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

June 18, 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

$$\frac{\partial \rho}{\partial f_k} = -\frac{-\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} \tag{7}$$

$$= \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} \tag{8}$$

$$\frac{\partial^2 \rho}{\partial f_j \partial f_k} = \frac{\delta_{jk}(-\alpha_k \exp(-f_k)) \sum_{i=1}^{K} \alpha_i \exp(-f_i) - \alpha_k \exp(-f_k)(-\alpha_j \exp(-f_j))}{(\sum_{i=1}^{K} \alpha_i \exp(-f_i))^2} \tag{9}$$

$$= \frac{-\delta_{jk}(\alpha_k \exp(-f_k)) \sum_{i=1}^{K} \alpha_i \exp(-f_i) + \alpha_k \alpha_j \exp(-f_k) \exp(-f_j)}{(\sum_{i=1}^{K} \alpha_i \exp(-f_i))^2} \tag{10}$$

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),

$$\frac{\partial J}{\partial \mathbf{x}} = \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k} \frac{\partial f_k}{\partial \mathbf{x}} \tag{12}$$

$$= \sum_{k=1}^{K} \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} \mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} \tag{13}$$

$$= \sum_{k=1}^{K} \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{K} \alpha_i \exp(-f_i)} \mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}}. \tag{14}$$

$$\tag{15}$$

The Hessian of the loss is given by (6)

$$\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} = \sum_{k=1}^{K} \left(\frac{\partial \rho}{\partial f_k} \frac{\partial^2 f_k}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}} + \frac{\partial f_k}{\partial \mathbf{x}^\mathsf{T}} \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \frac{\partial f_j}{\partial \mathbf{x}} \right) \tag{16}$$

$$= \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} + \left(\mathbf{e}_k^\mathsf{T} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} \right)^\mathsf{T} \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \left(\mathbf{e}_j^\mathsf{T} \frac{\partial \mathbf{e}_j}{\partial \mathbf{x}} \right) \tag{17}$$

$$= \sum_{k=1}^{K} \frac{\partial \rho}{\partial f_k} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}} + \sum_{j=1}^{K} \frac{\partial^2 \rho}{\partial f_j f_k} \frac{\partial \mathbf{e}_k}{\partial \mathbf{x}^\mathsf{T}} \mathbf{e}_k \mathbf{e}_j^\mathsf{T} \frac{\partial \mathbf{e}_j}{\partial \mathbf{x}}. \tag{18}$$

4 Derivation of Normalization Constant

Here we derive the result related to to the lower bound on ΔJ presented in Section IV of the paper. The solver error is given by

$$\mathbf{e}_{\text{solver}}^{\mathsf{T}} = \begin{bmatrix} \mathbf{e}_{\text{solver},1}^{\mathsf{T}} & \sqrt{2(\gamma_{\text{HSM}} + \Delta J)} \end{bmatrix}, \tag{1}$$

where

$$\mathbf{e}_{\text{solver},1}^{\mathsf{T}} = \begin{bmatrix} \sqrt{\frac{\partial \rho}{\partial f_1}}\mathbf{e}_1^{\mathsf{T}} & \dots & \sqrt{\frac{\partial \rho}{\partial f_K}}\mathbf{e}_K^{\mathsf{T}} \end{bmatrix}, \tag{2}$$

and $\frac{\partial \rho}{\partial f_k}$ is defined such that

$$\frac{\partial \rho}{\partial f_k} = \frac{\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^{\mathsf{T}}\mathbf{e}_k\right)}{\sum_{i=1}^{n_k} \alpha_i \exp\left(-\frac{1}{2}\mathbf{e}_i^{\mathsf{T}}\mathbf{e}_i\right)}. \tag{3}$$

The desired error is given by the negative log-likelihood of the GMM as

$$J_{\text{GMM}}(\mathbf{x}) = -\log \sum_{k=1}^{n_k} \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k(\mathbf{x})^{\mathsf{T}}\mathbf{e}_k(\mathbf{x})\right). \tag{4}$$

By setting $\Delta J = J_{\text{GMM}} - \frac{1}{2}\mathbf{e}_{\text{solver},1}^{\mathsf{T}}\mathbf{e}_{\text{solver},1}$, the evaluated cost becomes

$$\begin{aligned} \frac{1}{2}\mathbf{e}_{\text{solver}}^{\mathsf{T}}\mathbf{e}_{\text{solver}} &= \frac{1}{2}\mathbf{e}_{\text{solver},1}^{\mathsf{T}}\mathbf{e}_{\text{solver},1} + \gamma_{\text{HSM}} + \Delta J - \frac{1}{2}\mathbf{e}_{\text{solver},1}^{\mathsf{T}}\mathbf{e}_{\text{solver},1} && (5)\\ &= \gamma_{\text{HSM}} + J_{\text{GMM}}. && (6) \end{aligned}$$

