C Jitter Calculations

The following calculations are used in chapter 17.

C.1 Jitter: Small–Perturbation Approximation

For small noise amplitudes, the network output $y(\mathbf{x}+\mathbf{n})$ can be approximated by

$$y(\mathbf{x}+\mathbf{n}) \approx y(\mathbf{x}) + \left(\frac{\partial y}{\partial \mathbf{x}}\right)^T \mathbf{n} + \frac{1}{2}\mathbf{n}^T\mathbf{H}\mathbf{n} \tag{C.1}$$

where $\mathbf{H}$ is the Hessian matrix with elements $h_{ij} = \partial^2 y/(\partial x_i \partial x_j)$. Assuming an even noise distribution so that $\langle n^k \rangle = 0$ for k odd, one can write

$$\begin{aligned}\mathcal{E} \approx\ & \left\{(t-y)^2\right\} + \sigma^2 \left\{\left\|\frac{\partial y}{\partial \mathbf{x}}\right\|^2\right\} \\ & + \left\{\sigma^2(y-t)\mathrm{Tr}(\mathbf{H}) + \frac{\sigma^4}{4}\mathrm{Tr}(\mathbf{H})^2 + \frac{\sigma^4}{2}\mathrm{Tr}(\mathbf{H}^2) + \frac{m_4 - 3\sigma^4}{4}(\Sigma_i h_{ii}^2)\right\}\end{aligned}$$

where m_4 is the fourth moment $\langle n^4 \rangle$. Dropping all terms higher than second order in σ gives

$$\mathcal{E} \approx \left\{(t-y)^2\right\} + \sigma^2\left\{(y-t)\mathrm{Tr}(\mathbf{H})\right\} + \sigma^2\left\{\left\|\frac{\partial y}{\partial \mathbf{x}}\right\|^2\right\} \tag{C.2}$$

and when $\mathbf{H}$ is assumed to be zero, this reduces to (17.15). The Laplacian term, $\mathrm{Tr}(\mathbf{H}) = \nabla^2 y$, omitted in (17.15), can be described as an approximate measure of the difference between the average surrounding values and the precise value of the field at a point [100]. The third term in (C.2) is the first order regularization term in (17.15).

Training with nonjittered data simply minimizes the error at the training points and puts no constraints on the function at other points. In contrast, training with jitter minimizes the error while also forcing the approximating function to have small derivatives and a local average that approaches the target in the vicinity of each training point.

C.2 Jitter: CDF–PDF Convolution in n Dimensions

The following shows that the convolution of an n–dimensional spherical Gaussian probability density function (PDF) and a Gaussian cumulative distribution function (CDF) results in another Gaussian CDF.

Let $f_1(\mathbf{x})$ be a spherical Gaussian PDF in n–dimensions

$$f_1(\mathbf{x}) = \frac{1}{\sigma_1^n (2\pi)^{n/2}} \exp\left(\frac{-\|\mathbf{x}\|^2}{2\sigma_1^2}\right) \tag{C.3}$$

and let $F_2(\mathbf{x})$ be a Gaussian CDF of the form

$$F_2(\mathbf{x}) = \int_{-\infty}^{w^T\mathbf{x}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\tau^2}{2}\right) d\tau. \tag{C.4}$$

This can be written as

$$F_2(\mathbf{x}) = \int_{-\infty}^{\hat{w}^T\mathbf{x}} \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(\frac{-\tau^2}{2\sigma_2^2}\right) d\tau \tag{C.5}$$

where $\hat{w} = w/\|w\|$ and $\sigma_2 = 1/\|w\|$.

The convolution of F_2 and f_1 is the n–dimensional integral

$$F_2(\mathbf{x}) * f_1(\mathbf{x}) = \int_\alpha F_2(\alpha) f_1(\mathbf{x} - \alpha)\, d\alpha, \tag{C.6}$$

Separate $\mathbf{x}$ and α into components parallel and orthogonal to $\hat{w}$

$$\begin{aligned}
\mathbf{x} &= \ell\hat{w} + \gamma \\
\ell &= \hat{w}^T\mathbf{x} \\
\hat{w}^T\gamma &= 0 \\
\alpha &= k\hat{w} + \beta \\
k &= \hat{w}^T\alpha \\
\hat{w}^T\beta &= 0 \\
\|\mathbf{x} - \alpha\|^2 &= (\ell - k)^2\|\hat{w}\| + 2(\ell - k)\hat{w}^T(\gamma - \beta) + \|\gamma - \beta\|^2 \\
&= (\ell - k)^2 + \|\gamma - \beta\|^2.
\end{aligned}$$

where ℓ and k are scalars and γ and β are n–dimensional vectors orthogonal to $\hat{w}$.

