1 Setting

We can mostly explain symmetry breaking and tiling, and so we finally want to explain localization. To do this, I focus on a single-neuron model with ReLU activation.

The task is to discriminate between two classes of inputs:

$$X_1 \sim p(\xi_1), \quad X_0 \sim p(\xi_0), \tag{1}$$

which are n-dimensional vectors. We only assume that p is *translation-invariant* (this is explained more below) and that it is symmetric about 0, i.e. $p(x) = p(-x)$. Inputs X_i have scalar label $Y = i$, for $i = 0, 1$. The distribution p is paramaterized by $\xi > 0$, which defines the length-scale of correlations in the input. Specifically, we construct p so that

$$\mathrm{Cov}(X) = \Sigma(\xi), \quad \Sigma(\xi)_{ij} = k(i,j) \triangleq \exp(-(i-j)^2/\xi^2). \tag{2}$$

Later, we will consider more general kernels k.

We consider one neuron without bias and ReLU activation, since it's the simplest model that captures the localization phenomenon.

$$\hat{y}(x) = \mathrm{ReLU}(\langle w, x \rangle), \tag{3}$$

where w is our receptive field.

2 Dynamics

The dynamics of w is given by

$$\tau \frac{d}{dt} w = -\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{2}\big[\underbrace{\mathbb{E}_{X|Y=1}\left[\mathbb{1}(\langle w, X\rangle \geq 0) X\right]}_{\triangleq f(w)}\big] - \frac{1}{4}(\Sigma_0 + \Sigma_1) w, \tag{4}$$

This elucidates some universal structure and lets us establish some properties of f.

1. *It is invariant to scaling* w: $f(w) = f(\alpha w)$ for $\alpha > 0$.
2. *It is sign equivariant*: $f(-w) = -f(w)$.
3. *It is translation equivariant*[1]: $f(\mathcal{C}\, w) = \mathcal{C}\, f(w)$, where $\mathcal{C}$ is a circular shift.

Thus, f only depends on the shape of w, not it's magnitude or position. So, it is precisely the object we need to understand. Note we can reduce properties 2 and 3 to the more general statement:

4. *It can preserve symmetry in* p: If p is symmetric w.r.t. some invertible linear transformation A, i.e. $p_\xi(x) = p_\xi(Ax)$ for all x, then $f(Aw) = A^{-\top} f(w)$. If A is orthogonal, then $f(Aw) = Af(w)$.

What is the second term, $(\Sigma_0 + \Sigma_1)w$, doing? It's immediately clear that it is a *shrinkage term* since Σ is positive semi-definite.

[1] Here, "translation" refers to shifts of the *entries* of a vector. So, if $\mathcal{C}$ is a shift down by 1, $x_i = (\mathcal{C}\, x)_{i+1}$

One could try to understand in terms of the convolutions implemented by Σ_0 and Σ_1, which we call k_0 and k_1, respectively[2]. So, we can write the second term as $(k_0 + k_1) * w$. However, I still find it hard to understand how this drives the dynamics. *I could use some help here.*

Another way to interpret the second term is by using the discrete Fourier transform. Because Σ_0 and Σ_1 are circulant matrices, they both diagonalize in the discrete Fourier basis[3], which we denote P. Let Λ_i be the diagonal matrix of eigenvalues for Σ_i. Representing w in Fourier space by $u = P^\top w$, we have

$$\tau \frac{d}{dt} u = \frac{1}{2} P^\top f(Pu) - \frac{1}{2}(\Lambda_0 + \Lambda_1)u. \tag{5}$$

Note the eigenvalues in Λ_0 and Λ_1 are all positive, but Λ_1 weights lower-frequency oscillations more heavily than Λ_0, and higher-frequency ones less. Together, they act to shrink low-frequency oscillations faster than high-frequency ones[4].