The constant γ_{HSM} is set such that $\gamma_{\text{HSM}} + \Delta J \geq 0$, which requires a lower bound on ΔJ. The lower bound is obtained by first manipulating ΔJ as

$$\begin{aligned} \Delta J &= J_{\text{GMM}} - \frac{1}{2}\mathbf{e}_{\text{solver},1}^{\mathsf{T}}\mathbf{e}_{\text{solver},1} && (7)\\ &= -\log \sum_{k=1}^{n_k} \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k(\mathbf{x})^{\mathsf{T}}\mathbf{e}_k(\mathbf{x})\right) - \frac{1}{2}\mathbf{e}_{\text{solver},1}^{\mathsf{T}}\mathbf{e}_{\text{solver},1} && (8)\\ &= -\log \sum_{k=1}^{n_k} \alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k(\mathbf{x})^{\mathsf{T}}\mathbf{e}_k(\mathbf{x})\right) - \frac{1}{2}\sum_{k=1}^{n_k} \frac{\alpha_k \exp\left(-\frac{1}{2}\mathbf{e}_k^{\mathsf{T}}\mathbf{e}_k\right)}{\sum_{i=1}^{n_k} \alpha_i \exp\left(-\frac{1}{2}\mathbf{e}_i^{\mathsf{T}}\mathbf{e}_i\right)}\mathbf{e}_k^{\mathsf{T}}\mathbf{e}_k. && (9) \end{aligned}$$

Changing variables to $f_k = \frac{1}{2}\mathbf{e}_k(\mathbf{x})^{\mathsf{T}}\mathbf{e}_k(\mathbf{x})$ yields

$$\begin{aligned} \Delta J &= -\log \sum_{k=1}^{n_k} \alpha_k \exp\left(-f_k\right) - \frac{1}{2}\sum_{k=1}^{n_k} \frac{\alpha_k \exp\left(-f_k\right)}{\sum_{i=1}^{n_k} \alpha_i \exp\left(-f_i\right)} 2 f_k && (10)\\ &= -\log \sum_{k=1}^{n_k} \alpha_k \exp\left(-f_k\right) - \sum_{k=1}^{n_k} \frac{\alpha_k \exp\left(-f_k\right)}{\sum_{i=1}^{n_k} \alpha_i \exp\left(-f_i\right)} f_k, && (11) \end{aligned}$$

with $\alpha_k > 0, f_k \geq 0$. Then, exponentiating and taking logarithm of the second term yields

$$\Delta J = -\log \sum_{k=1}^{n_k} \alpha_k \exp(-f_k) - \sum_{k=1}^{n_k} \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} f_k \tag{12}$$

$$= -\log \left(\sum_{k=1}^{n_k} \alpha_k \exp(-f_k) \right) - \log \left(\exp \sum_{k=1}^{n_k} \frac{\alpha_k \exp(-f_k)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} f_k \right) \tag{13}$$

$$= -\log \sum_{k=1}^{n_k} \left(\alpha_k \exp(-f_k) \exp \sum_{j=1}^{n_k} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} f_j \right) \tag{14}$$

$$= -\log \sum_{k=1}^{n_k} \left(\alpha_k \exp \left(-f_k + \sum_{j=1}^{n_k} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} f_j \right) \right) \tag{15}$$

$$= -\log \sum_{k=1}^{n_k} \left(\alpha_k \exp \left(\sum_{j=1}^{n_k} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} (f_j - f_k) \right) \right). \tag{16}$$

The overall expression in (16) is decreasing as a function of the exponent argument S_k,

$$S_k = \sum_{j=1}^{n_k} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} (f_j - f_k). \tag{17}$$

Therefore, an upper bound on S_k is required. In the worst case scenario, $f_j \geq f_k$ for all j in the summation. This can be clarified as follows. Defining $A_{f_j \geq f_k}$ as the set of indices j with $f_j \geq f_k$, as well as the set $A_{f_j < f_k}$ with $f_j < f_k$ with $f_j < f_k$, and splitting S_k up into sums over these two sets yields

$$S_k = \sum_{A_{f_j \geq f_k}} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} (f_j - f_k) + \sum_{A_{f_j < f_k}} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} (f_j - f_k) \tag{18}$$

$$\leq \sum_{A_{f_j \geq f_k}} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} (f_j - f_k), \tag{19}$$

since $f_j - f_k < 0$ in the second term of (18). Thus, proceeding with $f_j - f_k \geq 0$ allows to write

$$S_k = \sum_{j=1}^{n_k} \frac{\alpha_j \exp(-f_j)}{\sum_{i=1}^{n_k} \alpha_i \exp(-f_i)} (f_j - f_k) \tag{20}$$

