Bayesian-Optimal Multi-Classification implies Abstract Representations (Theory)

Aman Bhargava

March 20, 2024

This document outlines the theoretical result from the paper "Disentangling Representations in RNNs through Multi-task Learning" by Pantelis Vafidis, Aman Bhargava, Antonio Rangel (coming soon to a preprint server near you!)

Consider an intelligent agent making decisions in some environment. The agent receives noisy observations conditioned on the environment state, and must produce optimal decisions (i.e., learn a multi-classification objective). We show that **the agent must represent an estimate of the de-noised environment state coordinate if it optimally estimates decision output from noisy observations**. I thought this was rather surprising, as it generalizes to to any optimal estimator. We focus on proving that optimal classification estimation implies state estimation in the case of linear decision boundaries in the multi-classification objective. We conclude with a discussion of the implications for non-linear decision boundaries via linear approximation and the introduction of non-linear observation maps.

1 General, Non-Linear Problem Statement

Noisy Multi-Classifier: Formalize the "environment state" as $X \sim P(X)$ with sample space $\mathcal{X}$ and a corresponding ground truth decision set $P(Y_i|X)$ for $i \in [N]$ (e.g., multi-classification on the environment state). Denote the i.i.d. noise process $X_i \sim P(\tilde{X}_i|X)$ from which observations $\tilde{X}_i$ are sampled. We consider optimal estimators of the ground truth readout Y given noisy measurements $\tilde{X}$ denoted $P(\hat{Y}|\tilde{X}_1, \ldots, \tilde{X}_T)$.

$$X \xrightarrow{\text{noise}} \{\tilde{X}_t\}_{t\in[T]} \xrightarrow{\text{agent}} \{\hat{Y}_i\}_{i\in[N]} \qquad (1)$$
$$\searrow \{Y_i\}_{i\in[N]}$$

Geometry: Let X reside in a metric space $\mathcal{X}$. Let each Y_i be defined in terms of a binary discriminator $\phi_i : \mathcal{X} \to \{0, 1\}$. Let the equivalence classes of $\mathcal{X}$ under each discriminator ϕ_i be connected (i.e., $\{x|\phi_i(x) = 1, x \in \mathcal{X}\}$ is connected for each ϕ_i).

Linearity: We will begin by presenting a proof on the case where the decision boundaries are linear. Note that this may be readily generalized to non-linear decision boundaries by considering the linear approximation of the decision boundaries in the vicinity of the true environment state, or by inserting an injective non-linear map from some abstract coordinate space $\mathcal{X}^*$ to observation space $\mathcal{X}$.

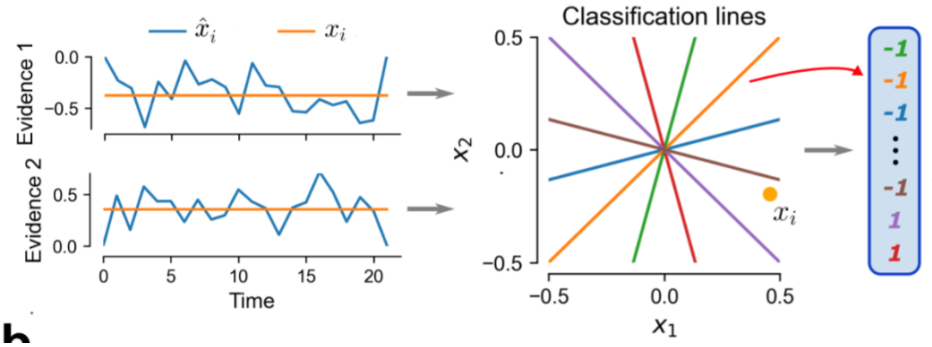

Figure 1: **Multitasking RNN learns abstract representations.** Data generating process. The task is to simultaneously report whether the true joint evidence (x_1, x_2) (yellow dot) lies above $(+1)$ or below (-1) a number of classification lines (here 6).

We seek to understand the properties of optimal multi-classifiers in the paradigm illustrated in Figure 2 where an estimate of classification certainty is made. We denote the set of classification estimates as $\hat{\mathbf{Y}}$, a vector of

Bernoulli random variables. We prove that **any optimal multi-task classifier** with i.i.d. noisy inputs $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ **implicitly estimates the ground truth coordinate $\mathbf{x}^*$ in its latent state $\mathbf{Z}(t)$.**

$$\mathbf{x}^* \rightarrow \{\mathbf{X}(t)\} \rightarrow \mathbf{Z}(t) \rightarrow \hat{\mathbf{Y}}(t) \tag{2}$$

Specifically, we prove that $\mathbf{Z}(t)$ represents the optimal estimate of $\mathbf{x}^*$ given the noisy measurements as long as the classification boundary normal vectors span the input space. For random decision boundaries in D-dimensional input space, we would expect D classification boundaries to satisfy this condition.