Extra fact One additional thing we can say about f is that

$$\frac{\partial}{\partial w} f(w)_i \perp w \quad \forall i, \qquad \text{and} \qquad \frac{\partial}{\partial w_j} f(w) \perp w \quad \forall j.$$

This is either something important about interpreting f or entirely obvious. Not sure which, but this seems like a special consequence of using ReLU activation. To see why:

$$\frac{\partial}{\partial w_j} f(w)_i = \frac{\partial}{\partial w_j} \int_{\mathbb{R}^n} p_\xi(x) \mathbb{1}(\langle w, x \rangle \geq 0) x_i dx = \int_{\mathbb{R}^n} p_\xi(x) \delta(\langle w, x \rangle) x_i x_j dx,$$

where δ is the Dirac delta function.

2.1 General Setting

We isolate the i-th entry,

$$\begin{aligned} f(w)_i &= \mathbb{E}_{X|Y=1} \left[\mathbb{1}(\langle w, X \rangle \geq 0) X_i \right] \\ &= \mathbb{E}_{X_i|Y=1} \left[X_i \, \mathbb{E}_{X|Y=1, X_i} \left[\mathbb{1}(\langle w, X \rangle \geq 0) \right] \right] \\ &= \mathbb{E}_{X_i|Y=1} \left[X_i \Pr\left(\langle w, X \rangle \geq 0 \mid Y = 1, X_i \right) \right]. \end{aligned}$$

Let's focus on the probability term, starting by writing all the randomness on the left side:

$$\Pr\left(\langle w, X \rangle \geq 0 \mid Y = 1, X_i \right) = \Pr\left(\sum_{j \neq i} w_j X_j \geq -w_i X_i \mid Y = 1, X_i \right).$$

This probability is impossible to compute exactly in general, but we can use a Gaussian approximation on $X \mid Y = 1, X_i$ to make it tractable. Let $\mu_{|X_i}$ $\Sigma_{|X_i}$ be the mean and covariance for $X_{\not i} \mid Y = 1, X_i \in \mathbb{R}^{n-1}$, where $X_{\not i}$ is the vector X with the i-th entry removed. Then, $\sum_{j \neq i} w_j X_j \mid Y = 1, X_i \sim \mathcal{N}(w_{\not i}^\top \mu_{|X_i}, w_{\not i}^\top \Sigma_{|X_i} w_{\not i})$. So,

$$\Pr\left(\langle w, X \rangle \geq 0 \mid Y = 1, X_i \right) \approx 1 - \Phi\left(\frac{-w_i X_i - w_{\not i}^\top \mu_{|X_i}}{\sqrt{w_{\not i}^\top \Sigma_{|X_i} w_{\not i}}} \right) = \Phi\left(\frac{w_i X_i + w_{\not i}^\top \mu_{|X_i}}{\sqrt{w_{\not i}^\top \Sigma_{|X_i} w_{\not i}}} \right).$$

[2] Recall Σ is circulant.

[3] In particular, we can take the real and imaginary parts of the complex Fourier modes because Σ is always real and symmetric.

[4] Shouldn't it be the other way around?

Thus,

$$f(w)_i \approx \mathbb{E}_{X_i|Y=1}\left[X_i\Phi\left(\frac{w_iX_i + w_{\not i}^\top \mu_{|X_i}}{\sqrt{w_{\not i}^\top \Sigma_{|X_i} w_{\not i}}}\right)\right]. \tag{6}$$

Rewriting Finally, let us rewrite this a bit to highlight its similarity to the Gaussian case. Let us define $\tilde{\mu}_{|X_i}$ and $\tilde{\Sigma}_{|X_i}$ to be the mean and covariance for $X \mid Y = 1, X_i \in \mathbb{R}^n$. Note the i-th entry of $\tilde{\mu}_{|X_i}$ is X_i and the j-th row and column of $\tilde{\Sigma}_{|X_i}$ are 0. Then, $w_iX_i + w_{\not i}^\top \mu_{|X_i} = w^\top \tilde{\mu}_{|X_i}$ and $w_{\not i}^\top \Sigma_{|X_i} w_{\not i} = w^\top \tilde{\Sigma}_{|X_i} w$. So,

$$f(w)_i \approx \mathbb{E}_{X_i|Y=1}\left[X_i\Phi\left(\frac{w^\top \tilde{\mu}_{|X_i}}{\sqrt{w^\top \tilde{\Sigma}_{|X_i} w}}\right)\right] \tag{7}$$

$$= \mathbb{E}_{X_i|Y=1}\left[X_i\Phi\left(\frac{w^\top \tilde{\mu}_{|X_i}}{\sqrt{\mathbb{E}_{X|X_i,Y=1}\left[(w^\top X)^2\right] - (w^\top \tilde{\mu}_{|X_i})^2}}\right)\right] \tag{8}$$

