# [UltraReps] Bayesian-Optimal Multi-Classification with Noisy Input Necessitates General Input Space Representation

Aman Bhargava

January 11, 2024

## 1 Introduction

*Notation: lower case variables denote scalars (e.g., $x$), upper case variables denote random variables (e.g., $X$), and boldfaced variables denote vector quantities (e.g., $\mathbf{x}, \mathbf{X}$). We denote the $d \times d$ identity matrix as $\mathbf{I}_d$.*

Here we analyze the latent representations in optimal Bayesian filter models trained to perform multi-classification on some ground truth input vector $\mathbf{x}^* \in \mathbb{R}^d$ based on noisy discrete-time measurement signals $\mathbf{X}(1), \dots, \mathbf{X}(t)$ for i.i.d. $\mathbf{X}(i) = \mathbf{x}^* + \mathcal{N}(\mathbf{0}, \mathbf{I_d})$. Classification boundaries are denoted $\{(\mathbf{c}_i, b_i)\}_{i=0}^M$ where $\mathbf{c}_i \in \mathbb{R}^d$ is the normal vector to the hyperplane and $b_i \in \mathbb{R}$ is the offset from the origin. For boundary $i$, the separating plane defined as the affine set $\{\mathbf{x} | \mathbf{c}_i^\top \mathbf{x} = b_i\}$. The ground truth classification of a given point $\mathbf{x}$ is given by decision rule $y(\mathbf{x})$:

$$y_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{c}_i^\top \mathbf{x} > b_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A restricted form of the problem is depicted in Figure 1. Note that we are interested in the general case where the hyperplanes do not necessarily pass through the origin.

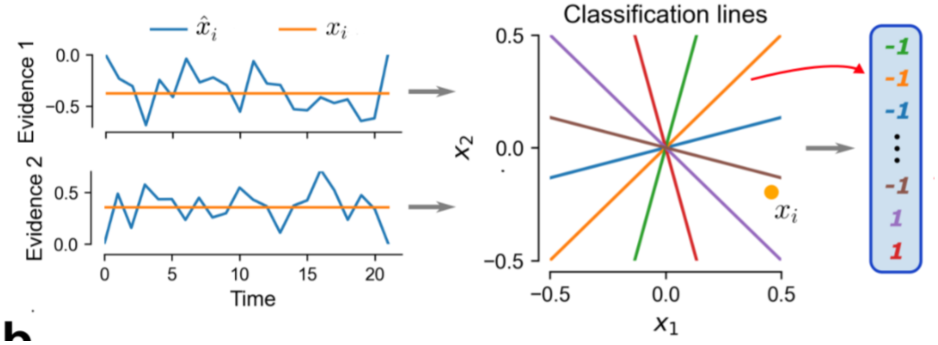

Figure 1: **Multitasking RNN learns abstract representations.** Data generating process. The task is to simultaneously report whether the true joint evidence $(x_1, x_2)$ (yellow dot) lies above $(+1)$ or below $(-1)$ a number of classification lines (here 6).

## 2 Bayesian Filtering Framework

We are interested in the properties of optimal Bayesian filter models that sequentially process inputs $\mathbf{X}(t)$ to estimate each classification $y_i(\mathbf{x}^*), i \in [M]$. 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 $\mathbf{Y}(t) = f(\mathbf{Z}(t))$.

**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:*

$$P\big(\mathbf{z}(t)|\mathbf{x}(t), \mathbf{z}(t-1)\big) = \frac{P\big(\mathbf{x}(t)|\mathbf{z}(t), \mathbf{z}(t-1)\big)P\big(\mathbf{z}(t)|\mathbf{z}(t-1)\big)}{P\big(\mathbf{x}(t)|\mathbf{z}(t-1)\big)} \tag{2}$$

$$\propto P\big(\mathbf{x}(t)|\mathbf{z}(t)\big)P\big(\mathbf{z}(t)|\mathbf{z}(t-1)\big) \tag{3}$$

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

The readout $\hat{\mathbf{Y}}(t)$ is a vector of Bernoulli random variables with index $i$ corresponds to the estimate of classification $(\mathbf{c}_i, b_i)$.

# 3 Main Result

In this section we show that an optimal Bayesian filter performing multi-classification as described in Section 1 must store a representation of $x^*$ in the latent variable $\mathbf{Z}(t)$. More specifically, we show that the maximum likelihood estimate of $x^*$ based on $\mathbf{X}(1), \dots, \mathbf{X}(t)$ is recoverable from $\mathbf{Z}(t)$. We will start by investigating estimator $P(\hat{\mathbf{Y}}(t)|\mathbf{X}(1), \dots, \mathbf{X}(t))$, assuming that optimal $\mathbf{Z}(t)$ retains all relevant information from $\mathbf{X}(1), \dots, \mathbf{X}(t)$, and thus $P(\hat{\mathbf{Y}(t)}|\mathbf{Z}(t)) = P(\hat{\mathbf{Y}(t)}|\mathbf{X}(1), \dots, \mathbf{X}(t))$.

