Improving Normalization With James-Stein Estimator
Supplementary Materials

Seyedalireza Khoshsirat   Chandra Kambhamettu
Video/Image Modeling and Synthesis (VIMS) Lab, University of Delaware
{alireza, chandrak}@udel.edu

Our JSNorm layers require the same computational resources as the original normalization layers. However, the current implementations of normalization layers employ an optimized mathematical expression tailored for computing derivatives during the backpropagation phase. This optimization is achieved through manual derivation and operates independently of the automatic differentiation mechanisms that are integral to many machine learning frameworks.

In this supplementary material, we present the optimized expression for the derivative of our method. Specifically, we provide the details for our batch normalization layer, while noting that similar expressions can be used for our layer normalization. We maintain the same notation as in the main text, with the addition of the symbol $\mathcal{S}$ denoting the sum of the squares of a vector, or $\|\cdot\|_2^2$. Therefore, $\mathcal{S}_{\mu_\mathcal{B}} = \|\mu_\mathcal{B}\|_2^2$.

1. Chain Rule Expansions

Below, we present the partial derivatives derived using the chain rule, starting from the final output y_i. The main three outputs are the partial derivatives of the loss function ℓ with respect to γ, β, and x_i, given an input $x \in \mathbb{R}^{n \times c}$.

$$\frac{\partial \ell}{\partial \gamma} = \frac{\partial \ell}{\partial y_i} \cdot \frac{\partial y_i}{\partial \gamma} \tag{1}$$

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial \ell}{\partial y_i} \cdot \frac{\partial y_i}{\partial \beta} \tag{2}$$

$$\frac{\partial \ell}{\partial x_i} = \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial \ell}{\partial \mu_\mathcal{B}} \cdot \frac{\partial \mu_\mathcal{B}}{\partial x_i} + \frac{\partial \ell}{\partial \sigma_\mathcal{B}^2} \cdot \frac{\partial \sigma_\mathcal{B}^2}{\partial x_i} \tag{3}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_\mathcal{B}} = {} & \frac{\partial \ell}{\partial \mu_\mathcal{JS}} \cdot \frac{\partial \mu_\mathcal{JS}}{\partial \mu_\mathcal{B}} + \frac{\partial \ell}{\partial \mathcal{S}_{\mu_\mathcal{B}}} \cdot \frac{\partial \mathcal{S}_{\mu_\mathcal{B}}}{\partial \mu_\mathcal{B}} + \frac{\partial \ell}{\partial \sigma_{\mu_\mathcal{B}}^2} \cdot \frac{\partial \sigma_{\mu_\mathcal{B}}^2}{\partial \mu_\mathcal{B}} \\ & + \frac{\partial \ell}{\partial \mu_{\mu_\mathcal{B}}} \cdot \frac{\partial \mu_{\mu_\mathcal{B}}}{\partial \mu_\mathcal{B}} + \frac{\partial \ell}{\partial \sigma_\mathcal{B}^2} \cdot \frac{\partial \sigma_\mathcal{B}^2}{\partial \mu_\mathcal{B}}\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma_\mathcal{B}^2} = {} & \frac{\partial \ell}{\partial \sigma_\mathcal{JS}^2} \cdot \frac{\partial \sigma_\mathcal{JS}^2}{\partial \sigma_\mathcal{B}^2} + \frac{\partial \ell}{\partial \mathcal{S}_{\sigma_\mathcal{B}^2}} \cdot \frac{\partial \mathcal{S}_{\sigma_\mathcal{B}^2}}{\partial \sigma_\mathcal{B}^2} + \frac{\partial \ell}{\partial \sigma_{\sigma_\mathcal{B}^2}^2} \cdot \frac{\partial \sigma_{\sigma_\mathcal{B}^2}^2}{\partial \sigma_\mathcal{B}^2} \\ & + \frac{\partial \ell}{\partial \mu_{\sigma_\mathcal{B}^2}} \cdot \frac{\partial \mu_{\sigma_\mathcal{B}^2}}{\partial \sigma_\mathcal{B}^2}\end{aligned} \tag{5}$$