What can we say about the conditional mean? We can relate it to the unconditional covariance:

$$[(\Sigma_1)_{ji}]_{j\in[n]} = \mathbb{E}_{X|Y=1}\left[X_iX\right] = \mathbb{E}_{X_i|Y=1}\left[X_i\,\mathbb{E}_{X|X_i,Y=1}\left[X\right]\right] = \mathbb{E}_{X_i|Y=1}\left[X_i\tilde{\mu}_{|X_i}\right] = \mathbb{E}_{X_i|X_i>0,Y=1}\left[X_i\tilde{\mu}_{|X_i}\right],$$

where last step follows by the symmetry of p about 0, and the term on the left is the i-th column of Σ_1.

Note that $(\tilde{\mu}_{|X_i})_i = X_i$ (duh). Because all the entries are symmetric about 0, we'd expect that $(\tilde{\mu}_{|X_i})_j \leq X_i$ for $j \neq i$. In particular, we'd expect that $(\tilde{\mu}_{|X_i})_j \approx 0$ for j such that the $k_i(j) \approx 0$, i.e. they are uncorrelated. Of course, this is not true in general, uncorrelatedness need not imply independence, but when is it sorta true? In fact, this is not true in the single-pulse case, where far apart points can be anti-correlated.

We can also relate the conditional and unconditional covariances. Note,

$$\Sigma_1 = \mathbb{E}_{X|Y=1}\left[XX^\top\right] = \mathbb{E}_{X_i|Y=1}\left[\mathbb{E}_{X|X_i,Y=1}\left[XX^\top\right]\right] = \mathbb{E}_{X_i|Y=1}\left[\tilde{\Sigma}_{|X_i} + \tilde{\mu}_{|X_i}\tilde{\mu}_{|X_i}^\top\right],$$

so,

$$\mathbb{E}_{X_i|Y=1}\left[\tilde{\Sigma}_{|X_i}\right] = \Sigma_1 - \mathbb{E}_{X_i|Y=1}\left[\tilde{\mu}_{|X_i}\tilde{\mu}_{|X_i}^\top\right].$$

I don't know how much more we can say about $\tilde{\mu}_{|X_i}$ or $\tilde{\Sigma}_{|X_i}$ in general.

2.2 Gaussian Data

For Gaussian data, we can evaluate $f(w)$ exactly. Using simpler analytical methods, we get,

$$f(w) = \frac{1}{\sqrt{2\pi}} \cdot \frac{\Sigma_1 w}{\sqrt{w^\top \Sigma_1 w}}. \tag{9}$$

But how do we arrive at this result from our expression for $f(w)_i$ above, and does this elucidate why localization does not occur for Gaussian data? Starting from equation (6), we have

$$f(w)_i = \frac{1}{2}\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}\left[x_i \operatorname{erf}\left(\frac{w_ix_i + w_{\not i}^\top \mu_{|X_i=x_i}}{\sqrt{2w_{\not i}^\top \Sigma_{|X_i=x_i} w_{\not i}}}\right)\right] dx_i.$$

We need to express $\mu_{|X_i=x_i}$ and $\Sigma_{|X_i=x_i}$ in terms of x_i. Note that

$$\mu_{|X_i=x_i} = \left[(\Sigma_1)_{ij}\right]_{j\neq i} x_i,$$

and

$$\Sigma_{|X_i=x_i} = \left[(\Sigma_1)_{kj}\right]_{k,j\neq i} - \left[(\Sigma_1)_{ij}\right]_{j\neq i} \left[(\Sigma_1)_{ij}\right]_{j\neq i}^\top.$$