$$\leq \sum_{j=1}^{n_k} \frac{\alpha_j \exp(-f_j)}{\alpha_j \exp(-f_j) + \alpha_k \exp(-f_k)} (f_j - f_k) \tag{21}$$

$$= \sum_{j=1}^{n_k} \frac{1}{1 + \frac{\alpha_k}{\alpha_j} \exp(f_j - f_k)} (f_j - f_k) \tag{22}$$

$$\leq \sum_{j=1}^{n_k} \frac{1}{\frac{\alpha_k}{\alpha_j} \exp(f_j - f_k)} (f_j - f_k) \tag{23}$$

$$= \frac{1}{\alpha_k} \sum_{j=1}^{n_k} \alpha_j \exp(-(f_j - f_k)) (f_j - f_k). \tag{24}$$

Using the change of variables $t = f_j - f_k$, with $t > 0$, makes explicit that

$$\exp\left(-(f_j - f_k)\right)(f_j - f_k) = \exp\left(-t\right)(t), \tag{25}$$

which has a single maximum at $t = 1$ with $\exp\left(-t\right)(t) = \exp(-1)$. Therefore,

$$S_k \leq \frac{1}{\alpha_k} \sum_{j=1}^{n_k} \alpha_j \exp(-1) \tag{26}$$

$$\leq \frac{1}{\alpha_k} \sum_{j=1}^{n_k} \alpha_j. \tag{27}$$

Therefore,

$$\Delta J = -\log \sum_{k=1}^{n_k} \left(\alpha_k \exp\left(S_k\right)\right) \geq -\log \sum_{k=1}^{n_k} \left(\alpha_k \exp\left(\frac{1}{\alpha_k} \sum_{j=1}^{n_k} \alpha_j\right)\right), \tag{28}$$

and the normalization constant γ_{HSM} may be set to

$$\gamma_{\text{HSM}} = \log \sum_{k=1}^{n_k} \left(\alpha_k \exp\left(\frac{1}{\alpha_k} \sum_{j=1}^{n_k} \alpha_j\right)\right). \tag{29}$$

5 Identities

5.1 Jacobian of vector-scalar product

For $a(\mathbf{x})$ a scalar function of $\mathbf{x}$ and $\mathbf{v}(\mathbf{x})$ a column vector function of $\mathbf{x}$, the Jacobian of their product is obtained as

$$\left.\frac{\partial a\mathbf{v}}{\partial \mathbf{x}}\right|_{i,j} = \frac{\partial}{\partial x_j} a v_i \tag{1}$$

$$= \frac{\partial a}{\partial x_j} v_i + a \frac{\partial v_i}{\partial x_j} \tag{2}$$

$$\frac{\partial a\mathbf{v}}{\partial \mathbf{x}} = \mathbf{v}\frac{\partial a}{\partial \mathbf{x}} + a\frac{\partial \mathbf{v}}{\partial \mathbf{x}} \tag{3}$$

5.2 Hessian of a function through Jacobian of gradient

The Hessian $\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}}$ of a function $J(x)$ is given by a matrix with entries

$$\left.\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^\mathsf{T}}\right|_{i,j} = \frac{\partial J}{\partial x_i \partial x_j}. \tag{4}$$

The gradient $\frac{\partial J}{\partial \mathbf{x}}^\mathsf{T}$ of J has entries given by

$$\left.\frac{\partial J}{\partial \mathbf{x}}^\mathsf{T}\right|_{i} = \frac{\partial J}{\partial x_i}, \tag{5}$$

and the Jacobian of the gradient is given by

$$\frac{\partial}{\partial \mathbf{x}} \frac{\partial J}{\partial \mathbf{x}}^{\mathsf{T}} \Big|_{i,j} = \frac{\partial J}{\partial x_j \partial x_i}. \tag{6}$$

Therefore,

$$\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^{\mathsf{T}}} = \left(\frac{\partial}{\partial \mathbf{x}} \frac{\partial J}{\partial \mathbf{x}}^{\mathsf{T}} \right)^{\mathsf{T}}. \tag{7}$$

5.3 2D wedge operator multiplication flip

Consider a general vector $\mathbf{v} \in \mathbb{R}^2$ such that

$$\delta\xi^{\phi\wedge}\mathbf{v} = \begin{bmatrix} 0 & -\delta\xi^{\phi} \\ \delta\xi^{\phi} & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{8}$$

$$= \begin{bmatrix} -\delta\xi^{\phi} v_2 \\ \delta\xi^{\phi} v_1 \end{bmatrix} \tag{9}$$

$$= \begin{bmatrix} -v_2 \\ v_1 \end{bmatrix} \delta\xi^{\phi} \tag{10}$$

$$= \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{v} \delta\xi^{\phi}. \tag{11}$$

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.