Node perturbation in vanilla deep networks

Peter Latham

February 25, 2019

1 Setup

Our goal is to compare node perturbation to stochastic gradient descent (SGD). We'll do this in the context of a vanilla feedforward network. We'll start with a network with added noise, because we'll need it for node perturbation,

$$\mathbf{x}^{k+1} = \phi(\mathbf{h}^k + \boldsymbol{\xi}^k) \tag{1a}$$

$$\mathbf{h}^k = \mathbf{w}^k \cdot \mathbf{x}^k \,. \tag{1b}$$

Here everything in bold is a vector or matrix, the nonlinearity ϕ is pointwise, k labels layer (note that $\mathbf{w}^k$ is the weight from layer k to layer $k+1$), and $\boldsymbol{\xi}^k$ is a zero mean, uncorrelated Gaussian random variable with variance σ^2,

$$\langle \boldsymbol{\xi}^k \boldsymbol{\xi}^l \rangle = \sigma^2 \, \mathbf{I}^{kl} \tag{2}$$

where $\mathbf{I}^{kl}$ is the identity matrix if $k = l$ and 0 otherwise,

$$\mathbf{I}^{kl} = \begin{cases} \text{identity matrix with the dimension of layer } k & k = l \\ 0 & k \neq l \,. \end{cases} \tag{3}$$

We'll take bias into account by setting x_0^k to 1 and not updating that variable. With this convention, w_{i0}^k is the bias for unit i in layer l. We'll use $\mathcal{L}$ to denote the loss; it depends on the weights, $\mathbf{w} \equiv (\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^L)$, noise, $\boldsymbol{\xi} \equiv (\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, ..., \boldsymbol{\xi}^L)$, and input, which we'll denote $\mathbf{x}_{\text{in}}$; we thus write

$$\text{Loss} = \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi}). \tag{4}$$

This is the loss on a single input. The true loss, for which we'll use $\overline{\mathcal{L}}(\mathbf{w})$, is the noise-free loss averaged over all the input examples,

$$\overline{\mathcal{L}}(\mathbf{w}) \equiv \langle \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0) \rangle \tag{5}$$

where here and in what follow, angle brackets without a subscript indicate an average over the inputs.

What we want to know is how this quantity changes under SGD and node perturbation. For that we need to define the true weight update, $\Delta \mathbf{w}_{\text{true}}^k$, which is given by

$$\Delta \mathbf{w}_{\text{true}}^k \equiv -\eta \, \frac{\partial \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k} \tag{6}$$

where η is the learning rate. The SGD weight update is the standard one,

$$\Delta \mathbf{w}_{\text{sgd}}^k \equiv -\eta \, \frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial \mathbf{w}^k} \,. \tag{7}$$

And the node perturbation update rule is

$$\Delta \mathbf{w}^k_{\text{np}} \equiv -\frac{\eta}{\sigma^2}\big[\mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi}) - \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)\big]\, \boldsymbol{\xi}^k \mathbf{x}^k\,. \tag{8}$$

In the latter equation, $\mathbf{x}^k$ is the activity in the network when $\boldsymbol{\xi} = 0$. So the procedure for node perturbation goes like this: draw a sample, $\boldsymbol{\xi}$, from the noise distribution; run the network with that noise using Eq. (1); compute the loss; do the same thing for the network without noise (but with the same input, $\mathbf{x}_{\text{in}}$); again compute the loss; subtract the two losses; divide by σ^2; then multiply by $\boldsymbol{\xi}^k \mathbf{x}^k$.

In both cases the quantity of interest is the change in the loss, which we'll call $\Delta\overline{\mathcal{L}}(\mathbf{w}, \Delta\mathbf{w})$, is

$$\Delta\overline{\mathcal{L}}(\mathbf{w}, \Delta\mathbf{w}) \equiv \overline{\mathcal{L}}(\mathbf{w} + \Delta\mathbf{w}) - \overline{\mathcal{L}}(\mathbf{w})\,. \tag{9}$$

We'll compute this quantity, to second order in $\Delta\mathbf{w}$, for $\Delta\mathbf{w} = \Delta\mathbf{w}_{\text{sgd}}$ and $\Delta\mathbf{w} = \Delta\mathbf{w}_{\text{np}}$, then average over input, $\mathbf{x}_{\text{in}}$, and, for node perturbation, noise, $\boldsymbol{\xi}$.

