Controlled Average, Variance, and Extrema (CAVE) Functions: Theory, Implementation, and Applications

1 Introduction

We introduce a new group of differentiable functions that can control the average, variance, and extrema (CAVE). With a CAVE function, one can have tight control over certain statistics, namely

$$l = \min(\text{CAVE}(\mathbf{X})) \tag{1}$$
$$h = \max(\text{CAVE}(\mathbf{X})) \tag{2}$$
$$\mu = \mathbb{E}[\text{CAVE}(\mathbf{X})] \tag{3}$$
$$\nu = \text{var}(\text{CAVE}(\mathbf{X})) \tag{4}$$

where l, h, μ, and ν are all chosen by the user. While it is easy to derive functions that control each individually, deriving functions for any combination is not trivial. Here we'll address how each combination can be achieved.

2 Pre-derived Functions

Here we discuss methods that are used to control certain combinations of the minimum, maximum, mean, and variance. Assume that any function $f(\cdot)$ operates element-wise over the input tensor.

2.1 Controlling l and h

The minimum and maximum values are easy to control. ReLU and Softplus for example are unbounded in one direction: any input will have an output in the range $[0, \infty)$. We use functions like these to control the minimum or maximum. More generally, we can denote $f_U(\cdot)$ as any semi-infinitely bound function. Without loss of generality, we'll assume that f_U is lower-bounded by value U. Any function with a finite maximum we can transform by taking the negative.

To limit the output range to our liking, we use f_U as a base function, then appropriately shift it vertically. For setting minimum or maximum values, the CAVE functions are

$$\text{CAVE}\left(\mathbf{X};\ l\right) = f_U(\mathbf{X}) - U + l \tag{5}$$

$$\text{CAVE}\left(\mathbf{X};\ h\right) = -\left(f_U(\mathbf{X}) - U\right) + h \tag{6}$$

respectively. The new ranges are as desired.

If one wanted to control both the minimum and maximum together, the sigmoid function for example is a common activation function that does just that. Again, we generalize these strictly bounded functions as $f_B\left(\cdot\right)$ with lower bound B_l and upper bound B_h. The CAVE function controlling the upper and lower bound is then

$$\text{CAVE}\left(\mathbf{X};\ l, h\right) = \left(h - l\right)\frac{f_B\left(\mathbf{X}\right) - B_l}{B_h - B_l} + l. \tag{7}$$

The output range is scaled and shifted to the desired range.

2.2 Controlling μ and ν

We'll use properties of the mean and variance to control these properties. For the mean, we know that

$$\mathbb{E}\left[\mathbf{X} + b\right] = \mathbb{E}\left[\mathbf{X}\right] + b. \tag{8}$$

It's easy to show that the CAVE function for mean adjustment is then

$$\text{CAVE}\left(\mathbf{X};\ \mu\right) = \mathbf{X} - \mathbb{E}\left[\mathbf{X}\right] + \mu. \tag{9}$$

The variance is likewise easy to derive. We use the property

$$\text{var}\left(a\,\mathbf{X}\right) = a^2\,\text{var}\left(\mathbf{X}\right) \tag{10}$$

to then derive the CAVE equation

$$\text{CAVE}\left(\mathbf{X};\ \nu\right) = \sqrt{\frac{\nu}{\text{var}\left(\mathbf{X}\right)}}\,\mathbf{X}. \tag{11}$$

Controlling both mean and variance, we can combine the properties to aid in that CAVE derivation:

$$\mathbb{E}\left[a\mathbf{X} + b\right] = a\mathbb{E}\left[\mathbf{X}\right] + b \tag{12}$$

$$\text{var}\left(a\mathbf{X} + b\right) = a^2\,\text{var}\left(\mathbf{X}\right). \tag{13}$$

Using both properties, the CAVE function to control mean and variance is

$$\text{CAVE}\left(\mathbf{X};\ \mu, \nu\right) = \sqrt{\frac{\nu}{\text{var}\left(\mathbf{X}\right)}}\left(\mathbf{X} - \mathbb{E}\left[\mathbf{X}\right]\right) + \mu. \tag{14}$$

3 Controlling the Remaining Combinations

We previously showed how we can limit the range, but it does not take into account the resulting mean and variance. For example, after applying a sigmoid function, it might be impossible to then apply a linear transform that achieves the desired mean and variance that is guaranteed to be within the desired output range. In order to combat this, we instead first apply a linear transform on the data such that after applying a range limiting function, we achieve all the desired statistics:

$$\text{CAVE}\left(\mathbf{X};\ l, \mu, \nu\right) = f_U\left(a\mathbf{X} + b\right) + l \tag{15}$$

$$\text{CAVE}\left(\mathbf{X};\ h, \mu, \nu\right) = -f_U\left(a\mathbf{X} + b\right) + h \tag{16}$$

$$\text{CAVE}\left(\mathbf{X};\ l, h, \mu, \nu\right) = (h - l)\ f_B\left(a\mathbf{X} + b\right) + l \tag{17}$$

where a and b are scalars such that the correct statistics are achieved. Note that the mean and variance cannot be simplified in a useful manner in order to solve for a and b. Thus, we solve it numerically.