$$\frac{\partial \ell}{\partial \mathcal{S}_{\mu_\mathcal{B}}} = \frac{\partial \ell}{\partial \mu_\mathcal{JS}} \cdot \frac{\partial \mu_\mathcal{JS}}{\partial \mathcal{S}_{\mu_\mathcal{B}}} \tag{6}$$

$$\frac{\partial \ell}{\partial \mathcal{S}_{\sigma_\mathcal{B}^2}} = \frac{\partial \ell}{\partial \sigma_\mathcal{JS}^2} \cdot \frac{\partial \sigma_\mathcal{JS}^2}{\partial \mathcal{S}_{\sigma_\mathcal{B}^2}} \tag{7}$$

$$\frac{\partial \ell}{\partial \mu_{\mu_\mathcal{B}}} = \frac{\partial \ell}{\partial \sigma_{\mu_\mathcal{B}}^2} \cdot \frac{\partial \sigma_{\mu_\mathcal{B}}^2}{\partial \mu_{\mu_\mathcal{B}}} \tag{8}$$

$$\frac{\partial \ell}{\partial \sigma_{\mu_\mathcal{B}}^2} = \frac{\partial \ell}{\partial \mu_\mathcal{JS}} \cdot \frac{\partial \mu_\mathcal{JS}}{\partial \sigma_{\mu_\mathcal{B}}^2} \tag{9}$$

$$\frac{\partial \ell}{\partial \mu_{\sigma_\mathcal{B}^2}} = \frac{\partial \ell}{\partial \sigma_{\sigma_\mathcal{B}^2}^2} \cdot \frac{\partial \sigma_{\sigma_\mathcal{B}^2}^2}{\partial \mu_{\sigma_\mathcal{B}^2}} \tag{10}$$

$$\frac{\partial \ell}{\partial \sigma_{\sigma_\mathcal{B}^2}^2} = \frac{\partial \ell}{\partial \sigma_\mathcal{JS}^2} \cdot \frac{\partial \sigma_\mathcal{JS}^2}{\partial \sigma_{\sigma_\mathcal{B}^2}^2} \tag{11}$$

$$\frac{\partial \ell}{\partial \mu_\mathcal{JS}} = \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu_\mathcal{JS}} \tag{12}$$

$$\frac{\partial \ell}{\partial \sigma_\mathcal{JS}^2} = \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma_\mathcal{JS}^2} \tag{13}$$

2. Partial Derivatives

The actual derivatives for the partials are calculated as follows:

$$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial y_i} \cdot \hat{x}_i \tag{14}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial y_i} \tag{15}$$

$$\frac{\partial \ell}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial y_i} \cdot \gamma \tag{16}$$

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma_{\mathcal{J}\mathcal{S}}^2 + \epsilon}} \tag{17}$$

$$\frac{\partial \hat{x}_i}{\partial \mu_{\mathcal{J}\mathcal{S}}} = \frac{-1}{\sqrt{\sigma_{\mathcal{J}\mathcal{S}}^2 + \epsilon}} \tag{18}$$

$$\frac{\partial \hat{x}_i}{\partial \sigma_{\mathcal{J}\mathcal{S}}^2} = -0.5 \sum_{i=1}^{n} (x_i - \mu_{\mathcal{J}\mathcal{S}}) \cdot (\sigma_{\mathcal{J}\mathcal{S}}^2 + \epsilon)^{-1.5} \tag{19}$$

$$\frac{\partial \mu_{\mathcal{J}\mathcal{S}}}{\partial \mu_{\mathcal{B}}} = 1 - \frac{(c-2)\sigma_{\mu_{\mathcal{B}}}^2}{\mathcal{S}_{\mu_{\mathcal{B}}}} \tag{20}$$