## 3.1 Single Decision Boundary

First, we will consider the Bayesian filtering of a single classification output $y(x^*)$ from noisy inputs $X(1), \dots, X(n)$ with decision boundary parameters $(\mathbf{c}, b)$. We begin by scaling our coordinates such that the Gaussian noise has unit variance:

$$\mathbf{X}(t) = \mathbf{x}^* + \mathcal{N}(\mathbf{0}, \mathbf{I_d}) \tag{4}$$

The problem is further simplified by setting $b = 0$. We are interested in $P(\hat{Y}(t)|\mathbf{X}(1), \dots, \mathbf{X}(t))$ where $\hat{Y}(t)$ is a Bernoulli random variable representing the probability that $y(\mathbf{x}^*) = 1$ given evidence $\mathbf{X}(1), \dots, \mathbf{X}(t)$.
Since $y(\mathbf{x}^*)$ is a deterministic function of non-random variable $x^*$, we will first derive the probability distribution over $\mathbf{x}^*$ (denoted $\hat{\mathbf{X}}$) to determine $y(\hat{\mathbf{X}})$.

**Lemma 1.** *For $\mathbf{X}(t) = \mathbf{x}^* + \mathcal{N}(\mathbf{0}, \mathbf{I_d})$ and no prior on $x^*$, the conditional probability distribution $\hat{\mathbf{X}} = P(x^*|\mathbf{X}(t), \dots, \mathbf{X}(t))$ is given by*

$$\hat{\mathbf{X}} = \mathcal{N}(\hat{\mu}, \hat{\Sigma}) \tag{5}$$

*where $\hat{\mu}$ is the mean of $\mathbf{X}(1), \dots, \mathbf{X}(t)$ and $\hat{\Sigma} = t^{-1/2}\mathbf{I}_d$.*

*Proof.* Since $\mathbf{X}(1), \dots, \mathbf{X}(t)$ are i.i.d. from a Gaussian distribution with mean $\mathbf{x}^*$ and identity covariance, the sample mean is known to be distributed normally centered at the ground truth $\mathbf{x}^*$. We apply the known standard deviation of the underlying distribution (identity covariance) to arrive at $\hat{\Sigma} = t^{-1/2}\mathbf{I}_d$ as the standard deviation on the sample mean (derived from the central limit theorem). □

We can use estimator $\hat{\mathbf{X}}$ to construct $\hat{\mathbf{Y}}$ by expanding $\hat{\mathbf{Y}} = y(\hat{\mathbf{X}})$ as

$$y(\hat{\mathbf{X}}) = \begin{cases} 1 & \text{if } \mathbf{c}^\top \hat{\mathbf{X}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In essence, we we are interested in the amount of the probability density of $\hat{\mathbf{X}}$ that lies on each side of the decision boundary. Deriving this probability is simplified by the fact that $\hat{\mathbf{X}}$ is isotropic – i.e., it inherits the spherical covariance of the underlying data generation process.

**Lemma 2.** *Consider Gaussian-distributed d-dimensional random variable $\hat{\mathbf{X}}$ with isotropic covariance $\Sigma = t^{-1/2}\mathbf{I}_d$ and mean $\mu \in \mathbb{R}^d$. The probability density of $\hat{\mathbf{X}}$ on each side of the positive side of the decision boundary $\{\mathbf{x} : \mathbf{c}^\top \mathbf{x} > 0\}$ can be expressed as*

$$\Pr\{\mathbf{c}^\top \mathbf{x} > 0\} = \Phi(k\sqrt{t}) \tag{7}$$

*where $\Phi$ is the CDF of the normal distribution and $k = (\mathbf{c}^\top \mu)/\|\mathbf{c}\|$ is the signed projection distance between the plane and the mean of $\hat{\mathbf{X}}$.*

*Proof.* Since the $\hat{\mathbf{X}}$ is isotropic, the variance on every axis is equal and independent. We may rotate our coordinate system such that the projection line between the plane and the mean of $\hat{\mathbf{X}}$ aligns with an axis we denote as "axis 0". The rest of the axes must be orthogonal to the plane. Since each component of an isotropic Gaussian is independent, the marginal distribution of $\hat{\mathbf{X}}$ on axis 0 is a univariate Gaussian with variance $t^{-1/2}$ mean at distance $k$ from the boundary. Equation 7 applies the normal distribution CDF $\Phi$ to determine the probability mass on the positive side of the boundary. □

Thus we are able to construct our optimal estimator $\hat{Y}$ for $P(y(\mathbf{x}^*)|X(1), \ldots, X(t))$ using lemma 1 as

$$\hat{Y} = \begin{cases} 1 & \text{with probability } \Phi\big(\frac{(\mathbf{c}^\top \mu)}{\|\mathbf{c}\|}\sqrt{t}\big) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Observe that the probability that $\hat{Y} = 1$ in Equation 8 **monotonically scales** with the signed distance between the hyperplane and $\mu$ (CDFs are monotonic).