Importantly, this is independent of x_i. Note that we can write $w_i x_i + w_{\not i}^\top \mu_{|X_i=x_i} = w^\top \left[(\Sigma_1)_{ij}\right]_{j\in[n]} x_i$ (we noted this for the general case above as well). We will denote $\left[(\Sigma_1)_{ij}\right]_{j\in[n]}$ by k_i. So,

$$\begin{aligned}
f(w)_i &= \frac{1}{2}\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}\left[x_i \operatorname{erf}\left(\underbrace{\frac{k_i^\top w}{\sqrt{2w_{\not i}^\top \Sigma_{|X_i=x_i} w_{\not i}}}}_{\triangleq a} x_i\right)\right] dx_i \\
&= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-b^2x_i^2}\left[x_i \operatorname{erf}(ax_i)\right] dx_i && \text{where } b = \frac{1}{\sqrt{2}} \\
&= \frac{1}{\sqrt{2\pi}}\frac{a}{2b^2}(a^2+b^2)^{-\frac{1}{2}} && \text{via Ng \& Geller (1968)} \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{2b^2}\frac{k_i^\top w}{\sqrt{(k_i^\top w)^2 + 2b^2 w_{\not i}^\top \Sigma_{|X_i=x_i} w_{\not i}}} \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{(2b^2)^{\frac{3}{2}}}\frac{k_i^\top w}{\sqrt{w^\top \Sigma_1 w - (1-\frac{1}{2b^2})(k_i^\top w)^2}}.
\end{aligned}$$

Plugging in b,

$$f(w)_i = \frac{1}{\sqrt{2\pi}}\frac{k_i^\top w}{\sqrt{w^\top \Sigma_1 w}},$$

precisely as desired. The scaling term here *no longer depends on* k_i, which is what drives localization for the high-gain case. This happened because of the b term in the denominator, which corresponds to the inverse of the standard deviation of the Gaussian distribution. Could we tinker with b and still get localization? We would need to make b large, which would require decreasing the variance of the Gaussian. If we scale Σ_1, the covariance of X, to $\alpha\Sigma_1$, then b^{-2} would scale to αb^{-2}. Importantly, $(k_i^\top w)^2$ would scale to $\alpha^2(k_i^\top w)^2$. So, after scaling by α, we get

$$f(w)_i = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\alpha}}{(2b^2)^{\frac{3}{2}}}\frac{k_i^\top w}{\sqrt{w^\top \Sigma_1 w - \alpha(1-\frac{\alpha}{2b^2})(k_i^\top w)^2}}.$$

As $\alpha \to 0$, the localization term still disappears because of the extra power of α we pick up from k_i being squared. Thus, scaling down the variance still will not yield localization. Making α larger also won't yield localization, albeit differently. In this case, the $(k_i^\top w)^2$ dominates, but it does not have a negative sign, so as w gets more localized, $f(w)_i$ decreases, thus actively working against localization. (Also, the integral's expression is only valid when $|b| > |a|$, so messing with α too much will break this.)

2.3 Elliptical Data

We cannot solve $f(w)$ exactly for most distributions, but for data from an elliptical distribution, we can at least give f enough structure to understand its behavior. An elliptical distribution $\mathrm{El}(\mu, \Sigma, \psi)$

has characteristic function $\phi_{X-\mu}(t) = \psi(t^\top \Sigma t)$. This gives the property that if $X \sim \text{El}(\mu, \Sigma, \psi)$, then $c + AX \sim \text{El}(c + A\mu, A\Sigma A^\top, \psi)$. We can use this to conclude that $\langle w, X\rangle \sim \text{El}(\langle w, \mu\rangle, w^\top \Sigma w, \psi)$. We can write the density of $S = \langle w, X\rangle$ as

$$p(s) = k \cdot g\left(\frac{s^2}{w^\top \Sigma w}\right),$$

where k is a normalizing constant. Then,

$$\mathbb{E}_X\left[\text{ReLU}(\langle w, X\rangle)\right] = \mathbb{E}_S\left[\text{ReLU}(S)\right] = \int_0^\infty s p(s) ds = k \int_0^\infty s g\left(\frac{s^2}{w^\top \Sigma w}\right) ds.$$

