1 Setting

We are interested in characterizing the structure learned in the first layer of a deep neural network when trained on a simple task.

Data In the most general case, our data are sampled from zero-mean stochastic processes defined on the probability space $(\Omega, \mathcal{F}, p)$ with values in the measurable space $(\mathbb{R}, S)$. The measure p is parameterized by $\xi > 0$, which controls the length-scale of the correlations, and $g > 0$, which controls the degree of non-Gaussianity. (I'm not sure how to rigorously add more structure to this.) A single realization of this stochastic process is denoted by $X(t)$.

Informally, ξ controls the frequency of oscillations[1] in $X(t)$, but it is important to note that $X(t)$ is not actually periodic. See Figure 1 for some samples.

As $g \to 0$, $X(t)$ converges in distribution to a Gaussian process with covariance function $\Sigma(\xi)(x, x')$. As g increases, $X(t)$ becomes increasingly non-Gaussian[2].

We take our index set to be $\Omega = \mathbb{T}^d$, the d-dimensional torus, so that our samples are periodic. This naturally introduces Fourier series[3].

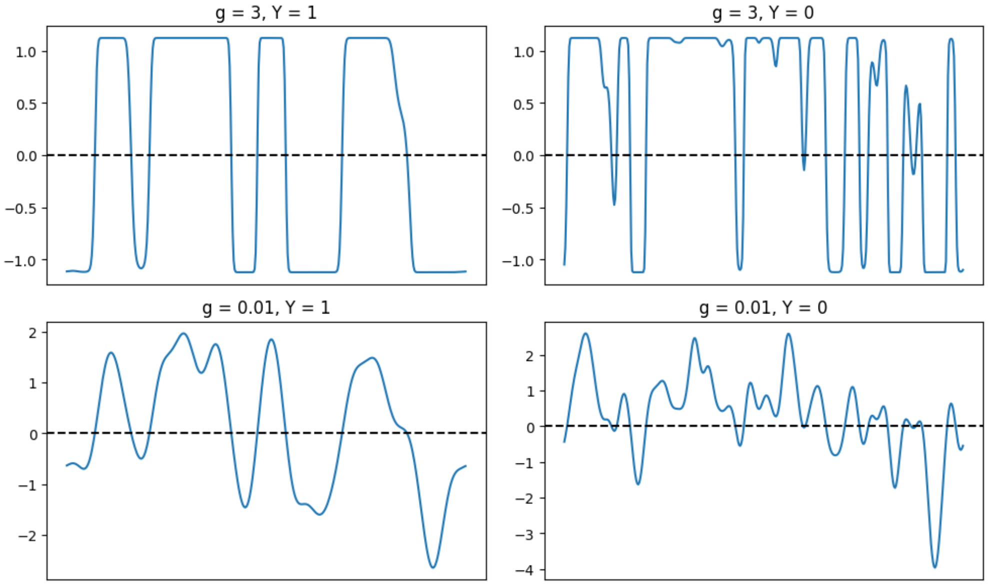

Figure 1: 1-d samples $X(t)$ from discretization with $n = 400$, $(\xi_1, \xi_0) = (20, 10)$.

Model Our model is a two-layer neural network with K hidden neurons and activation function σ.

$$\hat{y}(x) = \frac{1}{K} \sum_{k \in [K]} \sigma(\langle w_k, x \rangle + b_k). \tag{1}$$

[1] "Oscillation" isn't the right word, but ξ certainly controls the number of times $X(t)$ crosses zero, which is what I mean by frequency. Importantly, this definition of "frequency" is invariant to g! (Note that the correlation between two points is not invariant to choice of g, so this might actually be a useful definition.)

[2] What does this mean, exactly? If it's like what Alessandro did, then it means super-Gaussian tails, kind of.

[3] But do we need them?

We call w_k the receptive field (RF) of the k-th neuron and b_k its bias. Here, $\langle w_k, x\rangle = \int_{\mathbb{T}^d} w_k(t)x(t)dt$.