$$\frac{\partial \sigma_{\mathcal{J}\mathcal{S}}^2}{\partial \sigma_{\mathcal{B}}^2} = 1 - \frac{(c-2)\sigma_{\sigma_{\mathcal{B}}^2}^2}{\mathcal{S}_{\sigma_{\mathcal{B}}^2}} \tag{21}$$

$$\frac{\partial \mu_{\mathcal{J}\mathcal{S}}}{\partial \mathcal{S}_{\mu_{\mathcal{B}}}} = \frac{\mu_{\mathcal{B}} \cdot (c-2) \cdot \sigma_{\mu_{\mathcal{B}}}^2}{\mathcal{S}_{\mu_{\mathcal{B}}}^2} \tag{22}$$

$$\frac{\partial \sigma_{\mathcal{J}\mathcal{S}}^2}{\partial \mathcal{S}_{\sigma_{\mathcal{B}}^2}} = \frac{\sigma_{\mathcal{B}}^2 \cdot (c-2) \cdot \sigma_{\sigma_{\mathcal{B}}^2}^2}{\mathcal{S}_{\sigma_{\mathcal{B}}^2}^2} \tag{23}$$

$$\frac{\partial \mathcal{S}_{\mu_{\mathcal{B}}}}{\partial \mu_{\mathcal{B}}} = 2 \cdot \mu_{\mathcal{B}} \tag{24}$$

$$\frac{\partial \mathcal{S}_{\sigma_{\mathcal{B}}^2}}{\partial \sigma_{\mathcal{B}}^2} = 2 \cdot \sigma_{\mathcal{B}}^2 \tag{25}$$

$$\frac{\partial \mu_{\mathcal{J}\mathcal{S}}}{\partial \sigma_{\mu_{\mathcal{B}}}^2} = \frac{-\mu_{\mathcal{B}} \cdot (c-2)}{\mathcal{S}_{\mu_{\mathcal{B}}}} \tag{26}$$

$$\frac{\partial \sigma_{\mathcal{J}\mathcal{S}}^2}{\partial \sigma_{\sigma_{\mathcal{B}}^2}^2} = \frac{-\sigma_{\mathcal{B}}^2 \cdot (c-2)}{\mathcal{S}_{\sigma_{\mathcal{B}}^2}} \tag{27}$$

$$\frac{\partial \mu_{\mathcal{B}}}{\partial x_i} = \frac{1}{n} \tag{28}$$

$$\frac{\partial \mu_{\mu_{\mathcal{B}}}}{\partial \mu_{\mathcal{B}}} = \frac{1}{c} \tag{29}$$

$$\frac{\partial \mu_{\sigma_{\mathcal{B}}^2}}{\partial \sigma_{\mathcal{B}}^2} = \frac{1}{c} \tag{30}$$

$$\frac{\partial \sigma_{\mathcal{B}}^2}{\partial x_i} = \frac{2(x_i - \mu_{\mathcal{B}})}{n} \tag{31}$$

$$\frac{\partial \sigma_{\sigma_{\mathcal{B}}^2}^2}{\partial \sigma_{\mathcal{B}}^2} = \frac{2(\sigma_{\mathcal{B}}^2 - \mu_{\sigma_{\mathcal{B}}^2})}{c} \tag{32}$$

$$\frac{\partial \sigma_{\mu_{\mathcal{B}}}^2}{\partial \mu_{\mathcal{B}}} = \frac{2(\mu_{\mathcal{B}} - \mu_{\mu_{\mathcal{B}}})}{c} \tag{33}$$

$$\frac{\partial \sigma_{\mathcal{B}}^2}{\partial \mu_{\mathcal{B}}} = \frac{1}{n} \sum_{i=1}^{n} -2 \cdot (x_i - \mu_{\mathcal{B}}) \tag{34}$$