3.1 CAVE Loss Function

We want to minimize the loss between the calculated mean and/or variance and the desired mean and/or variance by adjusting linear weights a and b. The losses we define are

$$\mathcal{L}\left(\mathbf{X}, a, b\right) = \mathcal{L}_\mu(\mathbf{X}, a, b) + \mathcal{L}_\nu(\mathbf{X}, a, b) \tag{18}$$

$$\mathcal{L}_\mu\left(\mathbf{X}, a, b\right) = \left\|E_\mu\left(\mathbf{X}, a, b\right)\right\|_2^2 \tag{19}$$

$$\mathcal{L}_\nu\left(\mathbf{X}, a, b\right) = \left\|E_\nu\left(\mathbf{X}, a, b\right)\right\|_2^2. \tag{20}$$

where

$$E_\mu\left(\mathbf{X}, a, b\right) = \mathbb{E}\left[f\left(a\,\mathbf{X}_j + b\right)\right] - \mu \tag{21}$$

$$= \frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right) - \mu \tag{22}$$

$$E_\nu\left(\mathbf{X}, a, b\right) = \text{var}\left(f\left(a\,\mathbf{X}_j + b\right)\right) - \nu \tag{23}$$

$$= \frac{1}{N}\sum_{j=1}^{N} f^2\left(a\,\mathbf{X}_j + b\right) - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)^2 - \nu \tag{24}$$

and $f\left(\cdot\right)$ is a range-limited function. Our loss function $\mathcal{L}$ minimizes the l_2 norm of both errors.

Since there are only two variables we are optimizing, we turn to a combination of gradient descent and Newton's method.

3.2 CAVE Optimization

Here we derive how the CAVE function works.

3.2.1 CAVE Loss

We can solve for the gradient analytically by

$$\nabla \mathcal{L}(a,b) = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a} \\ \frac{\partial \mathcal{L}}{\partial b} \end{bmatrix} \tag{25}$$

$$= \begin{bmatrix} \frac{\partial \mathcal{L}_\mu}{\partial a} + \frac{\partial \mathcal{L}_\nu}{\partial a} \\ \frac{\partial \mathcal{L}_\mu}{\partial b} + \frac{\partial \mathcal{L}_\nu}{\partial b} \end{bmatrix} \tag{26}$$

where

$$\frac{\partial \mathcal{L}_\mu}{\partial a} = 2E_\mu \frac{\partial E_\mu}{\partial a} \tag{27}$$

$$\frac{\partial \mathcal{L}_\nu}{\partial a} = 2E_\nu \frac{\partial E_\nu}{\partial a} \tag{28}$$

$$\frac{\partial \mathcal{L}_\mu}{\partial b} = 2E_\mu \frac{\partial E_\mu}{\partial b} \tag{29}$$

$$\frac{\partial \mathcal{L}_\nu}{\partial b} = 2E_\nu \frac{\partial E_\nu}{\partial b}. \tag{30}$$

The Hessian we find by

$$\nabla^2 \mathcal{L}(a,b) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial a^2} & \frac{\partial^2 \mathcal{L}}{\partial a \partial b} \\ \frac{\partial^2 \mathcal{L}}{\partial a \partial b} & \frac{\partial^2 \mathcal{L}}{\partial b^2} \end{bmatrix} \tag{31}$$

$$= \begin{bmatrix} \frac{\partial^2 \mathcal{L}_\mu}{\partial a^2} + \frac{\partial^2 \mathcal{L}_\nu}{\partial a^2} & \frac{\partial^2 \mathcal{L}_\mu}{\partial a \partial b} + \frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial b} \\ \frac{\partial^2 \mathcal{L}_\mu}{\partial a \partial b} + \frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial b} & \frac{\partial^2 \mathcal{L}_\mu}{\partial b^2} + \frac{\partial^2 \mathcal{L}_\nu}{\partial b^2} \end{bmatrix} \tag{32}$$

where

$$\frac{\partial^2 \mathcal{L}_\mu}{\partial a^2} = 2\left(\left(\frac{\partial E_\mu}{\partial a}\right)^2 + E_\mu \frac{\partial^2 E_\mu}{\partial a^2}\right) \tag{33}$$

$$\frac{\partial^2 \mathcal{L}_\nu}{\partial a^2} = 2\left(\left(\frac{\partial E_\nu}{\partial a}\right)^2 + E_\nu \frac{\partial^2 E_\nu}{\partial a^2}\right) \tag{34}$$

$$\frac{\partial^2 \mathcal{L}_\mu}{\partial a \partial b} = 2\left(\frac{\partial E_\mu}{\partial a}\frac{\partial E_\mu}{\partial b} + E_\mu \frac{\partial^2 E_\mu}{\partial a \partial b}\right) \tag{35}$$

$$\frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial b} = 2\left(\frac{\partial E_\nu}{\partial a}\frac{\partial E_\nu}{\partial b} + E_\nu \frac{\partial^2 E_\nu}{\partial a \partial b}\right) \tag{36}$$

$$\frac{\partial^2 \mathcal{L}_\mu}{\partial b^2} = 2\left(\left(\frac{\partial E_\mu}{\partial b}\right)^2 + E_\mu \frac{\partial^2 E_\mu}{\partial b^2}\right) \tag{37}$$

$$\frac{\partial^2 \mathcal{L}_\nu}{\partial b^2} = 2\left(\left(\frac{\partial E_\nu}{\partial b}\right)^2 + E_\nu \frac{\partial^2 E_\nu}{\partial b^2}\right). \tag{38}$$