Objective The task is to discriminate which of two processes $\{X_1(t)\}$ and $\{X_0(t)\}$, parameterized by ξ_1 and ξ_0, respectively, for some fixed g, a given sample $X(t)$ was drawn from. If the former, the desired label is Y_1, otherwise it is Y_0. So, we seek weights $\{w_k, b_k\}_{k\in[K]}$ that minimize the loss

$$\mathcal{L} = \frac{1}{2}\,\mathbb{E}_{X,Y}\left[\left(\hat{y}(X) - Y\right)^2\right]. \tag{2}$$

Practice I've presented a very general version of the problem we have to strip away what I deem irrelevant details. However, in practice, we must add many more constraints.

First, we cannot train continuous RFs or sample from arbitrary stochastic processes. So, we restrict Ω to the finite set $\{\frac{i}{n} \mid 0 \leq i < n\}^d \subseteq \mathbb{T}^d$. In fact, we typically just take $d = 1$. So, $X(t)$ and w_k are vectors in $\mathbb{R}^n$. The inner product simplifies to the standard Euclidean inner product.

Secondly, we tune w_k and b_k using full-batch stochastic gradient descent, which is not guaranteed to minimize the loss in equation (2).

2 Functional Behavior

While analyzing this model exactly is nearly impossible, its functional behavior is quite easy to understand. We will consider sigmoid activation because it yields both oscillatory and localized RFs. We'll start by simplifying the model's structure a bit before explaining exactly how it solves the task.

Reducing the model

Frequency specialization With sigmoid activation, the model learns to solve the task by creating two populations of neurons roughly in equal size, one with RFs resembling localized weights (or long-range oscillations) and the other with RFs resembling short-range oscillations. We call this *frequency specialization.*

We'll attach the label $-$ to the former and $+$ to the latter. (This is because the former learns negative bias terms, while the latter learns positive ones. More on this later.) The biases within each class are all approximately the same. So, we'll reduce our model to

$$\hat{y}(x) = \frac{1}{K}\Bigg(\underbrace{\sum_{k\in[-]} \sigma(\langle w_k, x\rangle + b^-)}_{\text{localized}} + \underbrace{\sum_{k\in[+]} \sigma(\langle w_k, x\rangle + b^+)}_{\text{oscillatory}}\Bigg). \tag{3}$$

The facts that it constructs two populations of neurons and that they are equal in size seems to be arbitrary. It appears that either population could solve the task on its own, and balancing is not required.

[insert figure showing unbalanced classes]

(TODO: Test if just oscillatory neurons can do the task!) [insert figure showing just localized (and perhaps just oscillatory) solve the task]

Tiling In addition to the biases being approximately the same within each class, the RFs also exhibit a lot of symmetry. Within each class, the RFs differ only in where they are centered (for long-range neurons, at least) and their sign. That is, they implement a high-degree of *weight sharing* and *tile* the input space. As we increase K, the number of hidden neurons, to be very large, they effectively implement a convolution (but the integrand is wrapped by σ). We further reduce our model to

$$\begin{aligned}\hat{y}(x) = \frac{1}{4}\Bigg(&\underbrace{\int_{\mathbb{T}} \sigma((w^- * x)(\tau) + b^-)d\tau + \int_{\mathbb{T}} \sigma(-(w^- * x)(\tau) + b^-)d\tau}_{\text{localized}} + \\ &\underbrace{\int_{\mathbb{T}} \sigma((w^+ * x)(\tau) + b^+)d\tau + \int_{\mathbb{T}} \sigma(-(w^+ * x)(\tau) + b^+)d\tau}_{\text{oscillatory}} \Bigg).\end{aligned} \tag{4}$$

