mathbf{e}_{k^*}}{\partial \mathbf{x}} \end{bmatrix}. \tag{4}$$

1.2 Sum-Mixture

The log-likelihood is computed from the full Gaussian Mixture, and in the sum-mixture case is given by the squared norm of a single scalar term,

$$-\log p_{\text{GM}}(\mathbf{y}|\mathbf{x}) \sim \frac{1}{2}e_{\text{SM}}^2, \tag{5}$$

$$\frac{1}{2}e_{\text{SM}}^2 = \frac{1}{2}\sqrt{2\left(\log c - \log\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)\right)}^2 \tag{6}$$

$$= \frac{1}{2}\sqrt{2\left(\log c + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*} - \log\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*}\right)\right)}^2, \tag{7}$$

where the normalization constant c is given by $\sum_{k=1}^{K}\alpha_k$, and the Jacobian is given by

$$\frac{\partial e_{\text{SM}}}{\partial \mathbf{x}} = \frac{1}{2e_{\text{SM}}}\frac{-2}{\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)}\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)(-\mathbf{e}_k^\mathsf{T})\frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} \tag{8}$$

$$= \frac{1}{e_{\text{SM}}}\frac{\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)\mathbf{e}_k^\mathsf{T}\frac{\partial \mathbf{e}_k}{\partial \mathbf{x}}}{\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)}. \tag{9}$$

1.3 Max-Sum-Mixture

The algebraic value for the Max-Sum-Mixture is the same as for the sum-mixture, but the error term partitioning is different, such that

$$-\log p_{\text{GM}}(\mathbf{y}|\mathbf{x}) \sim \frac{1}{2}\mathbf{e}_{\text{MSM}}^\mathsf{T}\mathbf{e}_{\text{MSM}} \tag{10}$$

$$= \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*} + \frac{1}{2}\sqrt{2\left(\log c - \log\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*}\right)\right)}^2 \tag{11}$$

$$= \frac{1}{2}\left\|\begin{bmatrix}\sqrt{2\left(\log c - \log\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*}\right)\right)} \\ \mathbf{e}_{k^*}\end{bmatrix}\right\|_2^2 \tag{12}$$

$$= \left\|\begin{bmatrix}e_{\text{NL}} \\ \mathbf{e}_{k^*}\end{bmatrix}\right\|_2^2, \tag{13}$$

where $k^* = \arg\max \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)$ is the dominant component and

$$e_{\text{NL}} = \sqrt{2\left(\log c - \log\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*}\right)\right)}, \tag{14}$$

is defined similarly to e_{SM} in (6). The error Jacobian is given by

$$\frac{\partial \mathbf{e}_{\text{MSM}}}{\partial \mathbf{x}} = \begin{bmatrix}\frac{\partial \mathbf{e}_{k^*}}{\partial \mathbf{x}} \\ \frac{\partial e_{\text{NL}}}{\partial \mathbf{x}}\end{bmatrix} \tag{15}$$

and the Jacobian of the nonlinear term $\frac{\partial e_{\text{NL}}}{\partial \mathbf{x}}$ is given by

$$\frac{\partial e_{\text{NL}}}{\partial \mathbf{x}} = \frac{1}{e_{\text{NL}}} \frac{-\sum_{k=1}^{K} \alpha_k \exp(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*}) \left(-\mathbf{e}_k^\mathsf{T}\frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} + \mathbf{e}_{k^*}^\mathsf{T}\frac{\partial \mathbf{e}_{k^*}}{\partial \mathbf{x}}\right)}{\sum_{k=1}^{K} \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k + \frac{1}{2}\mathbf{e}_{k^*}^\mathsf{T}\mathbf{e}_{k^*}\right)}. \tag{16}$$

The normalization constant c is given by

$$c = K \max \alpha_k + \delta, \tag{17}$$

where δ is a damping constant [1] that controls the influence of the nonlinear term (14).

2 Robust Loss: Iterative Reweighted Least Squares

The robust loss formulation for a single factor may be written as

$$J = \rho(f(\mathbf{x})) \tag{1}$$

$$= \rho\left(\frac{1}{2}\mathbf{e}(\mathbf{x})^\mathsf{T}\mathbf{e}(\mathbf{x})\right), \tag{2}$$

where $\mathbf{e} : \mathbb{R}^{n_\text{x}} \to n_\text{e}$ is the error function, $\rho : \mathbb{R} \to \mathbb{R}$ is a robust loss function, and $f(\mathbf{x}) = \frac{1}{2}\mathbf{e}(\mathbf{x})^\mathsf{T}\mathbf{e}(\mathbf{x})$ is a convenient intermediate quantity for the use of the chain rule.