3.2.2 CAVE Error

The mean error function and its derivatives are found as

$$E_\mu = \frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right) - \mu \tag{39}$$

$$\frac{\partial E_\mu}{\partial a} = \frac{1}{N}\sum_{j=1}^{N} \frac{\partial f}{\partial a} \tag{40}$$

$$\frac{\partial E_\mu}{\partial b} = \frac{1}{N}\sum_{j=1}^{N} \frac{\partial f}{\partial b} \tag{41}$$

$$\frac{\partial^2 E_\mu}{\partial a^2} = \frac{1}{N}\sum_{j=1}^{N} \frac{\partial^2 f}{\partial a^2} \tag{42}$$

$$\frac{\partial^2 E_\mu}{\partial a \partial b} = \frac{1}{N}\sum_{j=1}^{N} \frac{\partial^2 f}{\partial a \partial b} \tag{43}$$

$$\frac{\partial^2 E_\mu}{\partial b^2} = \frac{1}{N}\sum_{j=1}^{N} \frac{\partial^2 f}{\partial b^2}. \tag{44}$$

The variance error function and its derivatives are also found as

$$E_\nu = \frac{1}{N}\sum_{j=1}^{N} f^2 - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)^2 - \nu \tag{45}$$

$$\frac{\partial E_\nu}{\partial a} = 2\left\{\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\frac{\partial f}{\partial a} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial E_\mu}{\partial a}\right\} \tag{46}$$

$$\frac{\partial E_\nu}{\partial b} = 2\left\{\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\frac{\partial f}{\partial b} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial E_\mu}{\partial b}\right\} \tag{47}$$

$$\begin{aligned}\frac{\partial^2 E_\nu}{\partial a^2} = 2\Bigg\{&\frac{1}{N}\sum_{j=1}^{N}\left[\left(\frac{\partial f}{\partial a}\right)^2 + f\left(a\,\mathbf{X}_j + b\right)\frac{\partial^2 f}{\partial a^2}\right] - \cdots \\ &\left(\frac{\partial E_\mu}{\partial a}\right)^2 - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^2 E_\mu}{\partial a^2}\Bigg\}\end{aligned} \tag{48}$$

$$\begin{aligned}\frac{\partial^2 E_\nu}{\partial a \partial b} = 2\Bigg\{&\frac{1}{N}\sum_{j=1}^{N}\left[\frac{\partial f}{\partial a}\frac{\partial f}{\partial b} + f\left(a\,\mathbf{X}_j + b\right)\frac{\partial^2 f}{\partial a \partial b}\right] - \cdots \\ &\frac{\partial E_\mu}{\partial a}\frac{\partial E_\mu}{\partial b} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^2 E_\mu}{\partial a \partial b}\Bigg\}\end{aligned} \tag{49}$$

$$\begin{aligned}\frac{\partial^2 E_\nu}{\partial b^2} = 2\Bigg\{&\frac{1}{N}\sum_{j=1}^{N}\left[\left(\frac{\partial f}{\partial b}\right)^2 + f\left(a\,\mathbf{X}_j + b\right)\frac{\partial^2 f}{\partial b^2}\right] - \cdots \\ &\left(\frac{\partial E_\mu}{\partial b}\right)^2 - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^2 E_\mu}{\partial b^2}\Bigg\}.\end{aligned} \tag{50}$$

3.2.3 CAVE Base Function

Lastly, we have the derivatives of the base function f:

$$\frac{\partial f}{\partial a} = \frac{\partial}{\partial a}\left[f\left(a\,\mathbf{X}_j + b\right)\right] = f'\left(a\,\mathbf{X}_j + b\right)\mathbf{X}_j \tag{51}$$

$$\frac{\partial f}{\partial b} = \frac{\partial}{\partial b}\left[f\left(a\,\mathbf{X}_j + b\right)\right] = f'\left(a\,\mathbf{X}_j + b\right) \tag{52}$$

$$\frac{\partial^2 f}{\partial a^2} = \frac{\partial^2}{\partial a^2}\left[f\left(a\,\mathbf{X}_j + b\right)\right] = f''\left(a\,\mathbf{X}_j + b\right)\mathbf{X}_j^2 \tag{53}$$

$$\frac{\partial^2 f}{\partial a \partial b} = \frac{\partial^2}{\partial a \partial b}\left[f\left(a\,\mathbf{X}_j + b\right)\right] = f''\left(a\,\mathbf{X}_j + b\right)\mathbf{X}_j \tag{54}$$

$$\frac{\partial^2 f}{\partial b^2} = \frac{\partial^2}{\partial b^2}\left[f\left(a\,\mathbf{X}_j + b\right)\right] = f''\left(a\,\mathbf{X}_j + b\right). \tag{55}$$

Given all these values, we can minimize the loss function.