Note that when $(w^- * x)(\tau) + b^-$ is large enough to be significant in the sigmoid, then $-(w^- * x)(\tau) + b^-$ is extremely small and thus insignificant. So, we can reduce the first two terms by considering just $|(w^- * x)(\tau)| + b^-$. The same holds for w^+. Thus,

$$\hat{y}(x) = \frac{1}{2}\Bigg(\underbrace{\int_{\mathbb{T}} \sigma(|(w^- * x)(\tau)| + b^-)d\tau}_{\text{localized}} + \underbrace{\int_{\mathbb{T}} \sigma(|(w^+ * x)(\tau)| + b^+)d\tau}_{\text{oscillatory}} \Bigg). \tag{5}$$

Localization The final phenomenon we observe is that w^- is highly localized, while w^+ is not. To try to explain this, we'll consider how the model solves the task. Let's just focus on w^- for now. We want w^- to be structured so that it yields outputs closer to 1 upon observing data from p_{ξ_1} and -1 otherwise (assuming $\sigma(x) = \mathrm{erf}(\frac{x}{\sqrt{2}})$). Ideally,

$$\mathbb{P}_{X_1,X_0}\left(\int \sigma(|(w^- * X_1)(\tau)| + b^-)d\tau > \int \sigma(|(w^- * X_0)(\tau)| + b^-)d\tau \right) \approx 1, \tag{6}$$

where the probability is over independent draws $X_1 \sim p_{\xi_1}$ and $X_0 \sim p_{\xi_0}$. This condition is equivalent to having w^- *maximize the spread in the preactivations corresponding to* p_{ξ_1} *relative to those for* p_{ξ_0}. I'll explain why this is true now.

In the context of equation (6), let's define the optimal RF w^- as

$$w^- \triangleq \arg\max_w \mathbb{P}_{X_1,X_0}\left(\int \sigma(|(w * X_1)(\tau)| + b^-)d\tau > \int \sigma(|(w * X_0)(\tau)| + b^-)d\tau \right).$$

Considering independent uniform draws τ, τ' from $\mathbb{T}$, we can (I think) rewrite this as

$$\begin{aligned}
w^- &= \arg\max_w \mathbb{P}_{X_1,X_0,\tau,\tau'}\left(\sigma(|(w * X_1)(\tau)| + b^-) > \sigma(|(w * X_0)(\tau')| + b^-) \right) && (7)\\
&= \arg\max_w \mathbb{P}_{X_1,X_0,\tau,\tau'}\left(|(w * X_1)(\tau)| > |(w * X_0)(\tau')| \right) && \text{monotonicity} \quad (8)\\
&= \arg\max_w \mathbb{P}_{X_1,X_0}\left(|\langle w, X_1\rangle| > |\langle w, X_0\rangle| \right), && \text{translation invariance} \quad (9)\\
&= \arg\max_w \mathbb{P}_{X_1,X_0}\left((\langle w, X_1\rangle)^2 > (\langle w, X_0\rangle)^2 \right), && (10)
\end{aligned}$$

Equations (9) and (10) are obviously equivalent, but equation (10) just makes it a bit clearer that we want to think about the *spread* of the preactivations. Both $\langle w, X_1\rangle$ and $\langle w, X_0\rangle$ are symmetric, zero-mean random variables. So, we cannot try to maximize the difference in their means. However, because we have a (learnable) bias term, we can exploit the difference in their variances. Specifically, we want to maximize the variance of $\langle w, X_1\rangle$ relative to that of $\langle w, X_0\rangle$.

Understanding the model

Equation (9) gives us a reduced model to understand the optimization problem the model is solving. Now, let's try to understand how it might solve this problem. First, a caveat.

A cheat In moving from equation (7) to equation (8), we made the assumption that b^- was fixed. In practice, we tune b^- along with w^-. So, if the model is able to achieve *some* advantage in equation (9), it can just increase the norms of w and b^- to achieve a larger advantage in equation (8). This is a perfectly valid strategy, and one I would expect it to exploit. This yields an important question: *Does the model take the easy route by scaling up w, or does it properly optimize w?*

Empirically, we see that the norms of w^- and b^- do indeed grow throughout training. The shape of the RF w^- emerges relatively early on in training before it is "smoothed out" and scaled up in norm.