2 Node perturbation

Before getting to the loss, we need to cast the node perturbation weight update into a more intuitive form. For that we'll Taylor expand the loss around zero noise,

$$\mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi}) \approx \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0) + \sum_m \left.\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^m}\right|_{\boldsymbol{\xi}=0} \cdot \boldsymbol{\xi}^m\,. \tag{10}$$

Taylor expanding in high dimensions seems like a bad idea, because the number of terms explodes as you go to higher order. Turns out it's OK, though; see Sec. 5.

Note that

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^m}\right|_{\boldsymbol{\xi}=0} = \frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial \mathbf{h}^m} \tag{11}$$

where the first equality follows from Eq. (1). This suggests the definition

$$\boldsymbol{\delta}^m \equiv \frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial \mathbf{h}^m}\,, \tag{12}$$

which is the standard backprop error signal. (I think; I can never remember. But it should be.) Inserting the last two expressions into Eq. (10), and then inserting that into Eq. (8), we arrive at

$$\Delta \mathbf{w}^k_{\text{np}} \approx -\frac{\eta}{\sigma^2}\sum_m \boldsymbol{\delta}^m \cdot \boldsymbol{\xi}^m \boldsymbol{\xi}^k \mathbf{x}^k\,. \tag{13}$$

Averaging over the noise, and using Eq. (2), the node perturbation weight update becomes

$$\langle \Delta \mathbf{w}^k_{\text{np}} \rangle_{\boldsymbol{\xi}} = -\eta\, \boldsymbol{\delta}^k \mathbf{x}^k\,. \tag{14}$$

As an aside, combining the definition of $\Delta \mathbf{w}_{\text{sgd}}$ with Eq. (12), we see that

$$\Delta \mathbf{w}^k_{\text{sgd}} = -\eta\, \boldsymbol{\delta}^k \mathbf{x}^k \,. \tag{15}$$

Consequently,

$$\langle \Delta \mathbf{w}^k_{\text{np}} \rangle_{\boldsymbol{\xi}} = \Delta \mathbf{w}^k_{\text{sgd}} \,, \tag{16}$$

a result that will come in handy later.

Below (see Sec. 4), we compute the next higher order correction to the node perturbation update rule; that computation yields

$$\langle \Delta \mathbf{w}_{\text{np}} \rangle \approx -\eta \left[\boldsymbol{\delta}^k \mathbf{x}^k + 3\sigma^2 \sum_{mi} \frac{\partial^2 \boldsymbol{\delta}^k}{\partial h_i^{m\,2}} \, \mathbf{x}^k \right] . \tag{17}$$

The number of terms in the sum over m and i is equal to the number of units in the network. Thus, if we want the noise to remain small as we go to larger networks, σ^2 should scale inversely with the number of units. In what follows, we'll assume that σ^2 is small enough that we can use Eq. (13).

3 Change in loss

We are now in a position to compute $\Delta \overline{\mathcal{L}}(\mathbf{w}, \Delta \mathbf{w})$. To second order, it is given by

$$\Delta \overline{\mathcal{L}}(\mathbf{w}, \Delta \mathbf{w}) = \sum_k \frac{\partial \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k} \cdot \Delta \mathbf{w}^k + \frac{1}{2} \sum_{kl} \Delta \mathbf{w}^k \cdot \frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} \cdot \Delta \mathbf{w}^l \,. \tag{18}$$

The "=" sign should really be "≈", but we should just interpret the equal sign as meaning equal through second order. (Alternatively, if we think of the Hessian as a suitably defined average, it is equal; see Sec. 5.)

The dot products are tricky; they involve sums over both components of the weight matrices. But ambiguity can be avoided. As is easy to show from Eq. (1),

$$\frac{\partial \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k} = \langle \boldsymbol{\delta}^k \, \mathbf{x}^k \rangle \tag{19}$$

where we used Eq. (12) for $\boldsymbol{\delta}^k$, and, recall, angle brackets indicate an average over input examples, $\mathbf{x}_{\text{in}}$. Similarly, we may write

$$\frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} = \langle \boldsymbol{\Delta}^{kl} \, \mathbf{x}^k \mathbf{x}^l \rangle \,. \tag{20}$$

where

$$\boldsymbol{\Delta}^{kl} \equiv \frac{\partial^2 \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial \mathbf{h}^k \partial \mathbf{h}^l} \,. \tag{21}$$

Inserting Eqs. (19) and (20) into Eq. (18), and being creative about colors, we have

$$\Delta\overline{\mathcal{L}}(\mathbf{w}, \Delta\mathbf{w}) = \sum_k \langle \boldsymbol{\delta}^k \cdot \Delta\mathbf{w}^k \cdot \mathbf{x}^k \rangle + \frac{1}{2} \sum_{kl} \langle \mathbf{x}^k \cdot \Delta\mathbf{w}^{kT} \cdot \boldsymbol{\Delta}^{kl} \cdot \Delta\mathbf{w}^l \cdot \mathbf{x}^l \rangle \tag{22}$$

where T means transpose and the red angle brackets tell us to average over only red quantities.