Then

$$
\begin{aligned}
F_2(\alpha) &= \int_{-\infty}^{\hat{w}^T\alpha} \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(\frac{-\tau^2}{2\sigma_2^2}\right) d\tau \\
&= \int_{-\infty}^{k} \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(\frac{-\tau^2}{2\sigma_2^2}\right) d\tau \\
f_1(\mathbf{x}-\alpha) &= \frac{1}{\sigma_1^n(2\pi)^{n/2}} \exp\left(\frac{-\|\mathbf{x}-\alpha\|^2}{2\sigma_1^2}\right) \\
&= \frac{1}{\sigma_1^n(2\pi)^{n/2}} \exp\left(\frac{-(\ell-k)^2}{2\sigma_1^2}\right) \exp\left(\frac{-\|\gamma-\beta\|^2}{2\sigma_1^2}\right) \\
&= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(\frac{-(\ell-k)^2}{2\sigma_1^2}\right) \cdot \frac{1}{\sigma_1^{n-1}(2\pi)^{(n-1)/2}} \exp\left(\frac{-\|\gamma-\beta\|^2}{2\sigma_1^2}\right)
\end{aligned}
\tag{C.7}
$$

and

$$
\begin{aligned}
F_2(\mathbf{x}) * f_1(\mathbf{x}) &= \int_k \int_\beta \int_{-\infty}^{k} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\tau^2/(2\sigma_2^2)}\, d\tau \cdot \left(\frac{1}{\sigma_1\sqrt{2\pi}} e^{-(\ell-k)^2/(2\sigma_1^2)}\right) \\
&\qquad \times \left(\frac{1}{\sigma_1^{n-1}(2\pi)^{(n-1)/2}} e^{-\|\gamma-\beta\|^2/(2\sigma_1^2)}\right) d\beta dk \\
&= \int_k \int_{-\infty}^{k} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\tau^2/(2\sigma_2^2)}\, d\tau \cdot \left(\frac{1}{\sigma_1\sqrt{2\pi}} e^{-(\ell-k)^2/(2\sigma_1^2)}\right) dk \\
&\qquad \times \int_\beta \frac{1}{\sigma_1^{n-1}(2\pi)^{(n-1)/2}} e^{-\|\gamma-\beta\|^2/(2\sigma_1^2)}\, d\beta \\
&= \int_k \int_{-\infty}^{k} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\tau^2/(2\sigma_2^2)}\, d\tau \cdot \left(\frac{1}{\sigma_1\sqrt{2\pi}} e^{-(\ell-k)^2/(2\sigma_1^2)}\right) dk.
\end{aligned}
$$

Thus $F_2(\mathbf{x}) * f_1(\mathbf{x})$ reduces to a one-dimensional convolution of a Gaussian CDF with standard deviation $\sigma_2 = 1/\|w\|$ and a Gaussian PDF with standard deviation σ_1. It can be shown (see section C.3) that this is a Gaussian CDF with variance $\sigma_3^2 = \sigma_1^2 + \sigma_2^2$.

Letting Z_a denote the Gaussian CDF function with standard deviation a,

$$\begin{aligned} F_2(\mathbf{x}) * f_1(\mathbf{x}) &= Z_{\sigma_2}(\ell) * g(\ell) \\ &= Z_{\sigma_3}(\ell) \\ &= Z_1\left(\frac{\ell}{\sigma_3}\right) \\ &= Z_1\left(\frac{\hat{w}^T\mathbf{x}}{\sqrt{\sigma_1^2 + (1/\|w\|)^2}}\right) \\ &= Z_1\left(\frac{\|w\|\ \hat{w}^T\mathbf{x}}{\sqrt{\|w\|^2\sigma_1^2 + 1}}\right) \\ &= Z_1\left(\frac{w^T\mathbf{x}}{\sqrt{\|w\|^2\sigma_1^2 + 1}}\right) \\ &= F_2\left(\frac{\mathbf{x}}{\sqrt{\|w\|^2\sigma_1^2 + 1}}\right) \end{aligned} \tag{C.8}$$

Thus, the convolution of a Gaussian CDF and a Gaussian PDF can be computed by a simple scaling of the original CDF.

C.3 Jitter: CDF–PDF Convolution in One Dimension

The following demonstrates that the convolution of Gaussian PDF with variance σ_1^2 and a Gaussian CDF with variance σ_2^2 results in a Gaussian CDF with variance $\sigma_3^2 = \sigma_1^2 + \sigma_2^2$. All the functions are one–dimensional.

Consider two independent random variables X_1 and X_2 with PDFs f_1 and f_2 and CDF's F_1 and F_2. The random variable $Y = X_1 + X_2$ has the PDF $f_1 * f_2$ and consequently its CDF is $F_1 * f_2 = f_1 * F_2$. Let X_1 and X_2 be zero mean Gaussian, $X_1 \sim N(0, \sigma_1^2)$ and $X_2 \sim N(0, \sigma_2^2)$, then, clearly, $Y \sim N(0, \sigma_1^2 + \sigma_2^2)$ has a Gaussian PDF with variance $\sigma_3^2 = \sigma_1^2 + \sigma_2^2$. Because Y has the CDF $f_1 * F_2$, $f_1 * F_2$ is a Gaussian CDF with zero mean and variance $\sigma_3^2 = \sigma_1^2 + \sigma_2^2$.