[TODO: insert plots showing this transition between the two states]

Gaussian data Let's momentarily assume X_1 and X_0 are Gaussian (i.e. $g \approx 0$) and go back to discrete land (it's a bit easier for my brain). Then, $\langle w, X_i \rangle \sim \mathcal{N}\left(0, w^\top \Sigma_i w\right)$. Recall that X_1 and X_0 are drawn independently. So,

$$\begin{aligned}
\mathbb{P}_{X_1,X_0}\left(|\langle w, X_1\rangle| > |\langle w, X_0\rangle|\right) &= \int_{\mathbb{R}} \int_{[-|x_1|,|x_1|]} p_0(x_0) p_1(x_1) dx_0 dx_1 \\
&= \int_{\mathbb{R}} \left[\Phi\left(\frac{|x_1|}{\sqrt{w^\top \Sigma_0 w}}\right) - \Phi\left(\frac{-|x_1|}{\sqrt{w^\top \Sigma_0 w}}\right)\right] p_1(x_1) dx_1 \\
&= 2\int_{\mathbb{R}_{\geq 0}} \operatorname{erf}\left(\frac{1}{\sqrt{2}} \cdot \frac{x_1}{\sqrt{w^\top \Sigma_0 w}}\right) p_1(x_1) dx_1 \\
&= 1 - \frac{2}{\pi} \tan^{-1}\left(\sqrt{\frac{w^\top \Sigma_0 w}{w^\top \Sigma_1 w}}\right).
\end{aligned}$$

Therefore, for Gaussian data,

$$\begin{aligned}
w^- &= \operatorname*{arg\,max}_{w \in S^{n-1}} \mathbb{P}_{X_1,X_0}\left(|\langle w, X_1\rangle| > |\langle w, X_0\rangle|\right) \\
&= \operatorname*{arg\,max}_{w \in S^{n-1}} 1 - \frac{2}{\pi} \tan^{-1}\left(\sqrt{\frac{w^\top \Sigma_0 w}{w^\top \Sigma_1 w}}\right) \\
&= \operatorname*{arg\,min}_{w \in S^{n-1}} \frac{w^\top \Sigma_0 w}{w^\top \Sigma_1 w} \\
&= F\left(\operatorname*{arg\,min}_{u \in S^{n-1}} \frac{u^\top \Lambda_0 u}{u^\top \Lambda_1 u}\right),
\end{aligned}$$

where F is our real DFT operator and Λ_i is the corresponding diagonal matrix of eigenvalues of Σ_i. Again, we can take u to be unit norm without loss of generality. Then, $u^\top \Lambda_0 u = \sum_{k \in [n]} c_k \lambda_0^{(k)}$, where $c_k \geq 0$ and $\sum_k c_k = 1$.

Polar coordinates Going back to discrete land, we can write the inner product in terms of the norm of w and the angle between w and X_i. Starting from equation (9),

$$\begin{aligned} w^- &= \arg\max_w \mathbb{P}_{X_1,X_0} \left(|\cos\theta(w, X_1)| \|X_1\|_2 > |\cos\theta(w, X_0)| \|X_0\|_2 \right) && (11) \\ &= \arg\max_w \mathbb{P}_{X_1,X_0} \left(\frac{|\cos\theta(w, X_1)|}{|\cos\theta(w, X_0)|} > \frac{\|X_0\|_2}{\|X_1\|_2} \right) && (12) \end{aligned}$$

We cannot control the norms on the right side of the inequality. So, we need to construct w to maximize the left side. We do this by making w be as parallel to X_1 as possible while making it as orthogonal as possible to ${X_0}^4$.