So,

$$\begin{aligned} f(w) = \frac{\partial}{\partial w}\left[\mathbb{E}_X\left[\text{ReLU}(\langle w, X\rangle)\right]\right] &= 2k \int_0^\infty s \frac{\partial}{\partial w}\left[g\left(\frac{s^2}{w^\top \Sigma w}\right)\right] ds \\ &= -2k\left[\int_0^\infty s^3 g'\left(\frac{s^2}{w^\top \Sigma w}\right) ds\right] \frac{\Sigma w}{(w^\top \Sigma w)^2}. \end{aligned}$$

Setting $s = (\sqrt{w^\top \Sigma w})u$, we have $ds = (\sqrt{w^\top \Sigma w})du$, so

$$f(w) = -2k\left[\int_0^\infty u^3 (\sqrt{w^\top \Sigma w})^3 g'\left(u^2\right) du\right] \frac{\Sigma w}{(w^\top \Sigma w)^2} \propto \frac{\Sigma w}{\sqrt{w^\top \Sigma w}},$$

which is similar in form to the Gaussian case. Given this form, we can write the update in Fourier space, given by $u = P^\top w$, where P is the (real) DFT matrix.

$$\tau \frac{d}{dt} u = \frac{c}{2} \frac{\Lambda_1 u}{\sqrt{u^\top \Lambda_1 u}} - \frac{1}{4}(\Lambda_0 + \Lambda_1) u, \tag{10}$$

where the Λ_i are diagonal matrices of eigenvalues of Σ_i. Because this is diagonal, we can compute steady states, which lets us show that the limiting solutions are a superposition of just one or two modes (i.e. u is sparse)[5]. This is insufficient for localization, which requires u to be sparse in the *spatial* domain, not the Fourier domain.

How can we connect this to the representation in equation (7), and reconcile it with the apparent observation that marginal support on $\{\pm 1\}$ is sufficient for localization? Let us consider an elliptical distribution with marginal that has support on $\{\pm 1\}$. This would be achieved by $g(x) = \mathbb{1}(x = 1)$. Then, $[(\Sigma_1)_{ji}]_{j\in[n]} = \tilde{\mu}_{|X_i=1}$ and $\tilde{\Sigma}_{|X_i=1} = \Sigma_1 - \tilde{\mu}_{|X_i=1}\tilde{\mu}_{|X_i=1}^\top$. So, we just need to understand what Σ_1, the covariance, looks like as a function of Σ, the shape parameter (which need not be the covariance).

$$(\Sigma_1)_{ij} = \mathbb{E}_X\left[X_i X_j\right] = \int_{\mathbb{R}^n} k\delta(x^\top \Sigma^{-1} x - 1) x_i x_j dx$$

$$\begin{aligned} (\tilde{\Sigma}_{|X_i=1})_{jk} &= \mathbb{E}_{X|X_i=1}\left[(X_j - \mathbb{E}_{X|X_i=1}[X_j])(X_k - \mathbb{E}_{X|X_i=1}[X_k])\right] \\ &= \mathbb{E}_{X|X_i=1}\left[X_j X_k\right] - \mathbb{E}_{X|X_i=1}\left[X_j\right]\mathbb{E}_{X|X_i=1}\left[X_k\right] \end{aligned}$$

If $j = i$, then this is zero. If $j \neq i$, then,

$$(\tilde{\Sigma}_{|X_i=1})_{jk} = \mathbb{E}_{X|X_i=1}\left[X_k\right].$$

[5] I need to confirm this with simulations.