3.3 CAVE Updates

The update function for gradient descent is

$$\begin{bmatrix} a \\ b \end{bmatrix}_{k+1} = \begin{bmatrix} a \\ b \end{bmatrix}_{k} - \eta\,\nabla\mathcal{L}\left(\begin{bmatrix} a \\ b \end{bmatrix}_{k}\right) \tag{56}$$

and the update function for Newton's method is

$$\begin{bmatrix} a \\ b \end{bmatrix}_{k+1} = \begin{bmatrix} a \\ b \end{bmatrix}_{k} - \eta\left(\nabla^2\mathcal{L}\left(\begin{bmatrix} a \\ b \end{bmatrix}_{k}\right)\right)^{-1}\nabla\mathcal{L}\left(\begin{bmatrix} a \\ b \end{bmatrix}_{k}\right). \tag{57}$$

The above two equations optimize both the mean and variance jointly. If only the mean is specified, the gradient descent step is

$$b_{k+1} = b_k - \eta\,\frac{\partial \mathcal{L}_\mu}{\partial b} \tag{58}$$

and the Newton step is

$$b_{k+1} = b_k - \eta\left(\frac{\partial^2 \mathcal{L}_\mu}{\partial b^2}\right)^{-1}\frac{\partial \mathcal{L}_\mu}{\partial b}. \tag{59}$$

Similarly, if only the variance is specified, the gradient descent step is

$$a_{k+1} = a_k - \eta\,\frac{\partial \mathcal{L}_\nu}{\partial a} \tag{60}$$

and the Newton step is

$$a_{k+1} = a_k - \eta\left(\frac{\partial^2 \mathcal{L}_\nu}{\partial a^2}\right)^{-1}\frac{\partial \mathcal{L}_\nu}{\partial a}. \tag{61}$$

To run CAVE functions, we take K_g gradient descent steps to approach the minimum followed by K_n Newton steps to rapidly approach the minimum. Once a and b are found, the data is finally transformed and fed through the base function as $f\left(t\left(\mathbf{X}_j, a, b\right)\right)$.

4 The Gradient of Gradient Descent and Newton's Method

This function, while effective, is memory-intensive when tracking the gradient: many operations require saving the input tensors for backpropagation. Here, we derive the gradient of one step of gradient descent and Newton's method with respect to $\mathbf{X}_i$ given only the input data. The previous section derived how to optimize linear transform variables a and b, but we'll derive how the steps to approach the minimum affect the input data.

Since it is impractical to unravel gradient descent and Newton's method, we'll find the gradient of $\nabla\mathcal{L}\left(\mathbf{X}, a, b\right)$ and $\left(\nabla^2\mathcal{L}\left(\mathbf{X}, a, b\right)\right)^{-1}\nabla\mathcal{L}\left(\mathbf{X}, a, b\right)$ both w.r.t. $\mathbf{X}_i$. These are the values multiplied by the learning rate.

4.1 Gradient of Newton's Method w.r.t. $\mathbf{X}_i$

We can expand the gradient step by

$$\left(\nabla^2\mathcal{L}\left(\mathbf{X}, a, b\right)\right)^{-1}\nabla\mathcal{L}\left(\mathbf{X}, a, b\right) = \frac{1}{D}\begin{bmatrix}N_a\\N_b\end{bmatrix}. \tag{62}$$

where

$$N_a\left(\mathbf{X}, a, b\right) = \frac{\partial\mathcal{L}}{\partial a}\frac{\partial^2\mathcal{L}}{\partial b^2} - \frac{\partial\mathcal{L}}{\partial b}\frac{\partial^2\mathcal{L}}{\partial a\partial b} \tag{63}$$

$$N_b\left(\mathbf{X}, a, b\right) = \frac{\partial\mathcal{L}}{\partial b}\frac{\partial^2\mathcal{L}}{\partial a^2} - \frac{\partial\mathcal{L}}{\partial a}\frac{\partial^2\mathcal{L}}{\partial a\partial b} \tag{64}$$

$$D\left(\mathbf{X}, a, b\right) = \frac{\partial^2\mathcal{L}}{\partial a^2}\frac{\partial^2\mathcal{L}}{\partial b^2} - \left(\frac{\partial^2\mathcal{L}}{\partial a\partial b}\right)^2. \tag{65}$$