**Lemma 3.** *Knowledge of time (number of samples) $t$ and the "success probability" for Bernoulli random variable $\hat{Y}$ as defined in Equation 8 is sufficient to determine the projection distance between $\mu = mean\big(X(1), \ldots, X(n)\big)$ and the decision boundary.*

*Proof.* Recall Equation 7 from Lemma 2. We may solve for projection distance $k$ separating the decision boundary and the mean $\mu$ of observations $\mathbf{X}(0), \ldots, \mathbf{X}(t)$ as

$$k = \frac{1}{\sqrt{t}} \Phi^{-1}(\Pr\{\hat{Y} = 1\}) \tag{9}$$

Since $\Phi$ is the CDF of the normal distribution, and the normal distribution is not zero except at $\pm\infty$, the inverse $\Phi^{-1}$ is well-defined. □

## 3.2 Translateration via Multiple Decision Boundaries

**To recap Section 3.1** : We derived an optimal estimator of $\mathbf{x}^*$ (denoted $\hat{\mathbf{X}}$) based on noisy i.i.d. measurements $\mathbf{X}(1), \ldots, \mathbf{X}(t) \sim \mathcal{N}(\mathbf{x}^*, \mathbf{I}_d)$ in Lemma 1. In Lemma 2 we derived the equation for Bernoulli variable $\hat{Y}$ to estimate $y(\mathbf{x}^*)$ based on the same noisy measurements via $\hat{\mathbf{X}}$. Finally, we showed in Lemma 3 that the uncertainty in $\hat{Y}$ and the time $t$ is sufficient to determine the projection distance between the decision boundary and $\mu = \text{mean}(X(0), \ldots, X(t))$ via Equation 9.

We now have the tools to prove our final result.

**Theorem 1.** *Consider the problem statement from Section 1 and the Bayesian filter from Section 2 with Bernoulli random variable vector readout $\hat{\mathbf{Y}}$.*

*Let $\mathbf{C}$ be the matrix of decision boundaries with row vectors $\mathbf{c}_i$ (cf Equation 1). If $\mathbf{C}$ is full-rank, then $\hat{\mathbf{Y}}$, $t$, and $\mathbf{C}$ are sufficient to reconstruct the exact value of $\mu$, the mean of $\mathbf{X}(1), \ldots, \mathbf{X}(t)$, which is also the optimal estimator for $\mathbf{x}^*$.*

*Proof.* We may prove this claim by providing an algorithm to reconstruct $\mu = \text{mean}(\mathbf{X}(1), \ldots, \mathbf{X}(t))$. Invoke Lemma 3 to compute the signed projection distance between $\mu$ and each decision plane $\mathbf{c}_i$. Let $\mathbf{k} = [k_1, \ldots, k_M]^\top$ where each $k_i$ corresponds to decision boundary $\mathbf{c}_i$. Assume each $c_i$ are normalized. Then the mean $\mu$ must satisfy

$$\mathbf{C}\mu = \mathbf{k} \tag{10}$$

Thus, for full rank $\mathbf{C}$, we will have a uniquely determined $\mu$ value. □

### 3.3 Data Processing Inequality

**Corollary 1.** *For any system implementing optimal Bayesian filtering to estimate $\hat{Y}(t)$ based on latent $\mathbf{Z}(t)$ and inputs $\{\mathbf{X}(0), \ldots, \mathbf{X}(t)\}$, the latent variable $Z(t)$ must necessarily encode a representation of the mean $\mu$ of optimal $\mathbf{x}^*$ estimator $\hat{\mathbf{X}}$, where $\hat{\mathbf{X}} \sim \mathcal{N}(\mu, \Sigma)$, and*

$$\begin{aligned} \mu &= mean\left(\mathbf{X}(1), \ldots, \mathbf{X}(n)\right), \\ \Sigma &= \hat{\Sigma} = t^{-\frac{1}{2}}\mathbf{I}_d. \end{aligned}$$

*Proof.* This follows from the data processing inequality, which in this context implies that if

$$X^* \to \{X(1), \ldots, X(n)\} \to \mathbf{Z}(t) \to \hat{\mathbf{Y}}(t) \to \hat{\mathbf{X}}(t) \tag{11}$$

forms a directed Markov chain, then the mutual information between $I\big(\hat{\mathbf{Y}}(t), \hat{\mathbf{X}}(t)\big)$ cannot exceed the mutual information between $I\big(\mathbf{Z}(t), \hat{\mathbf{X}}(t)\big)$. We recall that $\hat{\mathbf{X}}(t)$ is a deterministic function of $\hat{\mathbf{Y}}(t)$ and time $t$, so $I\big(\hat{\mathbf{Y}}(t), \hat{\mathbf{X}}(t)\big) = H(\hat{\mathbf{X}}(t))$. Since the mutual information between the latent $\mathbf{Z}(t)$ and $\hat{\mathbf{X}}(t)$ is at least as great as that between $\hat{\mathbf{Y}}(t)$ and $\hat{\mathbf{X}}(t)$, $\mathbf{Z}(t)$ also must contain a representation of $\hat{\mathbf{X}}(t)$. □