2.4 High-gain Data

In the high-gain setting, X_i has support on just $\{\pm 1\}$. So, we can write $\tilde{\mu}_{|X_i} = [(\Sigma_1)_{ij}]_{j\in[n]}$. Thus, $w^\top \tilde{\mu}_{|X_i} = e_i^\top \Sigma_1 w$ and $\tilde{\Sigma}_{|X_i} = \Sigma_1$. So,

$$f(w)_i \approx \Phi\left(\frac{e_i^\top \Sigma_1 w}{\sqrt{w^\top(\Sigma_1 - \tilde{\mu}_{|X_i}\tilde{\mu}_{|X_i}^\top)w}}\right) - \frac{1}{2} = \frac{1}{2}\operatorname{erf}\left(\frac{e_i^\top \Sigma_1 w}{\sqrt{2}\cdot\sqrt{w^\top\Sigma_1 w - (e_i^\top \Sigma_1 w)^2}}\right). \tag{11}$$

Intuition Let us momentarily ignore the effect of erf, since it just reduces the bias towards localization, and let's toss out the constant terms for good measure. Then, we can write f more simply as

$$f(w) \sim \left[\frac{1}{\sqrt{w^\top\Sigma_1 w - (e_i^\top \Sigma_1 w)^2}}\right]_{ii} \Sigma_1 w.$$

So, all the behavior driving localization is contained in the diagonal matrix in the middle. Recalling the circulant structure of Σ_1, let's write $e_i^\top \Sigma_1 w = k_1^\top \mathcal{C}^i\, w$, where k_1 is the kernel for Σ_1 centered at 0 and $\mathcal{C}$ is a circular shift up by 1. Then, we can rewrite the diagonal matrix in f as,

$$f(w) \sim \left[w^\top\Sigma_1 w - (k_1^\top \mathcal{C}^i\, w)^2\right]_{ii}^{-\frac{1}{2}} \Sigma_1 w.$$

As we vary i, the term $k_1^\top \mathcal{C}^i\, w$ compares the similarity of w to k_1 centered at i. So, it sweeps the kernel k_1 across w. As the similarity increases, the corresponding term in the diagonal matrix increases, weighting that entry of $\Sigma_1 w$ more heavily.

Then,

$$\tau\frac{d}{dt}u \sim \left[u^\top\Lambda_1 u - (k_1^\top \mathcal{C}^i\, Pu)^2\right]_{ii}^{-\frac{1}{2}} \Lambda_1 u - \frac{1}{2}(\Lambda_0 + \Lambda_1)u.$$

Setting each entry of this to 0, we have

$$\frac{1}{\sqrt{u^\top\Lambda_1 u - (k_i^\top Pu)^2}}\lambda_1^{(i)}u_i = \frac{1}{2}(\lambda_0^{(i)} + \lambda_1^{(i)})u_i.$$

If $u_i \neq 0$, then

$$\begin{aligned}
&\frac{1}{\sqrt{u^\top\Lambda_1 u - (k_i^\top Pu)^2}} = \frac{1}{2}\left(\frac{\lambda_0^{(i)}}{\lambda_1^{(i)}} + 1\right)\\
\iff &4\left(\frac{\lambda_0^{(i)}}{\lambda_1^{(i)}} + 1\right)^{-2} = u^\top\Lambda_1 u - (k_i^\top Pu)^2\\
\iff &4\left(\frac{\lambda_0^{(i)}}{\lambda_1^{(i)}} + 1\right)^{-2} + (k_i^\top Pu)^2 = u^\top\Lambda_1 u.
\end{aligned}$$

So, for all i such that $u_i \neq 0$,

$$u^\top\Lambda_1 u = 4\left(\frac{\lambda_0^{(i)}}{\lambda_1^{(i)}} + 1\right)^{-2} + (k_i^\top Pu)^2 = 4\left(\frac{\lambda_0^{(i)}}{\lambda_1^{(i)}} + 1\right)^{-2} + (k_i^\top w)^2 = w^\top\Sigma_1 w.$$

3 Signals on the hypercube

We will try to come up with some general sufficient conditions for localization. Let us introduce the data model of Alessandro as a starting point. There,

$$p_\xi(x) = \text{Law}(X), \quad X_i = \frac{1}{\sqrt{\mathcal{Z}(g)}} \text{erf}(gZ_i), \quad X \sim \mathcal{N}(0, \Sigma(\xi)), \tag{12}$$

where $g > 0$ is our gain parameter and $\mathcal{Z}(g)$ is a normalization constant that ensures $\text{Var}(X_i) = 1$ for all i. (Note that $\text{Cov}(X) \neq \Sigma(\xi)$, but it's pretty close.) Importantly, as $g \to 0$, $X \xrightarrow{d} Z$, i.e. the data is approximately Gaussian. However, as $g \to \infty$, X becomes supported on the vertices of the hypercube $\{\pm 1\}^n$. After staring at **??** for a while, I think that this is the key to understanding localization. I'll explain this below, but first, some analytical examples.