We need to find the derivative of each entry w.r.t. each $\mathbf{X}_i$:

$$\frac{\partial}{\partial\mathbf{X}_i}\left[\frac{N_a\left(\mathbf{X}, a, b\right)}{D\left(\mathbf{X}, a, b\right)}\right] = \frac{D\frac{\partial N_a}{\partial\mathbf{X}_i} - N_a\frac{\partial D}{\partial\mathbf{X}_i}}{D^2} \tag{66}$$

$$\frac{\partial}{\partial\mathbf{X}_i}\left[\frac{N_b\left(\mathbf{X}, a, b\right)}{D\left(\mathbf{X}, a, b\right)}\right] = \frac{D\frac{\partial N_b}{\partial\mathbf{X}_i} - N_b\frac{\partial D}{\partial\mathbf{X}_i}}{D^2} \tag{67}$$

where

$$\frac{\partial N_a}{\partial\mathbf{X}_i} = \frac{\partial^2\mathcal{L}}{\partial a\partial\mathbf{X}_i}\cdot\frac{\partial^2\mathcal{L}}{\partial b^2} + \frac{\partial\mathcal{L}}{\partial a}\cdot\frac{\partial^3\mathcal{L}}{\partial b^2\partial\mathbf{X}_i} - \frac{\partial^2\mathcal{L}}{\partial b\partial\mathbf{X}_i}\cdot\frac{\partial^2\mathcal{L}}{\partial a\partial b} - \frac{\partial\mathcal{L}}{\partial b}\cdot\frac{\partial^3\mathcal{L}}{\partial a\partial b\partial\mathbf{X}_i} \tag{68}$$

$$\frac{\partial N_b}{\partial\mathbf{X}_i} = \frac{\partial^2\mathcal{L}}{\partial b\partial\mathbf{X}_i}\cdot\frac{\partial^2\mathcal{L}}{\partial a^2} + \frac{\partial\mathcal{L}}{\partial b}\cdot\frac{\partial^3\mathcal{L}}{\partial a^2\partial\mathbf{X}_i} - \frac{\partial^2\mathcal{L}}{\partial a\partial\mathbf{X}_i}\cdot\frac{\partial^2\mathcal{L}}{\partial a\partial b} - \frac{\partial\mathcal{L}}{\partial a}\cdot\frac{\partial^3\mathcal{L}}{\partial a\partial b\partial\mathbf{X}_i} \tag{69}$$

$$\frac{\partial D}{\partial\mathbf{X}_i} = \frac{\partial^3\mathcal{L}}{\partial a^2\partial\mathbf{X}_i}\cdot\frac{\partial^2\mathcal{L}}{\partial b^2} + \frac{\partial^2\mathcal{L}}{\partial a^2}\cdot\frac{\partial^3\mathcal{L}}{\partial b^2\partial\mathbf{X}_i} - 2\frac{\partial^2\mathcal{L}}{\partial a\partial b}\cdot\frac{\partial^3\mathcal{L}}{\partial a\partial b\partial\mathbf{X}_i}. \tag{70}$$

Many of these derivatives have been found in the previous section, but any derivative w.r.t. $\mathbf{X}_i$ has yet to be found. These derivatives of $\mathcal{L}$ are decomposed into their $\mathcal{L}_\mu$ and $\mathcal{L}_\nu$ counterparts as

$$\frac{\partial^2 \mathcal{L}}{\partial a \partial \mathbf{X}_i} = \frac{\partial^2 \mathcal{L}_\mu}{\partial a \partial \mathbf{X}_i} + \frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial \mathbf{X}_i} \tag{71}$$

$$\frac{\partial^2 \mathcal{L}}{\partial b \partial \mathbf{X}_i} = \frac{\partial^2 \mathcal{L}_\mu}{\partial b \partial \mathbf{X}_i} + \frac{\partial^2 \mathcal{L}_\nu}{\partial b \partial \mathbf{X}_i} \tag{72}$$

$$\frac{\partial^3 \mathcal{L}}{\partial a^2 \partial \mathbf{X}_i} = \frac{\partial^3 \mathcal{L}_\mu}{\partial a^2 \partial \mathbf{X}_i} + \frac{\partial^3 \mathcal{L}_\nu}{\partial a^2 \partial \mathbf{X}_i} \tag{73}$$

$$\frac{\partial^3 \mathcal{L}}{\partial a \partial b \partial \mathbf{X}_i} = \frac{\partial^3 \mathcal{L}_\mu}{\partial a \partial b \partial \mathbf{X}_i} + \frac{\partial^3 \mathcal{L}_\nu}{\partial a \partial b \partial \mathbf{X}_i} \tag{74}$$

$$\frac{\partial^3 \mathcal{L}}{\partial b^2 \partial \mathbf{X}_i} = \frac{\partial^3 \mathcal{L}_\mu}{\partial b^2 \partial \mathbf{X}_i} + \frac{\partial^3 \mathcal{L}_\nu}{\partial b^2 \partial \mathbf{X}_i}. \tag{75}$$