$$\frac{\partial \sigma_{\sigma_{\mathcal{B}}^2}^2}{\partial \mu_{\sigma_{\mathcal{B}}^2}} = \frac{1}{c} \sum_{i=1}^{c} -2 \cdot (\sigma_{\mathcal{B}i}^2 - \mu_{\sigma_{\mathcal{B}}^2}) \tag{35}$$

$$\frac{\partial \sigma_{\mu_{\mathcal{B}}}^2}{\partial \mu_{\mu_{\mathcal{B}}}} = \frac{1}{c} \sum_{i=1}^{c} -2 \cdot (\mu_{\mathcal{B}i} - \mu_{\mu_{\mathcal{B}}}) \tag{36}$$

3. Simplifying Expressions

Multiple expressions can be simplified as follows:

$$\begin{aligned} \frac{\partial \sigma_{\mathcal{B}}^2}{\partial \mu_{\mathcal{B}}} &= \frac{1}{n} \sum_{i=1}^{n} -2 \cdot (x_i - \mu_{\mathcal{B}}) \\ &= -2 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \mu_{\mathcal{B}} \right) \\ &= -2 \cdot \left(\mu_{\mathcal{B}} - \frac{n \cdot \mu_{\mathcal{B}}}{n} \right) \\ &= -2 \cdot (\mu_{\mathcal{B}} - \mu_{\mathcal{B}}) \\ &= 0 \end{aligned} \tag{37}$$

Similarly:

$$\frac{\partial \sigma_{\sigma_{\mathcal{B}}^2}^2}{\partial \mu_{\sigma_{\mathcal{B}}^2}} = 0 \tag{38}$$

$$\frac{\partial \sigma_{\mu_{\mathcal{B}}}^2}{\partial \mu_{\mu_{\mathcal{B}}}} = 0 \tag{39}$$

Therefore:

$$\frac{\partial \ell}{\partial \mu_{\sigma_{\mathcal{B}}^2}} = 0 \tag{40}$$

$$\frac{\partial \ell}{\partial \mu_{\mu_{\mathcal{B}}}} = 0 \tag{41}$$

By utilizing the above results, the partial derivatives $\frac{\partial \ell}{\partial \mu_{\mathcal{B}}}$ and $\frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2}$ can be re-written as:

$$\frac{\partial \ell}{\partial \mu_{\mathcal{B}}} = \frac{\partial \ell}{\partial \mu_{\mathcal{J}\mathcal{S}}} \cdot \frac{\partial \mu_{\mathcal{J}\mathcal{S}}}{\partial \mu_{\mathcal{B}}} + \frac{\partial \ell}{\partial \mathcal{S}_{\mu_{\mathcal{B}}}} \cdot \frac{\partial \mathcal{S}_{\mu_{\mathcal{B}}}}{\partial \mu_{\mathcal{B}}} + \frac{\partial \ell}{\partial \sigma_{\mu_{\mathcal{B}}}^2} \cdot \frac{\partial \sigma_{\mu_{\mathcal{B}}}^2}{\partial \mu_{\mathcal{B}}} \tag{42}$$

$$\frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} = \frac{\partial \ell}{\partial \sigma_{\mathcal{J}\mathcal{S}}^2} \cdot \frac{\partial \sigma_{\mathcal{J}\mathcal{S}}^2}{\partial \sigma_{\mathcal{B}}^2} + \frac{\partial \ell}{\partial \mathcal{S}_{\sigma_{\mathcal{B}}^2}} \cdot \frac{\partial \mathcal{S}_{\sigma_{\mathcal{B}}^2}}{\partial \sigma_{\mathcal{B}}^2} + \frac{\partial \ell}{\partial \sigma_{\sigma_{\mathcal{B}}^2}^2} \cdot \frac{\partial \sigma_{\sigma_{\mathcal{B}}^2}^2}{\partial \sigma_{\mathcal{B}}^2} \tag{43}$$