Single bump If $w = e_i$, then

$$f(w) = \Sigma e_i. \tag{13}$$

Balanced bumps If $w = e_i + e_j$, then

$$f(w) = \Sigma(e_i + e_j). \tag{14}$$

Imbalanced bumps If $w = \alpha e_i + e_j$ for $\alpha > 1$, then

$$f(w) = \Sigma e_i. \tag{15}$$

Interesting!

If we flip the sign of the smaller bump so that If $w = \alpha e_i - e_j$ for $\alpha > 1$, then

$$f(w) = \Sigma e_i. \tag{16}$$

Cool!

Three bumps If $w = \alpha e_i + \beta e_j + e_k$ for $\alpha > \beta > 1$, then

$$f(w) \approx \Sigma e_i. \tag{17}$$

More generally? Assume $w_1 > 0$.

$$\mathbb{1}(\langle w, x \rangle \geq 0) = \mathbb{1}(x_1 \geq -\sum_{i=2}^{n} (\tfrac{w_i}{w_1}) x_i).$$

If $\sum_{i=2}^{n} |\frac{w_i}{w_1}| < 1$, this is equivalent to $\mathbb{1}(x_1 \geq 0) = \mathbb{1}(x_1 \geq 1)$. This is a pretty strong condition on w that is not usually true. However, it gives us a starting point for how we might be able to generally cut away a lot of the complexity of f when the data is supported on the vertices of the hypercube.

Let's consider some separation point k, (recall we assume $|w_1| > \ldots > |w_n|$, but this is just to make the sums easier to write—we just need to partition the entries of w into two sets).

$$\mathbb{1}(\langle w, x \rangle \geq 0) = \mathbb{1}(\sum_{i=1}^{k} w_i x_i \geq -\sum_{i=k+1}^{n} w_i x_i).$$

What is the smallest positive value the LHS produces? Define

$$\delta \triangleq \min_{x\in\{\pm 1\}^k} \left| \sum_{i=1}^{k} w_i x_i \right|. \tag{18}$$

Then, if $\left|\sum_{i=k+1}^{n} w_i x_i\right| < \delta$ for all x, which in this case is equivalent to $\sum_{i=k+1}^{n} |w_i| < \delta$, we have

$$\mathbb{1}(\langle w, x\rangle \geq 0) = \mathbb{1}(\sum_{i=1}^{k} w_i x_i \geq 0).$$

We want $\mathcal{X}$ s.t. we can make k small in the following inequality.

$$\min_{x\in\mathcal{X}} \left| \sum_{i=1}^{k} w_i x_i \right| > \max_{x\in\mathcal{X}} \left| \sum_{i=k+1}^{n} w_i x_i \right|$$

Again, we're making *universal statements* about f without assuming the underlying probability distribution (other than that it's supported on the hypercube). In practice, we can do better than having x in equation (18) range across the entire k-dimensional hypercube, *considering instead some subset of it that occurs with high probability under* p. This would allow us to cut away even more of the complexity of f. Obviously, we'd want to consider the smallest k such that the condition above holds.

If k is sufficiently small, then we can cut away a lot of the complexity of f. More specifically, if w_i is sufficiently small, the indicator function treats it as if it were zero. However, we need to understand how the remaining terms, which cannot be treated like zero, affect the indicator function.

SDP

Equation (18) is related to the integer optimization problem

$$\delta_{\text{INT}} \triangleq \max_{x\in\{\pm 1\}^k} \sum_{i,j=1}^{k} A_{ij} x_i x_j,$$

where $A = -ww^\top$. This is because $\sum_{i,j=1}^{k} A_{ij} x_i x_j = (x^\top A x) = -(\langle w, x\rangle)^2$. So,

$$\delta_{\text{INT}} = \max_{x\in\{\pm 1\}^k} -(\langle w, x\rangle)^2 = -\min_{x\in\{\pm 1\}^k} (\langle w, x\rangle)^2 = -\sqrt{\delta}.$$

It would be cool to bound δ from below in terms of w. Grothendieck's inequality might let us do this, albeit with a rather loose bound.