This result holds for *any* system that performs optimal multi-task classification with a latent variable separating the inputs from the outputs (e.g., RNNs, Bayesian filters, etc.), **regardless of the internal dynamics of the latent state**.

2 Trilateration Theorem for Linear Decision Boundaries

Presented in Appendix B of "Disentangling Representations in RNNs through Multi-task Learning"

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 $\mathbf{I}_d$.*

Variable Glossary:

- $\mathbf{x}^* \in \mathbb{R}^D$: Ground truth (un-noised) input variable.
- $\mathbf{X}(t) \sim \mathbf{x}^* + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ are i.i.d. noisy measurements of $\mathbf{x}^*$, where
  - σ is the amount of equivariant Gaussian noise, and
  - t is the discrete time index within a trial.
- N_{task} is the number of classification tasks,

- $\{(\mathbf{c}_i, b_i)\}_{i=1}^{N_{task}}$ are the classification boundary normal vectors and offsets respectively, with $\mathbf{c}_i \in \mathbb{R}^D$ and $b_i \in \mathbb{R}$. We assume each $\|\mathbf{c}_i\| = 1$.
- $(\mathbf{C}, \mathbf{b})$ are a matrix and vector representing each of the N_{task} classification tasks where $\mathbf{C} \in \mathbb{R}^{N \times D}$
- $\mathbf{y}(\mathbf{x}^*) \in \{-1, +1\}^{N_{task}}$: Ground truth classification outputs, where each ground truth classification $y_i(\mathbf{x}^*)$ is given by

$$y_i(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{c}_i^\top \mathbf{x} > b_i \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

- $\mathbf{Z}(t)$: Latent variable of a multi-task classification model, conditional on $\mathbf{X}(1), \ldots, \mathbf{X}(t)$.
- $\hat{\mathbf{Y}}(t) \in [0, 1]^{N_{task}}$: Output vector of the multi-task classification model at time t, where each $\hat{Y}_i(t)$ is a Bernoulli random variable parameter representing the conditional probability $\Pr\{y_i(\mathbf{x}^*) = +1\}$ given the noisy observations (via latent variable $\mathbf{Z}(t)$ – see graphical model in Equation 4).
- $\hat{\mathbf{X}}(t) = \mathcal{N}(\mu(t), \Sigma(t))$: Optimal estimate of $\mathbf{x}^*$ given measurements $\mathbf{X}(1), \ldots, \mathbf{X}(t)$, derived in Lemma 1.

Problem Statement: We consider optimal estimators of $\mathbf{y}(\mathbf{x}^*)$ in the multi-classification paradigm in Equation 4, shown graphically in Figure 2.

$$\mathbf{x}^* \to \{\mathbf{X}(1), \ldots, \mathbf{X}(t)\} \to \mathbf{Z}(t) \to \hat{\mathbf{Y}}(t) \tag{4}$$

Contribution: We derive optimal classification estimates $\hat{\mathbf{Y}}(t)$ in Lemma 2. We show that the mean μ of $\hat{\mathbf{X}}(t)$ – the optimal estimator of $\mathbf{x}^*$ – can be reconstructed from $\hat{\mathbf{Y}}(t)$ in Theorem 1. Finally, we show via the data processing inequality that latent variable $\mathbf{Z}(t)$ must contain all information about μ for any system that produces optimal $\hat{\mathbf{Y}}(t)$.

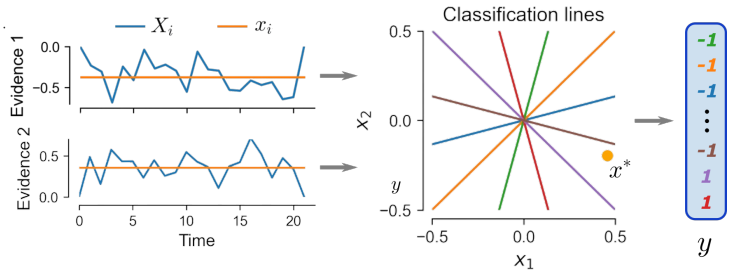

Figure 2: Data generating process. The task is to simultaneously report whether the true joint evidence $\mathbf{x}^* = [x_1, x_2]^T$ (yellow dot) lies above $(+1)$ or below (-1) a number of classification lines (here 6).