Fourier space Returning to the continuous realm, we use Parseval's theorem[5] to rewrite equation (9)

$$\begin{aligned} w^- &= \arg\max_w \mathbb{P}_{X_1,X_0} \left(|\langle \mathcal{F}\, w, \mathcal{F}\, X_1 \rangle| > |\langle \mathcal{F}\, w, \mathcal{F}\, X_0 \rangle| \right), && (13) \\ &= \arg\max_{\widehat{w}} \mathbb{P}_{\widehat{X}_1,\widehat{X}_0} \left(|\langle \widehat{w}, \widehat{X}_1 \rangle| > |\langle \widehat{w}, \widehat{X}_0 \rangle| \right). && (14) \end{aligned}$$

All we've done is represent the problem in Fourier space. I'm not sure if this is actually helpful.

Difference in means It's hard to make a ton of analytical progress here, so we'll have to defer to simulations (for now). We can, though, compute the means of the squared terms in equation (10). For $i = 0, 1$,

$$\begin{aligned} \mathbb{E}_{X_i} \left[(\langle w, X_i \rangle)^2 \right] &= \mathbb{E}_{X_i} \left[\left(\int_{\mathbb{T}} w(s) X_i(s) ds \right) \left(\int_{\mathbb{T}} w(t) X_i(t) dt \right) \right] \\ &= \int_{\mathbb{T}} \int_{\mathbb{T}} w(s) w(t)\, \mathbb{E}_{X_i} \left[X_i(s) X_i(t) \right] ds dt \\ &= \int_{\mathbb{T}} \int_{\mathbb{T}} w(s) w(t) k_i(s,t) ds dt \\ &= \langle K_i w, w \rangle_{L^2}, \end{aligned}$$

where $K_i w(t) = \int_{\mathbb{T}} k_i(s,t) w(s) ds$. Therefore,

$$\begin{aligned} \mathbb{E}_{X_1} \left[(\langle w, X_1 \rangle)^2 \right] - \mathbb{E}_{X_0} \left[(\langle w, X_0 \rangle)^2 \right] &= \langle K_1 w, w \rangle_{L^2} - \langle K_0 w, w \rangle_{L^2} \\ &= \langle (K_1 - K_0) w, w \rangle_{L^2}. \end{aligned}$$

Note that the probability in equation (10) is invariant to scaling w. So, let's fix $\|w\|_{L^2} = 1$. Then, maximizing this is just finding the leading eigenvector of the operator $K_1 - K_0$. We can show that this is the constant function, which is inconsistent with our observation that we get Gabor-like RFs. Let's test this result empirically to see has any merit (note that maximizing the difference in means is *not* the same as maximizing the probability dominance), and if Gabors are actually better.

Optimality

We've explored what it takes for our neural network model to be optimal, but we want to understand how close it gets to universal optimality. We'll consider accuracy instead of loss because it lets us invoke the Neyman-Pearson lemma.

[4] We haven't really said anything new here—this is already apparent in equations (9) and (10).

[5] This says the Fourier transform is unitary from $L^2(\mathbb{T})$ to $\ell^2(\mathbb{Z})$. The result for DFT is analogous.

Let us start by considering an arbitrary classifier $\hat{y}$ that sends an input X_i to either 0, if it believes $X_i \sim p_{\xi_0}$, or 1 otherwise. Then, the optimal classifier, in terms of accuracy, is

$$\hat{y}^* = \arg\max_{\hat{y}} \mathbb{P}_{X_1,X_0}(\hat{y}(X_i) = i)$$

We can pose this as a hypothesis testing problem.

$$\begin{aligned} H_0 &: X \sim p_{\xi_0} \\ H_1 &: X \sim p_{\xi_1} \end{aligned}$$

The Neyman-Pearson lemma tells us that the likelihood ratio test has the largest accuracy[6] among all tests. That is,

$$\hat{y}^*(x) = \begin{cases} 1 & \text{if } L(x) \triangleq \frac{p_{\xi_1}(x)}{p_{\xi_0}(x)} > \eta \\ 0 & \text{otherwise} \end{cases},$$

for some threshold $\eta > 0$. The corresponding accuracy is

$$\frac{1}{2}\left(\mathbb{P}_{X_1}\left(L(X_1) > \eta\right) + \mathbb{P}_{X_0}\left(L(X_0) \leq \eta\right)\right)$$

I think we can make some progress here for the general case, but let's just consider Gaussian data for now.