4.1.1 Gradient of Losses w.r.t. $\mathbf{X}_i$

Using previous definitions, we can find the above loss derivatives. The second order derivatives are

$$\frac{\partial^2 \mathcal{L}_\mu}{\partial a \partial \mathbf{X}_i} = 2 \left(\frac{\partial E_\mu}{\partial \mathbf{X}_i} \cdot \frac{\partial E_\mu}{\partial a} + E_\mu \cdot \frac{\partial^2 E_\mu}{\partial a \partial \mathbf{X}_i} \right) \tag{76}$$

$$\frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial \mathbf{X}_i} = 2 \left(\frac{\partial E_\nu}{\partial \mathbf{X}_i} \cdot \frac{\partial E_\nu}{\partial a} + E_\nu \cdot \frac{\partial^2 E_\nu}{\partial a \partial \mathbf{X}_i} \right) \tag{77}$$

$$\frac{\partial^2 \mathcal{L}_\mu}{\partial b \partial \mathbf{X}_i} = 2 \left(\frac{\partial E_\mu}{\partial \mathbf{X}_i} \cdot \frac{\partial E_\mu}{\partial b} + E_\mu \cdot \frac{\partial^2 E_\mu}{\partial b \partial \mathbf{X}_i} \right) \tag{78}$$

$$\frac{\partial^2 \mathcal{L}_\nu}{\partial b \partial \mathbf{X}_i} = 2 \left(\frac{\partial E_\nu}{\partial \mathbf{X}_i} \cdot \frac{\partial E_\nu}{\partial b} + E_\nu \cdot \frac{\partial^2 E_\nu}{\partial b \partial \mathbf{X}_i} \right) \tag{79}$$

and the third derivatives are

$$\frac{\partial^3 \mathcal{L}_\mu}{\partial a^2 \partial \mathbf{X}_i} = 2\left(2\frac{\partial E_\mu}{\partial a}\cdot\frac{\partial^2 E_\mu}{\partial a \partial \mathbf{X}_i} + \frac{\partial E_\mu}{\partial \mathbf{X}_i}\cdot\frac{\partial^2 E_\mu}{\partial a^2} + E_\mu \cdot \frac{\partial^3 E_\mu}{\partial a^2 \partial \mathbf{X}_i}\right) \tag{80}$$

$$\frac{\partial^3 \mathcal{L}_\nu}{\partial a^2 \partial \mathbf{X}_i} = 2\left(2\frac{\partial E_\nu}{\partial a}\cdot\frac{\partial^2 E_\nu}{\partial a \partial \mathbf{X}_i} + \frac{\partial E_\nu}{\partial \mathbf{X}_i}\cdot\frac{\partial^2 E_\nu}{\partial a^2} + E_\nu \cdot \frac{\partial^3 E_\nu}{\partial a^2 \partial \mathbf{X}_i}\right) \tag{81}$$

$$\frac{\partial^3 \mathcal{L}_\mu}{\partial a \partial b \partial \mathbf{X}_i} = 2\left(\frac{\partial^2 E_\mu}{\partial a \partial \mathbf{X}_i}\cdot\frac{\partial E_\mu}{\partial b} + \frac{\partial E_\mu}{\partial a}\cdot\frac{\partial^2 E_\mu}{\partial b \partial \mathbf{X}_i} + \frac{\partial E_\mu}{\partial \mathbf{X}_i}\cdot\frac{\partial^2 E_\mu}{\partial a \partial b} + E_\mu \cdot \frac{\partial^3 E_\mu}{\partial a \partial b \partial \mathbf{X}_i}\right) \tag{82}$$

$$\frac{\partial^3 \mathcal{L}_\nu}{\partial a \partial b \partial \mathbf{X}_i} = 2\left(\frac{\partial^2 E_\nu}{\partial a \partial \mathbf{X}_i}\cdot\frac{\partial E_\nu}{\partial b} + \frac{\partial E_\nu}{\partial a}\cdot\frac{\partial^2 E_\nu}{\partial b \partial \mathbf{X}_i} + \frac{\partial E_\nu}{\partial \mathbf{X}_i}\cdot\frac{\partial^2 E_\nu}{\partial a \partial b} + E_\nu \cdot \frac{\partial^3 E_\nu}{\partial a \partial b \partial \mathbf{X}_i}\right) \tag{83}$$

$$\frac{\partial^3 \mathcal{L}_\mu}{\partial b^2 \partial \mathbf{X}_i} = 2\left(2\frac{\partial E_\mu}{\partial b}\cdot\frac{\partial^2 E_\mu}{\partial b \partial \mathbf{X}_i} + \frac{\partial E_\mu}{\partial \mathbf{X}_i}\cdot\frac{\partial^2 E_\mu}{\partial b^2} + E_\mu \cdot \frac{\partial^3 E_\mu}{\partial b^2 \partial \mathbf{X}_i}\right) \tag{84}$$

$$\frac{\partial^3 \mathcal{L}_\nu}{\partial b^2 \partial \mathbf{X}_i} = 2\left(2\frac{\partial E_\nu}{\partial b}\cdot\frac{\partial^2 E_\nu}{\partial b \partial \mathbf{X}_i} + \frac{\partial E_\nu}{\partial \mathbf{X}_i}\cdot\frac{\partial^2 E_\nu}{\partial b^2} + E_\nu \cdot \frac{\partial^3 E_\nu}{\partial b^2 \partial \mathbf{X}_i}\right). \tag{85}$$