2.1 Single Decision Boundary

First, we will derive $\hat{Y}(t)$ for a single decision boundary with parameters $(\mathbf{c}, b)$. We focus on $P(\hat{Y}(t)|\mathbf{X}(1), \ldots, \mathbf{X}(t))$, reintroducing the latent variable $\mathbf{Z}(t)$ later on.
Since $y(\mathbf{x}^*)$ is a deterministic function of non-random variable $\mathbf{x}^*$, we will derive the probability distribution over $P(\mathbf{x}^*|\mathbf{X}(1), \ldots, \mathbf{X}(t))$ – denoted $\hat{\mathbf{X}}(t)$ – to determine $\hat{Y} = y(\hat{\mathbf{X}}(t))$ [1].

Lemma 1. *Assuming no prior on* $\mathbf{x}^*$*, the conditional probability distribution* $\hat{\mathbf{X}}(t) \sim P(\mathbf{x}^*|\mathbf{X}(1), \ldots, \mathbf{X}(t))$ *is given by*

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

where $\mu(t) = mean(\mathbf{X}(1), \ldots, \mathbf{X}(t))$ *and* $\Sigma(t) = t^{-1}\sigma^2\mathbf{I}_d$.

Proof. Since $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ 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 scaled by σ) to arrive at $\Sigma(t) = t^{-1}\sigma^2\mathbf{I}_d$ as the variance on the sample mean (derived from the central limit theorem). □

[1] Note that the intermediate computation of $\hat{\mathbf{X}}(t)$ does not imply that a system *must* compute this value to predict $\hat{Y}$, as the full computation of $\hat{\mathbf{X}}(t)$ may not be necessary to determine $\hat{Y}(t)$.

We can use estimator $\hat{\mathbf{X}}(t)$ to construct $\hat{\mathbf{Y}}(t)$ by expanding $\hat{\mathbf{Y}}(t) = y(\hat{\mathbf{X}}(t))$ via Equation 3.
In essence, 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).

Lemma 2. *$\hat{\mathbf{X}}(t) = \mathcal{N}(\mu(t), \Sigma(t))$ with isotropic covariance $\Sigma(t) = t^{-1}\sigma^2\mathbf{I}_d$ and mean $\mu(t) \in \mathbb{R}^D$. The probability density of $\hat{\mathbf{X}}(t)$ on the positive side of the decision boundary $\{\mathbf{x} : \mathbf{c}^\top\mathbf{x} > b\}$ can be expressed as*

$$\hat{Y}(t) \triangleq \Pr\{\mathbf{c}^\top\mathbf{x}^* > b\} = \Phi(k\sqrt{t}/\sigma) \tag{6}$$

where Φ is the CDF of the normal distribution and $k = \mathbf{c}^\top\mu(t) - b$ is the signed projection distance between the decision boundary and the mean $\mu(t)$ of $\hat{\mathbf{X}}(t)$.

Proof. Since the $\hat{\mathbf{X}}(t)$ 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 $\hat{\mathbf{X}}(t)$ 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 $\hat{\mathbf{X}}(t)$ on axis 0 is a univariate Gaussian with variance $t^{-1}\sigma^2$ mean at distance k from the boundary. Equation 6 applies the normal distribution CDF Φ to determine the probability mass on the positive side of the boundary. □

Observe that $\hat{Y}(t)$ in Equation 6 **monotonically scales** with the signed distance k between the hyperplane and $\mu(t)$ (CDFs are monotonic).

Lemma 3. *Knowledge of time t and optimal classification estimate $\hat{Y}(t)$ is sufficient to determine the projection distance k between $\mu(t) = mean\big(\mathbf{X}(1), \ldots, \mathbf{X}(t)\big)$ and the decision boundary $(\mathbf{c}, b)$.*

Proof. Recall Equation 6 from Lemma 2. We may solve for projection distance k separating the decision boundary and the mean $\mu(t)$ of observations $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ as

$$k = \frac{\sigma}{\sqrt{t}}\Phi^{-1}(\hat{Y}(t)) \tag{7}$$

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

2.2 Trilateration via Multiple Decision Boundaries

To recap Section 2.1 : We derived an optimal estimator of $\mathbf{x}^*$ (denoted $\hat{\mathbf{X}}(t)$) based on noisy i.i.d. measurements $\mathbf{X}(1), \ldots, \mathbf{X}(t) \sim \mathcal{N}(\mathbf{x}^*, \sigma^2 \mathbf{I}_d)$ in Lemma 1. In Lemma 2 we derived the equation for Bernoulli variable $\hat{Y}(t)$ to estimate a single classification output $y(\mathbf{x}^*)$ based on the same noisy measurements via $\hat{\mathbf{X}}(t)$. Finally, we showed in Lemma 3 that the uncertainty in $\hat{Y}(t)$ and the time t is sufficient to determine the projection distance between the decision boundary and $\mu(t) = \text{mean}(\mathbf{X}(1), \ldots, \mathbf{X}(t))$ via Equation 7.

