Bayesian-Optimal Multi-Classification with Noisy Input Necessitates General Input Space Representation

Aman Bhargava

December 1, 2023

1 Introduction

Here we analyze the latent representations in Bayesian filter models trained to perform multi-class classification on some ground truth input vector $\mathbf{x}^* = [x_1^*, x_2^*]^\top$ based on noisy discrete-time measurement signals $\mathbf{X}(t) = [X_1(t), X_2(t)]^\top$ defined as

$$X_1(t) = x_1^* + \eta\mathcal{N}(0,1) \tag{1}$$
$$X_2(t) = x_2^* + \eta\mathcal{N}(0,1) \tag{2}$$

The multi-classification task may be defined in terms of N classification boundary angles which we collectively denote as Θ:

$$\Theta = \{\alpha_i : \alpha_i \in [-\pi, \pi], i = 1, \ldots, N\} \tag{3}$$

as in Figure 1. Thus models must predict the ground truth label $\mathbf{y}^* = [y_1^* \ldots y_N^*]^\top$ corresponding to some $\mathbf{x}^*$ based on the resulting noisy measurement signals $\mathbf{X}(t)$ where

$$y_i^* = \begin{cases} +1 & \text{if } x_2^* > x_1^* \tan \alpha_i \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

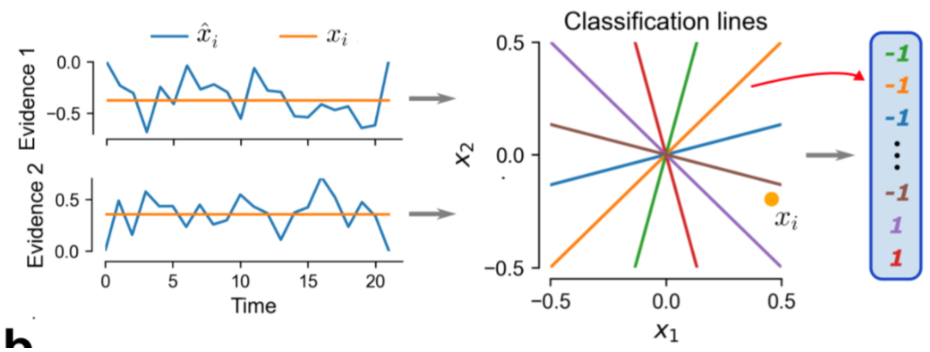

Figure 1: **Multitasking RNN learns abstract representations.** Data generating process. The task is to simultaneously report whether the true joint evidence (x_1, x_2) (yellow dot) lies above $(+1)$ or below (-1) a number of classification lines (here 6).

Contribution: In this document, I demonstrate that a Bayesian-optimal multi-classifier with noisy random input $\mathbf{X}(t)$ and classification estimate output $\hat{\mathbf{y}}(t) \in [-1, 1]^N$ must form latent representations $\mathbf{z}(t)$ that retain the 2-dimensional structural information of the input data space $[x_1^*, x_2^*] \in \mathbb{R}^2$ in the limit as $N \to \infty$ for evenly spaced $\alpha_i \in \Theta$. Intriguingly, without noise, we are unable to guarantee that optimal latent representations $\mathbf{z}(t)$ will retain sufficient information to estimate $\mathbf{x}^*$. While the proof is stated for 2-dimensional input, the argument should hold for any input dimensionality.

1.1 Bayesian Filtering Framework

Bayesian filters are a class of statistical models and algorithm that update a latent state based on noisy and uncertain observation signals. Rooted in principles of Bayesian inference, these filters combine aggregated "knowledge", represented by a latent state $\mathbf{Z}(t)$, with incoming observations $\mathbf{X}(t)$ to continually update the latent state to facilitate some prediction of some output $\mathbf{Y}(t) = f(Z(t))$.