The first observation is that for both SGD and node perturbation, the average value of $\Delta\mathbf{w}^k$ is $\Delta\mathbf{w}^k_{\text{true}}$. That follows from Eq. (16) and the definition of the true weight update, Eq. (6). Thus, for both SGD and node perturbation,

$$\langle \boldsymbol{\delta}^k \cdot \Delta\mathbf{w}^k \cdot \mathbf{x}^k \rangle = \langle \boldsymbol{\delta}^k \cdot \Delta\mathbf{w}^k_{\text{true}} \cdot \mathbf{x}^k \rangle = -\eta \sum_k \langle\langle \boldsymbol{\delta}^k \cdot \boldsymbol{\delta}^k \, \mathbf{x}^k \cdot \mathbf{x}^k \rangle\rangle \, . \tag{23}$$

The second order term does depend on whether we're using SGD or node perturbation. We'll start with SGD, as it's easier. Using Eq. (15),

$$\langle\langle \mathbf{x}^k \cdot \Delta\mathbf{w}^{kT}_{\text{sgd}} \cdot \boldsymbol{\Delta}^{kl} \cdot \Delta\mathbf{w}^l_{\text{sgd}} \cdot \mathbf{x}^l \rangle\rangle = \eta^2 \langle\langle \mathbf{x}^k \cdot \mathbf{x}^k \, \boldsymbol{\delta}^k \cdot \boldsymbol{\Delta}^{kl} \cdot \boldsymbol{\delta}^l \, \mathbf{x}^l \cdot \mathbf{x}^l \rangle\rangle \, . \tag{24}$$

Node perturbation is only slightly harder. Using Eq. (13),

$$\langle\langle \mathbf{x}^k \cdot \Delta\mathbf{w}^{kT}_{\text{np}} \cdot \boldsymbol{\Delta}^{kl} \cdot \Delta\mathbf{w}^l_{\text{np}} \cdot \mathbf{x}^l \rangle\rangle_{\boldsymbol{\xi}} = \frac{\eta^2}{\sigma^4} \sum_{mn} \langle\langle \mathbf{x}^k \cdot \mathbf{x}^k \, \boldsymbol{\delta}^m \cdot \boldsymbol{\xi}^m \boldsymbol{\xi}^k \cdot \boldsymbol{\Delta}^{kl} \cdot \boldsymbol{\xi}^l \boldsymbol{\xi}^n \cdot \boldsymbol{\delta}^n \, \mathbf{x}^l \cdot \mathbf{x}^l \rangle\rangle_{\boldsymbol{\xi}} \, . \tag{25}$$

The average over the noise is straightforward,

$$\begin{aligned} \langle\langle \mathbf{x}^k \cdot \Delta\mathbf{w}^{kT}_{\text{np}} \cdot \boldsymbol{\Delta}^{kl} \cdot \Delta\mathbf{w}^l_{\text{np}} \cdot \mathbf{x}^l \rangle\rangle_{\boldsymbol{\xi}} &= 2\eta^2 \langle \mathbf{x}^k \cdot \mathbf{x}^k \boldsymbol{\delta}^k \cdot \boldsymbol{\Delta}^{kl} \cdot \boldsymbol{\delta}^l \, \mathbf{x}^l \cdot \mathbf{x}^l \rangle \\ &\quad + \eta^2 \sum_m \langle \mathbf{x}^k \cdot \mathbf{x}^k \, \mathbf{x}^l \cdot \mathbf{x}^l \boldsymbol{\delta}^m \cdot \boldsymbol{\delta}^m \text{trace}\{\boldsymbol{\Delta}^{kl} \cdot \mathbf{I}^{lk}\} \rangle \, . \end{aligned} \tag{26}$$