The robustified Gauss-Newton least squares update is derived by considering Newton's method applied to (2), while assuming the Gauss-Newton Hessian approximation is valid for f,

$$\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} \approx \frac{\partial \mathbf{e}}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}}{\partial \mathbf{x}}. \tag{3}$$

The Jacobian of the loss function (2) is then given by

$$\frac{\partial J}{\partial \mathbf{x}} = \frac{\mathrm{d}\rho}{\mathrm{d}f} \frac{\partial f}{\partial \mathbf{x}}, \tag{4}$$

and the Hessian by

$$\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} = \frac{\mathrm{d}\rho}{\mathrm{d}f} \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} + \frac{\partial f}{\partial \mathbf{x}^\mathsf{T}} \frac{\mathrm{d}^2\rho}{\mathrm{d}f^2} \frac{\partial f}{\partial \mathbf{x}}. \tag{5}$$

The intermediate function $f = \frac{1}{2}\mathbf{e}^\mathsf{T}\mathbf{e}$ has Jacobian

$$\frac{\partial f}{\partial \mathbf{x}} = \mathbf{e}^\mathsf{T} \frac{\partial \mathbf{e}}{\partial \mathbf{x}}, \tag{6}$$

and Hessian approximated by (3). The Hessian of the loss is thus given by

$$\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} = \frac{\mathrm{d}\rho}{\mathrm{d}f} \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} + \frac{\partial f}{\partial \mathbf{x}^\mathsf{T}} \frac{\mathrm{d}^2\rho}{\mathrm{d}f^2} \frac{\partial f}{\partial \mathbf{x}} \tag{7}$$

$$= \frac{\mathrm{d}\rho}{\mathrm{d}f} \frac{\partial \mathbf{e}}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}}{\partial \mathbf{x}} + \left(2\mathbf{e}^\mathsf{T}\frac{\partial \mathbf{e}}{\partial \mathbf{x}}\right)^\mathsf{T} \frac{\mathrm{d}^2\rho}{\mathrm{d}f^2} \left(2\mathbf{e}^\mathsf{T}\frac{\partial \mathbf{e}}{\partial \mathbf{x}}\right) \tag{8}$$

$$= \frac{\mathrm{d}\rho}{\mathrm{d}f} \frac{\partial \mathbf{e}}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}}{\partial \mathbf{x}} + \frac{\mathrm{d}^2\rho}{\mathrm{d}f^2} \frac{\partial \mathbf{e}}{\partial \mathbf{x}^\mathsf{T}} \mathbf{e}\mathbf{e}^\mathsf{T} \frac{\partial \mathbf{e}}{\partial \mathbf{x}}. \tag{9}$$

Writing the Newton method update step with the Hessian approximation (9) and the Jacobian (6) yields

$$\left(\frac{\mathrm{d}\rho}{\mathrm{d}f}\frac{\partial \mathbf{e}}{\partial \mathbf{x}^\mathsf{T}}\frac{\partial \mathbf{e}}{\partial \mathbf{x}} + \frac{\mathrm{d}^2\rho}{\mathrm{d}f^2}\frac{\partial \mathbf{e}}{\partial \mathbf{x}^\mathsf{T}}\mathbf{e}\mathbf{e}^\mathsf{T}\frac{\partial \mathbf{e}}{\partial \mathbf{x}}\right)\delta\mathbf{x} = -\frac{\mathrm{d}\rho}{\mathrm{d}f}\mathbf{e}^\mathsf{T}\frac{\partial \mathbf{e}}{\partial \mathbf{x}}, \tag{10}$$

Using only the robust loss function first order term is exactly the iteratively reweighted least squares approach. Some solvers, such as Ceres, also use the second order term, which is called the Triggs correction. Note that this approach does *not* correspond to defininining an error term of the form

$$\tilde{e} = \sqrt{\rho(\mathbf{e}^\mathsf{T}\mathbf{e})}, \tag{11}$$

and using that in Gauss-Newton since that would be ill-conditioned, as the error term size reduces to one. Furthermore, for problems with many different factors, the robust loss ρ terms will all have different values.