Definition 1 (Bayesian Filter Operation)**.** *A discrete-time Bayesian filter updates latent variable $\mathbf{z}(t)$ based on incoming data $\mathbf{x}(t)$ by applying Bayes'* *theorem:*

$$P\big(\mathbf{z}(t)|\mathbf{x}(t), \mathbf{z}(t-1)\big) = \frac{P\big(\mathbf{x}(t)|\mathbf{z}(t), \mathbf{z}(t-1)\big)P\big(\mathbf{z}(t)|\mathbf{z}(t-1)\big)}{P\big(\mathbf{x}(t)|\mathbf{z}(t-1)\big)} \tag{5}$$

$$\propto P\big(\mathbf{x}(t)|\mathbf{z}(t)\big)P\big(\mathbf{z}(t)|\mathbf{z}(t-1)\big) \tag{6}$$

Bayesian filters are commonly equipped with a "decoder" or "readout map" f which maps latent $\mathbf{Z}(t)$ to readout estimation $\hat{\mathbf{Y}}(t) = f(\mathbf{Z}(t))$.

There is a deep structural similarity between RNNs and Bayesian filters, as both models update some latent state $\mathbf{z}(t)$ based on incoming datum $\mathbf{x}(t)$. Moreover, RNNs and Bayesian filters are both frequently used to predict some value $\mathbf{y}(t) = f(\mathbf{z}(t))$ (citation: Goodfellow for RNN, Bayesian inference textbook for filters). We leverage the structure in the Bayesian filter formulation to prove our main result in Section 2.

2 Main Results

Consider an optimal Bayesian filter for the multi-class classification task on noisy discrete time measurement signals.

Theorem 1. *An optimal Bayesian filter trained to perform multi-class classification on ground truth input $\mathbf{x}^*$ w.r.t. decision boundaries $\Theta = \{\alpha_1, \ldots, \alpha_N\}$ based on noisy measurement signals $\mathbf{X}(t)$ must have latent state $\mathbf{Z}(t)$ that retains a representation of the 2-dimensional input vector $\mathbf{x}^*$ in the limit as $N \to \infty$.*

Proof.

Lemma 1 (Equivalence to Angle Estimation)**.** *In the limit as $N \to \infty$ for uniformly distributed decision boundaries in $\Theta = \{\alpha_i\}_{i \in [N]}$, the multi-classification task of estimating $\mathbf{y}^*$ (Equation 3) for a given $\mathbf{x}^*$ given noisy observations $\mathbf{X}(t)$ is equivalent to estimating the angle $\theta = \angle \mathbf{x}^*$.*

Lemma 2 (Angle Estimation Requires Magnitude Estimation)**.** *An optimal Bayesian filter predicting the angle of some ground truth input $\angle \mathbf{x}^*$ based on noisy observations $\mathbf{X}(t)$ must implicitly estimate both the angle and the magnitude of $\mathbf{x}^*$ within its latent $\mathbf{Z}(t)$.*

Proof. Recall that the update rule latent $\mathbf{Z}(t)$ in Definition 1. For Bayesian-optimal estimation of θ using latent $\mathbf{Z}(t)$ with readout map f where $\hat{\theta}(t) = f(\mathbf{Z}(t))$, the likelihood term $P\big(\mathbf{X}(t)|\mathbf{Z}(t)\big)$ must be accurate. Therefore, the latent representation must contain not only angle information on $\mathbf{x}^*$, but also magnitude information, as this information can be leveraged to narrow down the likelihood term, thus optimizing the estimation. $\square$

Therefore, performing optimal multi-class classification in the limit as the number of tasks $N \to \infty$ with uniformly distributed boundary angles $\alpha_i \in \Theta$ is equivalent to estimating the angle of the ground truth $\mathbf{x}^*$ based on noisy $\mathbf{X}(t)$ (Lemma 1). Bayesian filtering for angle estimation requires that $Z(t)$ to maximally constrain the likelihood $P(\mathbf{X}(t)|\mathbf{Z}(t))$ to optimally weight the contribution of the new data $\mathbf{X}(t)$ in the new angle estimate $\hat{\theta} = f(Z(t))$.

Thus we have demonstrated that both the angle information and the magnitude information of $\mathbf{x}^*$ must be implicitly estimated by an optimal Bayesian filter in multi-class classification problem Θ with non-zero noise η in the measurements $\mathbf{X}(t)$. $\square$

3 Conclusion