Putting this all together, for SGD we have

$$\langle \Delta\overline{\mathcal{L}}(\mathbf{w}, \Delta\mathbf{w}_{\text{sgd}}) \rangle = -\eta \sum_k \langle\langle \boldsymbol{\delta}^k \cdot \boldsymbol{\delta}^k \, \mathbf{x}^k \cdot \mathbf{x}^k \rangle\rangle + \frac{\eta^2}{2} \sum_{kl} \langle\langle \mathbf{x}^k \cdot \mathbf{x}^k \, \boldsymbol{\delta}^k \cdot \boldsymbol{\Delta}^{kl} \cdot \boldsymbol{\delta}^l \, \mathbf{x}^l \cdot \mathbf{x}^l \rangle\rangle \tag{27}$$

and for node perturbation,

$$\begin{aligned} \langle\langle \Delta\overline{\mathcal{L}}(\mathbf{w}, \Delta\mathbf{w}_{\text{np}}) \rangle\rangle_{\boldsymbol{\xi}} &= \langle \Delta\overline{\mathcal{L}}(\mathbf{w}, \Delta\mathbf{w}_{\text{sgd}}) \rangle + \frac{\eta^2}{2} \sum_{kl} \langle\langle \mathbf{x}^k \cdot \mathbf{x}^k \, \boldsymbol{\delta}^k \cdot \boldsymbol{\Delta}^{kl} \cdot \boldsymbol{\delta}^l \, \mathbf{x}^l \cdot \mathbf{x}^l \rangle\rangle \\ &\quad + \frac{\eta^2}{2} \sum_{mk} \langle\langle (\mathbf{x}^k \cdot \mathbf{x}^k)^2 \boldsymbol{\delta}^m \cdot \boldsymbol{\delta}^m \text{trace}\{\boldsymbol{\Delta}^{kk}\} \rangle\rangle \, . \end{aligned} \tag{28}$$

The difference between SGD and node perturbation is mainly in the term involving the trace of $\boldsymbol{\Delta}^{kk}$. If you count terms in the sum over m and n versus the sum over k and l, however, they're the same. The difference is that the sum over k and l contains off-diagonal elements of $\boldsymbol{\Delta}^{kl}$, which may be either positive or negative (in principle; we don't really know), while the trace term has contributions only from the diagonal elements, which, near a local minimum, are presumably mainly positive. We need to work this out for a linear network to get an intuition for how these terms behave.

3.1 Minibatches

How much can minibatches reduce the variance? To address that, we rewrite the last term in Eq. (18) as

$$\Delta \mathbf{w}^k \cdot \frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} \cdot \Delta \mathbf{w}^l = (\Delta \mathbf{w}^k_{\text{true}} + \Delta \mathbf{w}^k - \Delta \mathbf{w}^k_{\text{true}}) \cdot \frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} \cdot (\Delta \mathbf{w}^l_{\text{true}} + \Delta \mathbf{w}^l - \Delta \mathbf{w}^l_{\text{true}}) \, . \tag{29}$$

The second term is zero mean (whether we're using SGD or node perturbation), so averaging eliminates the cross terms,

$$\begin{aligned} \left\langle \Delta \mathbf{w}^k \cdot \frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} \cdot \Delta \mathbf{w}^l \right\rangle &= \Delta \mathbf{w}^k_{\text{true}} \cdot \frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} \cdot \Delta \mathbf{w}^l_{\text{true}} \\ &\quad + \left\langle (\Delta \mathbf{w}^k - \Delta \mathbf{w}^k_{\text{true}}) \cdot \frac{\partial^2 \overline{\mathcal{L}}(\mathbf{w})}{\partial \mathbf{w}^k \partial \mathbf{w}^l} \cdot (\Delta \mathbf{w}^l - \Delta \mathbf{w}^l_{\text{true}}) \right\rangle . \end{aligned} \tag{30}$$

The second term can be reduced using minibatches – so long as the noise, $\boldsymbol{\xi}$, is different on every draw. However, that still leaves the first term. So there's a limit to what minibatches can do. That's true, however, for both SGD and node perturbation.

4 Higher order noise

Let us expand the loss to third order in the noise, which will allow us to compute the gradient to $\mathcal{O}(\sigma^2)$. Ignoring the second order term – which will average to zero once we multiply by $\boldsymbol{\xi}^k \mathbf{x}^k$ and average – we have

$$\mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi}) \approx \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0) + \sum_m \boldsymbol{\delta}^m \cdot \boldsymbol{\xi}^m + \sum_{mnp} \frac{\partial^3 \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial \mathbf{h}^m \partial \mathbf{h}^n \partial \mathbf{h}^p} \cdot \boldsymbol{\xi}^m \boldsymbol{\xi}^n \boldsymbol{\xi}^p \, . \tag{31}$$