Gaussian data

$$\begin{aligned} \frac{p_{\xi_1}(x)}{p_{\xi_0}(x)} &= \frac{\frac{1}{(2\pi)^{n/2}}|\Sigma_1|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^\top \Sigma_1^{-1} x\right)}{\frac{1}{(2\pi)^{n/2}}|\Sigma_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^\top \Sigma_0^{-1} x\right)} \\ &= \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}} \exp\left(-\frac{1}{2}x^\top(\Sigma_1^{-1} - \Sigma_0^{-1})x\right) \\ &\simeq x^\top(\Sigma_1^{-1} - \Sigma_0^{-1})x, \end{aligned}$$

where $\simeq$ means equal up to some monotonic transformation.

So, the optimal accuracy is given by

$$\min_{\eta>0} \frac{1}{2}\left(\mathbb{P}_{X_1}\left(X_1^\top(\Sigma_1^{-1} - \Sigma_0^{-1})X_1 < \eta\right) + \mathbb{P}_{X_0}\left(X_0^\top(\Sigma_1^{-1} - \Sigma_0^{-1})X_0 \geq \eta\right)\right)$$

[6] Note that 1−accuracy = Type 1 error + Type II error, which the likelihood ratio test minimizes.

3 Sigmoid → ReLU

Our data are sampled from zero-mean distributions of the form

$$X \sim p(\xi, g) \in \mathbb{R}^n, \tag{15}$$

where $\xi > 0$ controls the length-scale of the correlations while $g > 0$ controls the degree of non-Gaussianity. As $g \to 0$, $p \overset{d}{\to} \mathcal{N}(0, \Sigma)$ for some covariance matrix $\Sigma(\xi)$. The labels Y are a bijection with ξ.

For some general activation function σ, our model is

$$\hat{y}(x) \triangleq \frac{1}{K} \sum_{k \in [K]} \sigma(\langle w_k, x \rangle + b_k). \tag{16}$$

The loss is

$$\begin{aligned} \mathcal{L} &\triangleq \frac{1}{2}\, \mathbb{E}_{X,Y}\left[(\hat{y}(X) - Y)^2\right] && (17) \\ &= \frac{1}{4}\left(\mathbb{E}_{X \sim p(\xi_1, g)}\left[(\hat{y}(X) - Y_1)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g)}\left[(\hat{y}(X) - Y_0)^2\right]\right). && (18) \end{aligned}$$

We want to understand why, as $g \to \infty$, the starting loss for ReLU activation seems to decrease, while it (sensibly) stays fixed for sigmoid activation.

Sigmoid Our starting choice of activation function is $\sigma(x) = \operatorname{erf}(\frac{x}{\sqrt{2}})$, with which we use labels $Y_1, Y_0 = 1, -1$. Let us take $K = 1$ for simplicity. At initialization, $b_1 = 0$, so