4.1.2 Gradient of Errors w.r.t. $\mathbf{X}_i$

Similarly, we can find the derivatives of the error functions. The mean error functions are

$$\frac{\partial E_\mu}{\partial \mathbf{X}_i} = \frac{1}{N}\frac{\partial f}{\partial \mathbf{X}_i} \tag{86}$$

$$\frac{\partial^2 E_\mu}{\partial a \partial \mathbf{X}_i} = \frac{1}{N}\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i} \tag{87}$$

$$\frac{\partial^2 E_\mu}{\partial b \partial \mathbf{X}_i} = \frac{1}{N}\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i} \tag{88}$$

$$\frac{\partial^3 E_\mu}{\partial a^2 \partial \mathbf{X}_i} = \frac{1}{N}\frac{\partial^3 f}{\partial a^2 \partial \mathbf{X}_i} \tag{89}$$

$$\frac{\partial^3 E_\mu}{\partial a \partial b \partial \mathbf{X}_i} = \frac{1}{N}\frac{\partial^3 f}{\partial a \partial b \partial \mathbf{X}_i} \tag{90}$$

$$\frac{\partial^3 E_\mu}{\partial b^2 \partial \mathbf{X}_i} = \frac{1}{N}\frac{\partial^3 f}{\partial b^2 \partial \mathbf{X}_i} \tag{91}$$

and the variance error functions are

$$\frac{\partial E_\nu}{\partial \mathbf{X}_i} = \frac{2}{N}\left\{\frac{\partial f}{\partial \mathbf{X}_i} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial f}{\partial \mathbf{X}_i}\right\} \tag{92}$$

$$\begin{aligned}\frac{\partial^2 E_\nu}{\partial a \partial \mathbf{X}_i} = \frac{2}{N}\Bigg\{&\frac{\partial f}{\partial a}\frac{\partial f}{\partial \mathbf{X}_i} + f\left(a\,\mathbf{X}_i + b\right)\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i} - \cdots \\ &\frac{\partial E_\mu}{\partial a}\frac{\partial f}{\partial \mathbf{X}_i} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i}\Bigg\}\end{aligned} \tag{93}$$

$$\begin{aligned}\frac{\partial^2 E_\nu}{\partial b \partial \mathbf{X}_i} = \frac{2}{N}\Bigg\{&\frac{\partial f}{\partial b}\frac{\partial f}{\partial \mathbf{X}_i} + f\left(a\,\mathbf{X}_i + b\right)\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i} - \cdots \\ &\frac{\partial E_\mu}{\partial b}\frac{\partial f}{\partial \mathbf{X}_i} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i}\Bigg\}\end{aligned} \tag{94}$$

$$\begin{aligned}\frac{\partial^3 E_\nu}{\partial a^2 \partial \mathbf{X}_i} = \frac{2}{N}\Bigg\{&2\frac{\partial f}{\partial a}\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i} + \frac{\partial f}{\partial \mathbf{X}_i}\frac{\partial^2 f}{\partial a^2} + f\left(a\,\mathbf{X}_j + b\right)\frac{\partial^3 f}{\partial a^2 \partial \mathbf{X}_i} - \cdots \\ &2\frac{\partial E_\mu}{\partial a}\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i} - \frac{\partial^2 E_\mu}{\partial a^2}\frac{\partial f}{\partial \mathbf{X}_i} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^3 f}{\partial a^2 \partial \mathbf{X}_i}\Bigg\}\end{aligned} \tag{95}$$

$$\begin{aligned}\frac{\partial^3 E_\nu}{\partial a \partial b \partial \mathbf{X}_i} = \frac{2}{N}\Bigg\{&\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i}\frac{\partial f}{\partial b} + \frac{\partial^2 f}{\partial b \partial \mathbf{X}_i}\frac{\partial f}{\partial a} + \frac{\partial f}{\partial \mathbf{X}_i}\frac{\partial^2 f}{\partial a \partial b} + \frac{\partial^3 f}{\partial a \partial b \partial \mathbf{X}_i} f\left(a\,\mathbf{X}_j + b\right) - \cdots \\ &\frac{\partial E_\mu}{\partial a}\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i} - \frac{\partial E_\mu}{\partial b}\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i} - \frac{\partial^2 E_\mu}{\partial a \partial b}\frac{\partial f}{\partial \mathbf{X}_i} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^3 f}{\partial a \partial b \partial \mathbf{X}_i}\Bigg\}\end{aligned} \tag{96}$$

$$\begin{aligned}\frac{\partial^3 E_\nu}{\partial b^2 \partial \mathbf{X}_i} = \frac{2}{N}\Bigg\{&2\frac{\partial f}{\partial b}\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i} + \frac{\partial f}{\partial \mathbf{X}_i}\frac{\partial^2 f}{\partial b^2} + f\left(a\,\mathbf{X}_j + b\right)\frac{\partial^3 f}{\partial b^2 \partial \mathbf{X}_i} - \cdots \\ &2\frac{\partial E_\mu}{\partial b}\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i} - \frac{\partial^2 E_\mu}{\partial b^2}\frac{\partial f}{\partial \mathbf{X}_i} - \left(\frac{1}{N}\sum_{j=1}^{N} f\left(a\,\mathbf{X}_j + b\right)\right)\frac{\partial^3 f}{\partial b^2 \partial \mathbf{X}_i}\Bigg\}.\end{aligned} \tag{97}$$