As usual, we're abusing notation: the dot means the components of the $\boldsymbol{\xi}$'s must match up with the components in the derivatives. But we'll deal with that later. Multiplying both sides by $\boldsymbol{\xi}^k$ and averaging, we have

$$\langle (\mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi}) - \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)) \boldsymbol{\xi}^k \rangle \approx \sigma^2 \boldsymbol{\delta}^k + \sigma^4 \sum_{mnp} \frac{\partial^3 \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial \mathbf{h}^m \partial \mathbf{h}^n \partial \mathbf{h}^p} \cdot (\mathbf{I}^{mn} \mathbf{I}^{pk} + \mathbf{I}^{mp} \mathbf{I}^{nk} + \mathbf{I}^{mn} \mathbf{I}^{pk}) \, . \tag{32}$$

Because it doesn't matter what order we take the derivatives, all three terms involving $\mathbf{I}$ are the same, and so this simplifies to

$$\langle (\mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, \boldsymbol{\xi}) - \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)) \boldsymbol{\xi}^k \rangle \approx \sigma^2 \boldsymbol{\delta}^k + 3\sigma^4 \sum_{mi} \frac{\partial^3 \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial h_i^{m2} \partial \mathbf{h}^k} \, . \tag{33}$$

Using Eq. (8), we arrive at an expression for the node perturbation update to $\mathcal{O}(\sigma^2)$,

$$\langle \Delta \mathbf{w}_{\text{np}} \rangle \approx -\eta \left[\boldsymbol{\delta}^k \mathbf{x}^k + 3\sigma^2 \sum_{mi} \frac{\partial^3 \mathcal{L}(\mathbf{w}, \mathbf{x}_{\text{in}}, 0)}{\partial h_i^{m2} \partial \mathbf{h}^k} \, \mathbf{x}^k \right] . \tag{34}$$

This expression is copied, with a small tweak, to Eq. (17).

5 Taylor expansions in high dimensions

Taylor expansions in high dimensions are dangerous, as the number of terms increases geometrically with the order (e.g, in d dimension, the k^{th} term has $\mathcal{O}(k^d)$ terms). However, it turns out that with suitable averaging, this problem goes away. To show that, we use repeated integration by parts to write

$$f(t) - f(0) = \sum_{n=1}^{k} \frac{t^n}{n!} \left. \frac{d^n f(t)}{dt^n} \right|_{t=0} + \frac{1}{n!} \int_0^t ds \, (t-s)^n \, \frac{d^{n+1} f(s)}{ds^{n+1}} \,. \tag{35}$$

Let

$$\mathbf{x}(t) \equiv \mathbf{x} + t \Delta \mathbf{x} \,. \tag{36}$$

Then, in a slight abuse of notation,

$$f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) = \sum_{n=1}^{k} \frac{1}{n!} \left. \frac{d^n f(\mathbf{x}(t))}{dt^n} \right|_{t=0} + \frac{1}{n!} \int_0^1 ds \, (1-s)^n \, \frac{d^{n+1} f(\mathbf{x}(s))}{ds^{n+1}} \,. \tag{37}$$

Note that

$$\frac{d}{dt} = \Delta \mathbf{x} \cdot \frac{\partial}{\partial \mathbf{x}} \,, \tag{38}$$

which allows us to write

$$f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) = \sum_{n=1}^{k} \frac{1}{n!} \left(\Delta \mathbf{x} \cdot \frac{\partial}{\partial \mathbf{x}} \right)^n f(\mathbf{x}) + \frac{1}{n!} \left(\Delta \mathbf{x} \cdot \frac{\partial}{\partial \mathbf{x}} \right)^{n+1} \int_0^1 ds \, (1-s)^n \, f(\mathbf{x}(s)) \,. \tag{39}$$

Using $\int_0^1 (1-s)^n = 1/(n+1)$, the last term simplifies,

$$f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) = \sum_{n=1}^{k} \frac{1}{n!} \left(\Delta \mathbf{x} \cdot \frac{\partial}{\partial \mathbf{x}} \right)^n f(\mathbf{x}) + \frac{1}{(n+1)!} \left(\Delta \mathbf{x} \cdot \frac{\partial}{\partial \mathbf{x}} \right)^{n+1} \langle f(\mathbf{x}) \rangle \,. \tag{40}$$

where the average is along the path form $\mathbf{x}$ to $\mathbf{x} + s\Delta\mathbf{x}$, weighted by $(n+1)(1-s)^n$. We are interested in particular in second order expansions,

$$f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) = \Delta \mathbf{x} \cdot \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{2} \Delta \mathbf{x} \cdot \left\langle \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \right\rangle \cdot \Delta \mathbf{x} \,. \tag{41}$$

So the second derivatives in our expansions should really be averaged along a line in weight space. Since we don't know what the derivatives are anyway, that shouldn't be a big deal. He says hopefully.