Let $\hat{\mathbf{Y}}(t)$ denote the vector of classification estimates $\hat{Y}(t)$ from Equation 7. We now have the tools to prove our final result via **trilateration**. Much like distance information from cell towers can be used to trilaterate[2] one's position, we will leverage Lemma 3 and use distances from decision boundaries $\{(\mathbf{c}_i, b_i)\}_{i \in [N_{task}]}$ to constrain the positions.

Theorem 1 (Trilateration Theorem)**.** *If* $\mathbf{C}$ *is full-rank, then* $\hat{\mathbf{Y}}(t)$*,* t*,* $\mathbf{b}$*, and* $\mathbf{C}$ *are sufficient to reconstruct the exact value of* $\mu(t)$*, 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 $\mu(t) = \text{mean}(\mathbf{X}(1), \ldots, \mathbf{X}(t))$ from $\hat{\mathbf{Y}}(t), \mathbf{C}$, and t. Invoke Lemma 3 to compute the signed projection distance between $\mu(t)$ and each decision plane $(\mathbf{c}_i, b_i)$. Let $\mathbf{k} = [k_1, \ldots, k_M]^\top$ where each k_i corresponds to decision boundary $\mathbf{c}_i$. Then the mean $\mu(t)$ must satisfy

$$\mathbf{C}\mu(\mathbf{t}) = \mathbf{k} + \mathbf{b} \tag{8}$$

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

Theorem 2 (General Representation Theorem)**.** *Any system that optimally estimates* $\hat{\mathbf{Y}}$ *based on noisy measurements* $\{\mathbf{X}(1), \ldots, \mathbf{X}(t)\}$ *must implicitly encode a representation of optimal estimator* $\mu(t) = \textit{mean}(\mathbf{X}(1), \ldots, \mathbf{X}(t))$ *in its latent state* $\mathbf{Z}(t)$ *if decision boundary matrix* $\mathbf{C}$ *is full rank.*

[2]Trilateration differs from triangulation, and it is more frequently used in practice. Triangulation is when one has angle information w.r.t. the cell towers. Usually, this is not available – so one **trilaterates** their position [**?**]. This more closely matches our setting, where we just have distances information w.r.t. the decision boundaries and must determine the position.

Proof. This follows from the data processing inequality. We begin with the following Markov chain:

$$\mathbf{x}^* \to \{\mathbf{X}(1), \ldots, \mathbf{X}(t)\} \to \mathbf{Z}(t) \to \hat{\mathbf{Y}}(t) \to \mu(t) \tag{9}$$

Put more simply, we have $\mathbf{x}^* \to \mathbf{Z}(t) \to \mu(t)$. Applying the data processing inequality, we obtain

$$I\big(\mathbf{x}^*; \mathbf{Z}(t)\big) \geq I\big(\mathbf{x}^*; \mu(t)\big) \tag{10}$$

where $\mathbf{I}(\cdot;\cdot)$ denotes the mutual information between two variables.
Since $\hat{\mathbf{X}}(t)$ is the optimal estimator of $\mathbf{x}^*$ given measurements $\{\mathbf{X}(1), \ldots, \mathbf{X}(t)\}$, $I\big(\mathbf{x}^*; \mu(t)\big) = H(\mu(t))$. Therefore $I(\mathbf{x}^*; \mathbf{Z}(t)) \geq H(\mu(t))$, implying that $\mathbf{Z}(t)$ must contain all the information of $\mu(t)$. □

3 Discussion

The primary result boils down to the observation that the confidence associated with each $\hat{Y}_i$ in $\hat{\mathbf{Y}}$ are measures of distance between an implied estimate of $\mathbf{x}^*$ (denoted $\hat{\mathbf{X}}$) and classification boundary i (denoted $(\mathbf{c}_i, b)$). $\hat{\mathbf{Y}}$ specifies the position of $\hat{\mathbf{X}} = \mu$ via "coordinates" defined by decision boundaries $\mathbf{c}_1, \ldots, \mathbf{c}_{N_{task}}$.