3.1 Simulations

In the few examples above, we've been able to show analytically that, because X has support on the hypercube, f extracts the maximum value of w. This makes a lot of sense! But we'd like to show this in a more general setting.

Something like X_j is approximately independent of $\mathbb{1}(\langle w, X\rangle \geq 0)$ when j does not correspond to the maximum absolute entry in w. Otherwise, $X_j \approx \text{sgn}(w_j)$.

Let's consider the set of x

$$\mathbb{E}_X[\mathbb{1}(\langle w, x\rangle \geq 0)X] = \mathbb{E}_X\left[\sum_{x'\in\Theta} \mathbb{1}(X = x')X\right] \tag{19}$$

$$= \mathbb{E}_X\left[\sum_{x'\in\Theta}\left[\prod_{i=1}^{n} \mathbb{1}(X_i = x'_i)\right] X\right] \tag{20}$$

$$= \mathbb{E}_X\left[\sum_{x'\in\Theta}\left[\prod_{i=1}^{n} \frac{\operatorname{sgn}(x'_i)}{2}(X_i + x'_i)\right] X\right] \tag{21}$$

$$= \sum_{x'\in\Theta} \mathbb{E}_X\left[\left(\prod_{i=1}^{n} \frac{\operatorname{sgn}(x'_i)}{2}(X_i + x'_i)\right) X\right]. \tag{22}$$

Entrywise,

$$\sum_{x'\in\Theta} \mathbb{E}_X\left[\left(\prod_{i=1}^{n} \frac{\operatorname{sgn}(x'_i)}{2}(X_i + x'_i)\right) X_j\right] = \sum_{x'\in\Theta} \mathbb{E}_X\left[\left(\prod_{i=1}^{n} \frac{\operatorname{sgn}(x'_i)}{2}(X_i + x'_i)X_j\right)\right]$$

First, I want to understand how we could rigorously apply this intuition to simplify f. Let's start by defining the event

$$\Theta = \left\{x = (x_j)_{j\in[n]} \in \{\pm 1\}^n \mid \langle w, x\rangle \geq 0\right\}. \tag{23}$$

Then,

$$\mathbb{1}(\langle w, x\rangle \geq 0) = \sum_{x'\in\Theta} \mathbb{1}(x = x') = \sum_{x'\in\Theta}\prod_{j\in[n]} \mathbb{1}(x_j = x'_j) \tag{24}$$

So,

$$\mathbb{E}_{X|Y=1}\left[\mathbb{1}(\langle w, X\rangle \geq 0)X\right] = \sum_{j\in[n]} \mathbb{E}_{X|Y=1}\left[\Bigg(\underbrace{\sum_{x'\in\Theta} \mathbb{1}(X_j = x'_j)}_{\triangleq g_j(X)}\Bigg)X\right]. \tag{25}$$

With this perspective, we want to show

1. $g_j(x) \approx \mathbb{1}(x_j = \operatorname{sign}(w_j))$ when w_j is sufficiently large, and
2. $g_j(x) \approx 1$ when w_j is not sufficiently large.

We also want to understand what it means for w_j to be sufficiently large. Hopefully, we can show there is a pretty clear divide between the two cases, and that nothing falls in between. **Now, how do I do this?**

We'll start with the case where $w_j > 0$ for the sufficiently large case.

Let's consider the case where only one w_j is sufficiently large. WLOG, let's say this happens for $j = 1$. As we make it larger, how do the $g_j(x)$ change?

Note that

$$g_j(x) = \sum_{x'\in\Theta} \mathbb{1}(x_j = x'_j) \tag{26}$$

$$= |\{x' \in \Theta \mid x'_j = 1\}|\mathbb{1}(x_j = 1) + |\{x' \in \Theta \mid x'_j = -1\}|\mathbb{1}(x_j = -1) \tag{27}$$