3 Hessian Approximation for Sum-Mixtures

Consider an optimization problem consisting of a single factor of the form

$$J = \log c - \log \sum_{k=1^K} \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right) \tag{1}$$

$$= \sqrt{\log c - \log \sum_{k=1^K} \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k\right)}^2, \tag{2}$$

where the square root is present to make it similar to standard error evaluation for Gauss-Newton, and the normalization constant c is present for the square root argument to be positive.

Similarly to the robust loss case (2), the additional nonlinearity imposed by the Gaussian mixture needs to be taken into account for the Hessian approximation. Analogously to the robust loss case, $f_k = \frac{1}{2}\mathbf{e}_k^\mathsf{T}\mathbf{e}_k$ is assumed well-behaved such that

$$\frac{\partial^2 f_k}{\partial \mathbf{x}\partial \mathbf{x}^\mathsf{T}} \approx \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}}\frac{\partial \mathbf{e}_k}{\partial \mathbf{x}}, \tag{3}$$

and the definition for ρ follows as

$$\rho(f_1, \ldots, f_k) = \log c - \log \sum_{k=1}^{K} \alpha_k \exp(-f_k). \tag{4}$$

Results from the robust loss case are unapplicable here due to the additional nonlinearity due to the LogSumExp term and the presence of multiple f_k terms instead of a single one.

Nevertheless, a robustified Hessian approximation for the sum-mixture may be derived. The Jacobian is given by

$$\frac{\partial J}{\partial \mathbf{x}} = \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k}\frac{\partial f_k}{\partial \mathbf{x}}, \tag{5}$$

and the Hessian by

$$\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} = \sum_{k=1}^{K} \left(\frac{\partial \rho}{\partial f_k} \frac{\partial^2 f_k}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} + \frac{\partial f_k}{\partial \mathbf{x}^\mathsf{T}} \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \frac{\partial f_j}{\partial \mathbf{x}} \right). \tag{6}$$

The partial derivatives of ρ are given by

$$\begin{aligned}
\frac{\partial \rho}{\partial f_k} &= -\frac{-\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} && (7)\\
&= \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} && (8)\\
\frac{\partial^2 \rho}{\partial f_j \partial f_k} &= \frac{\delta_{jk}(-\alpha_k \exp(-f_k)) \sum_{i=1}^{K} \alpha_i \exp(-f_i) - \alpha_k \exp(-f_k)(-\alpha_j \exp(-f_j))}{(\sum_{i=1}^{K} \alpha_i \exp(-f_i))^2} && (9)\\
&= \frac{-\delta_{jk}(\alpha_k \exp(-f_k)) \sum_{i=1}^{K} \alpha_i \exp(-f_i) + \alpha_k \alpha_j \exp(-f_k) \exp(-f_j)}{(\sum_{i=1}^{K} \alpha_i \exp(-f_i))^2} && (10)
\end{aligned}$$

and the Jacobian of $f_k = \frac{1}{2} \mathbf{e}_k^\mathsf{T} \mathbf{e}_k$ is given by

$$\frac{\partial f_k}{\partial \mathbf{x}} = \mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}}. \tag{11}$$

The Jacobian of the loss J is then given by (5),

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{x}} &= \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k} \frac{\partial f_k}{\partial \mathbf{x}} && (12)\\
&= \sum_{k=1}^{K} \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} \mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} && (13)\\
&= \sum_{k=1}^{K} \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} \mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}}. && (14)\\
& && (15)
\end{aligned}$$

The Hessian of the loss is given by (6)

$$\begin{aligned}
\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} &= \sum_{k=1}^{K} \left(\frac{\partial \rho}{\partial f_k} \frac{\partial^2 f_k}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} + \frac{\partial f_k}{\partial \mathbf{x}^\mathsf{T}} \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \frac{\partial f_j}{\partial \mathbf{x}} \right) && (16)\\
&= \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} + \left(\mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} \right)^\mathsf{T} \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \left(\mathbf{e}_j^\mathsf{T} \frac{\partial \mathbf{e}_j}{\partial \mathbf{x}} \right) && (17)\\
&= \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} + \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}} \mathbf{e}_k \mathbf{e}_j^\mathsf{T} \frac{\partial \mathbf{e}_j}{\partial \mathbf{x}}. && (18)
\end{aligned}$$

The parallel to the classic result in the robust loss case (10) is aesthetically pleasing.

References

[1] T. Pfeifer, S. Lange, and P. Protzel, "Advancing Mixture Models for Least Squares Optimization," *IEEE Robotics and Automation Letters (RA-L)*, vol. 6, no. 2, pp. 3941–3948, 2021.