For sub-optimal estimators of $\hat{\mathbf{Y}}$, we may still obtain an understanding of the implied estimate $\hat{\mathbf{X}}$ using the same methods. In fact, the machinery of least-squares estimation for $\mathbf{Ax} = \mathbf{b}$ provides a readily accessible formula for $\tilde{\mu}$ in sub-optimal estimators of $\hat{\mathbf{Y}}$ in the form of the Moore-Penrose pseudoinverse:

$$\tilde{\mu} = (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^T (\mathbf{k} + \mathbf{b}) \tag{11}$$

Conveniently, if the estimation errors in sub-optimal $\hat{\mathbf{Y}}$ have a mean of zero, additional decision boundaries in $\mathbf{C}$ (e.g., beyond the minimum D linearly independent boundaries) result in improved estimation of $\mathbf{x}^*$ by the central limit theorem, thus generalizing our results to sub-optimal estimators.

3.1 Non-Linear Decision Boundaries

The result in Theorem 1 may be generalized to non-linear decision boundaries in two ways:

1. **Local Linear Approximation:** Assume that the world coordinate space $\mathcal{X}$ has decision boundaries ϕ_i that are locally linear in the neighbourhood of $\mathbf{x}^*$. If the space has a sufficient density of decision boundaries, we may apply the same trilateration theorem (1) to the local linear approximation of the decision boundaries.

2. **Inserting a Non-Linear Observation Map:** Assume there exists an abstract world coordinate space $\mathcal{X}^*$ with a non-linear map $\mathcal{F} : \mathcal{X}^* \to \mathcal{X}$ such that the decision boundaries ϕ_i are linear in $\mathcal{X}^*$. If $\mathcal{F}$ is injective, we may apply the same trilateration theorem (1) to the linear decision boundaries in $\mathcal{X}^*$ to obtain the same result.

These are left as an exercise for the reader.

3.2 Relationship to the Manifold Hypothesis

Our theoretical results have important implications for the manifold hypothesis, which posits that real-world high-dimensional data tend to lie on or near low-dimensional manifolds embedded in the high-dimensional space [1, 2]. The key insight is that our proofs show an optimal multi-task classifier must encode an estimate of the abstract coordinates of the true underlying environment state in its latent representation.

Consider an abstract space $\mathcal{X}^*$ and an injective observation map $\mathcal{F} : \mathcal{X}^* \to \mathcal{X}$, where the decision boundaries $\phi_i : \mathcal{X} \to 0, 1$ have a linear image when translated into $\mathcal{X}^*$ via $\mathcal{F}^{-1}$. Our results imply that an optimal classifier's latent state $Z(t)$ must encode an estimate of the abstract coordinates $x^* \in \mathcal{X}^*$, rather than the ambient coordinates $x \in \mathcal{X}$. This is a crucial distinction, as the abstract coordinates capture the intrinsic geometry of the data, which is invariant under coordinate transformations of the ambient space.

$$\mathcal{F} : \mathcal{X}^* \to \mathcal{X}, \quad \phi_i \circ \mathcal{F} : \mathcal{X}^* \to 0, 1 \text{ is linear} \tag{12}$$

The injective observation map $\mathcal{F}$ aligns closely with the typical conception of a data manifold. The abstract space $\mathcal{X}^*$ can be seen as the intrinsic coordinate system of the manifold, while $\mathcal{F}$ maps these coordinates to the high-dimensional observation space $\mathcal{X}$. Our findings suggest an optimal classifier will implicitly learn to invert this mapping and recover the abstract

coordinates.

Moreover, for natural data where the manifold hypothesis holds, the learned latent representation would plausibly capture the manifold structure, as this is essential for disambiguating noisy observations and estimating the true underlying state. The low-dimensional manifold structure is a key prior that an optimal classifier can exploit to improve its performance.

In conclusion, our theoretical results provide a novel perspective on the emergence of manifold-aligned representations in neural networks trained on multi-task classification. We show that these representations arise as a natural consequence of optimal multi-task classification, with the latent space encoding the abstract coordinates of the data manifold. This insight deepens our understanding of why multi-task learning encourages the discovery of geometrically meaningful representations and suggests a strong link between optimal classification and manifold learning.

3.3 References

1. Fefferman, C., Mitter, S., & Narayanan, H. (2016). Testing the manifold hypothesis. Journal of the American Mathematical Society, 29(4), 983-1049.

2. Chris Olah (2014). Neural Networks, Manifolds, and Topology. *colah's blog* https://colah.github.io/posts/2014-03-NN-Manifolds-Topology.