1 Model

We begin with a general model that encapsulates many of the works we will discuss. For an input $\mathbf{x}$, we want to predict a response $\mathbf{y}^*$. Our prediction is a function f_θ of $\mathbf{x}$. The function f_θ is parameterized by θ. We will consider models of the form

$$f_\theta(\mathbf{x}) = y_\theta\left(\sum_i \alpha_\theta^{(i)} \phi_\theta^{(i)}(\mathbf{x})\right). \tag{1}$$

Here, $\mathbf{x}$ is an L-dimensional real vector and $\mathbf{y}^*$ is an M-dimensional real vector. The ϕ_i are basis functions that map from $\mathbb{R}^L$ to $\mathbb{R}^K$, and the α_i are real scalars. Additionally, the sum over i need not be finite. The coefficients α_i may depend on $\mathbf{x}$ and our parameters θ, but we suppress the former dependence for notational simplicity. The function y_θ maps from $\mathbb{R}^K$ to $\mathbb{R}^M$ and is parameterized by θ as well.

We will consider a penalized mean squared error loss function

$$\mathcal{L}(\theta) = \frac{1}{2}\,\mathbb{E}_{\mathbf{x},\mathbf{y}^*}\left[\|\mathbf{y}^* - f_\theta(\mathbf{x})\|_2^2\right] + \lambda p(\{\alpha^{(i)}\}_i, \theta). \tag{2}$$

Our penalty function p serves to induce sparseness in the coefficients α_i, and perhaps also regularize the parameters θ.

We will show how this model encapsulates the works we are interested in understanding.

2 Ingrosso et al. (2022)

The model in Ingrosso et al. (2022) sets $M = 1$, uses linear basis functions, and sets y_θ to be the mean function after applying a nonlinearity. That is,

$$\sum_i \alpha_\theta^{(i)} \phi_\theta^{(i)}(\mathbf{x}) = \Theta\mathbf{x} + b_\theta \tag{3}$$

$$y_\theta(\mathbf{x}) = \frac{1}{K}\mathbf{1}^\top \sigma(\mathbf{x}). \tag{4}$$

Here, Θ is a $K \times L$ matrix, b_θ is a K-dimensional vector, and σ is a nonlinear function applied elementwise. It seems that in much of their work, b_θ is fixed at -1 or 0. I still need to confirm experimentally that this does not affect the results.

In this work, we also ignore the penalty term. I believe that our gradient update is of the form

$$a \tag{5}$$

3 Data Model

We assume that our data is generated by the following model. It is paramaterized by the length of the input, $L \in \mathbb{N}$. It is also parameterized by a scale parameter $\xi \leq L$. We construct data $\{X_l\}_l \subseteq \{0,1\}^L$ as follows:

1. Sample integers $l^* \sim \text{Uniform}[1, L]$ (starting position) and $T \sim \text{Uniform}[1, \xi)$ (length of pulse).
2. For $0 \leq i \leq T$, set $X_{l^*+i \pmod{L}} = 1$, and set all other X_l to 0.
3. Return the sequence $\{X_l\}_l$.

Now, we derive the conditional probability $p_{11} \triangleq \mathbb{P}(X_a = 1 \mid X_b = 1)$. For now, assume $d \triangleq b - a > 0$. If $d \geq \xi$, then $p_{11} = 0$. So, assume $d < \xi$.

Now, we count the number of values of T that result in both X_a and X_b being in the pulse for a given l^*. WLOG, assume $a = 0$. For $1 \leq l^* \leq d$, the range of values for T that results in both X_a and X_b being in the pulse is given by

$$L - l^* \leq T < \xi. \tag{6}$$

Thus, the number of values for T in this case is $\max(\xi - L + l^*, 0)$.

For $d < l^* \leq L$, the values of T that result in both X_a and X_b being in the pulse is given by

$$L - (l^* - d) \leq T < \xi. \tag{7}$$

So, the number of values for T in this case is $\max(\xi - (L - (l^* - d)), 0) = \max(\xi - L + l^* - d, 0)$.

Note that the first max condition is at least zero when $l^* \geq L - \xi$. This yields a sum over $\max(L - \xi, 1) \leq l^* \leq d$. Similarly, the second max condition is at least zero when $l^* \geq L - \xi + d$. This yields a sum over $\max(L - \xi + d, d + 1) \leq l^* \leq L$. Note that the first term dominates in both of these new max statements when $\xi \leq L - 1$. So, let us assume this is the case.

Now, we want to find the total number of values of T that result in both X_a and X_b being in the pulse.

$$T_1 = \sum_{L-\xi \leq l^* \leq d} (\xi - L + l^*) \tag{8}$$

$$T_2 = \sum_{L-\xi+d \leq l^* \leq L} (\xi - L + l^* - d) = \sum_{L-\xi \leq t \leq L-d} (\xi - L + t). \tag{9}$$

We compute T_1 as,

$$T_1 = \left[(\xi - L)(d - L + \xi + 1) + \frac{d(d+1)}{2} - \frac{(L-\xi)(L-\xi-1)}{2}\right] \mathbb{1}(d \geq L - \xi).$$

Next, we compute T_2 as,

$$T_2 = \left[(\xi - d + 1)(\xi - L) + \frac{(L-d)(L-d+1)}{2} - \frac{(L-\xi)(L-\xi-1)}{2}\right] \mathbb{1}(d \leq \xi).$$

Thus, the total number of values of T that result in both X_a and X_b being in the pulse is

$$T_1 + T_2 = \begin{cases} (\xi - d + 1)(\xi - L) + \frac{(L-d)(L-d+1)}{2} - \frac{(L-\xi)(L-\xi-1)}{2} & d \leq \xi, L - \xi + 1 \\ d^2 - dL + \frac{L^2}{2} + \frac{L}{2} - (L - \xi)(\xi + 1) & L - \xi \leq d \leq \xi \\ (\xi - L)(d - L + \xi + 1) + \frac{d(d+1)}{2} - \frac{(L-\xi)(L-\xi-1)}{2} & d \geq L - \xi, \xi + 1 \end{cases}$$

There are $\xi - 1$ possible values for T and L values for l^*. Recall we sample T and l^* uniformly and independently. Thus, the probability that $X_b = 1$ given $X_a = 1$ is

$$p_{11} = \frac{T_1 + T_2}{(\xi - 1)L} = \frac{d^2 - dL + \frac{L^2}{2} + \frac{L}{2} - (L - \xi)(\xi + 1)}{(\xi - 1)L}. \tag{10}$$

Note that we assumed $\xi \leq L - 1$. However, we can check that the above expression is still valid when $\xi = L$. (Check Desmos. Otherwise, an exercise left to the reader.) It also satisfies the "sanity check" that it is minimized at $x = \frac{L}{2}$.