4.1.3 Gradient of CAVE Base Functions w.r.t. $\mathbf{X}_i$

Lastly, we can find the additional function derivatives as

$$\frac{\partial f}{\partial \mathbf{X}_i} = \frac{\partial}{\partial \mathbf{X}_i}\left[f\left(a\,\mathbf{X}_i + b\right)\right] = f'\left(a\,\mathbf{X}_i + b\right)\cdot a \tag{98}$$

$$\frac{\partial^2 f}{\partial a \partial \mathbf{X}_i} = \frac{\partial^2}{\partial a \partial \mathbf{X}_i}\left[f\left(a\,\mathbf{X}_i + b\right)\right] = f''\left(a\,\mathbf{X}_i + b\right)\cdot a\,\mathbf{X}_i + f'\left(a\,\mathbf{X}_i + b\right) \tag{99}$$

$$\frac{\partial^2 f}{\partial b \partial \mathbf{X}_i} = \frac{\partial^2}{\partial b \partial \mathbf{X}_i}\left[f\left(a\,\mathbf{X}_i + b\right)\right] = f''\left(a\,\mathbf{X}_i + b\right)\cdot a \tag{100}$$

$$\frac{\partial^3 f}{\partial a^2 \partial \mathbf{X}_i} = \frac{\partial^3}{\partial a^2 \partial \mathbf{X}_i}\left[f\left(a\,\mathbf{X}_i + b\right)\right] = f'''\left(a\,\mathbf{X}_i + b\right)\cdot a\,\mathbf{X}_i^2 + f''\left(a\,\mathbf{X}_i + b\right)\cdot 2\,\mathbf{X}_i \tag{101}$$

$$\frac{\partial^3 f}{\partial a \partial b \partial \mathbf{X}_i} = \frac{\partial^3}{\partial a^2 \partial \mathbf{X}_i}\left[f\left(a\,\mathbf{X}_i + b\right)\right] = f'''\left(a\,\mathbf{X}_i + b\right)\cdot a\,\mathbf{X}_i + f''\left(a\,\mathbf{X}_i + b\right) \tag{102}$$

$$\frac{\partial^3 f}{\partial b^2 \partial \mathbf{X}_i} = \frac{\partial^3}{\partial a^2 \partial \mathbf{X}_i}\left[f\left(a\,\mathbf{X}_i + b\right)\right] = f'''\left(a\,\mathbf{X}_i + b\right)\cdot a. \tag{103}$$

All of these equations constitute the gradient of Newton's method.

4.2 Remaining Gradients

We derived the formula for the gradient of Newton's method for two variables, but there are still five other cases. We finish here with Newton's method with one variable, and gradient descent with both one and two variables.

Gradient descent only requires us to know some combination of $\frac{\partial^2 \mathcal{L}}{\partial a \partial \mathbf{X}_i}$, $\frac{\partial^2 \mathcal{L}}{\partial b \partial \mathbf{X}_i}$, $\frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial \mathbf{X}_i}$, and $\frac{\partial^2 \mathcal{L}_\mu}{\partial b \partial \mathbf{X}_i}$, all of which were found in the previous section, so we're done with gradient descent. Newton's method requires a little more effort. Recall that Newton's update for variance optimization is

$$\left(\frac{\partial \mathcal{L}_\nu}{\partial a}\right)\left(\frac{\partial^2 \mathcal{L}_\nu}{\partial a^2}\right)^{-1}. \tag{104}$$

Using the quotient rule, the gradient of Newton's method for variance optimization is

$$\left(\frac{\partial^2 \mathcal{L}_\nu}{\partial a^2}\frac{\partial^2 \mathcal{L}_\nu}{\partial a \partial \mathbf{X}_i} - \frac{\partial \mathcal{L}_\nu}{\partial a}\frac{\partial^3 \mathcal{L}_\nu}{\partial a^2 \partial \mathbf{X}_i}\right)\left(\frac{\partial^2 \mathcal{L}_\nu}{\partial a^2}\right)^{-2}. \tag{105}$$

Similarly for the mean, we have the optimization step as

$$\left(\frac{\partial \mathcal{L}_\mu}{\partial b}\right)\left(\frac{\partial^2 \mathcal{L}_\mu}{\partial b^2}\right)^{-1} \tag{106}$$

and the gradient as

$$\left(\frac{\partial^2 \mathcal{L}_\mu}{\partial b^2}\frac{\partial^2 \mathcal{L}_\mu}{\partial b \partial \mathbf{X}_i} - \frac{\partial \mathcal{L}_\mu}{\partial b}\frac{\partial^3 \mathcal{L}_\mu}{\partial b^2 \partial \mathbf{X}_i}\right)\left(\frac{\partial^2 \mathcal{L}_\mu}{\partial b^2}\right)^{-2}. \tag{107}$$