$$\begin{aligned} 4\,\mathcal{L}(0) &= \mathbb{E}_{X \sim p(\xi_1, g)}\left[\left(\hat{y}(X) - 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g)}\left[\left(\hat{y}(X) + 1\right)^2\right] && (19) \\ &= \mathbb{E}_{X \sim p(\xi_1, g)}\left[\left(\sigma(\langle w_1, X \rangle) - 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g)}\left[\left(\sigma(\langle w_1, X \rangle) + 1\right)^2\right] && (20) \\ &= \mathbb{E}_{X \sim p(\xi_1, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) - 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_1, g), \langle w_1, X \rangle < 0}\left[\left(\sigma(\langle w_1, X \rangle) - 1\right)^2\right] \\ &\quad + \mathbb{E}_{X \sim p(\xi_0, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) + 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g), \langle w_1, X \rangle < 0}\left[\left(\sigma(\langle w_1, X \rangle) + 1\right)^2\right] && (21) \\ &= \mathbb{E}_{X \sim p(\xi_1, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) - 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_1, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) + 1\right)^2\right] \\ &\quad + \mathbb{E}_{X \sim p(\xi_0, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) + 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) - 1\right)^2\right] && (22) \\ &= \mathbb{E}_{X \sim p(\xi_1, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) - 1\right)^2 + \left(\sigma(\langle w_1, X \rangle) + 1\right)^2\right] \\ &\quad + \mathbb{E}_{X \sim p(\xi_0, g), \langle w_1, X \rangle > 0}\left[\left(\sigma(\langle w_1, X \rangle) + 1\right)^2 + \left(\sigma(\langle w_1, X \rangle) - 1\right)^2\right] && (23) \\ &= 2\left(\mathbb{E}_{X \sim p(\xi_1, g), \langle w_1, X \rangle > 0}\left[\sigma(\langle w_1, X \rangle)^2 + 1\right] + \mathbb{E}_{X \sim p(\xi_0, g), \langle w_1, X \rangle > 0}\left[\sigma(\langle w_1, X \rangle)^2 + 1\right]\right) && (24) \\ &= && (25) \end{aligned}$$

ReLU With ReLU activation, we have

$$\begin{aligned} 4\,\mathcal{L} &= \mathbb{E}_{X \sim p(\xi_1, g)}\left[\left(\hat{y}(X) - 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g)}\left[\left(\hat{y}(X)\right)^2\right] && (26) \\ &= \mathbb{E}_{X \sim p(\xi_1, g)}\left[\left(\mathrm{ReLU}(w_1, X) - 1\right)^2\right] + \mathbb{E}_{X \sim p(\xi_0, g)}\left[\left(\mathrm{ReLU}(w_1, X)\right)^2\right] && (27) \end{aligned}$$

Weird idea: use L^1-loss

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{X,Y}\left[|Y - \hat{y}(X)|\right] &(28)\\ &= \mathbb{E}_{X|Y=1}\left[1 - \hat{y}(X)\right] + \mathbb{E}_{X|Y=-1}\left[1 + \hat{y}(X)\right]. &(29)\end{aligned}$$

Then, for sigmoid activation,

$$\frac{\partial\,\mathcal{L}}{\partial w_1} = -\,\mathbb{E}_{X|Y=1}\left[\frac{\partial \hat{y}(X)}{\partial w_1}\right] + \mathbb{E}_{X|Y=-1}\left[\frac{\partial \hat{y}(X)}{\partial w_1}\right]. \tag{30}$$

Note that

$$\begin{aligned}\mathbb{E}_{X|Y=1}\left[\frac{\partial \hat{y}(X)}{\partial w_1}\right] &= \frac{1}{K}\,\mathbb{E}_{X|Y=1}\left[\sigma'(\langle w_1, X\rangle + b_1)X\right] &(31)\\ &= \frac{1}{\sqrt{\pi}}\frac{1}{K}\,\mathbb{E}_{X|Y=1}\left[e^{-\frac{1}{2}(\langle w_1, X\rangle + b_1)^2}X\right]. &(32)\end{aligned}$$

Let's focus on a single entry of the expectation.

$$\mathbb{E}_{X|Y=1}\left[e^{-\frac{1}{2}(\sum_{j=1}^n w_1^{(j)}X_j + b_1)^2}X_i\right] \tag{33}$$

$$= \mathbb{E}_{X|Y=1}\left[\exp\left(-\frac{1}{2}\left(\sum_{j=1}^n (w_1^{(j)})^2 X_j^2 + 2\sum_{k>j} w_1^{(k)} w_1^{(j)} X_k X_j + 2b_1\sum_{j=1}^n w_1^{(j)} X_j + b_1^2\right)\right) X_i\right] \tag{34}$$