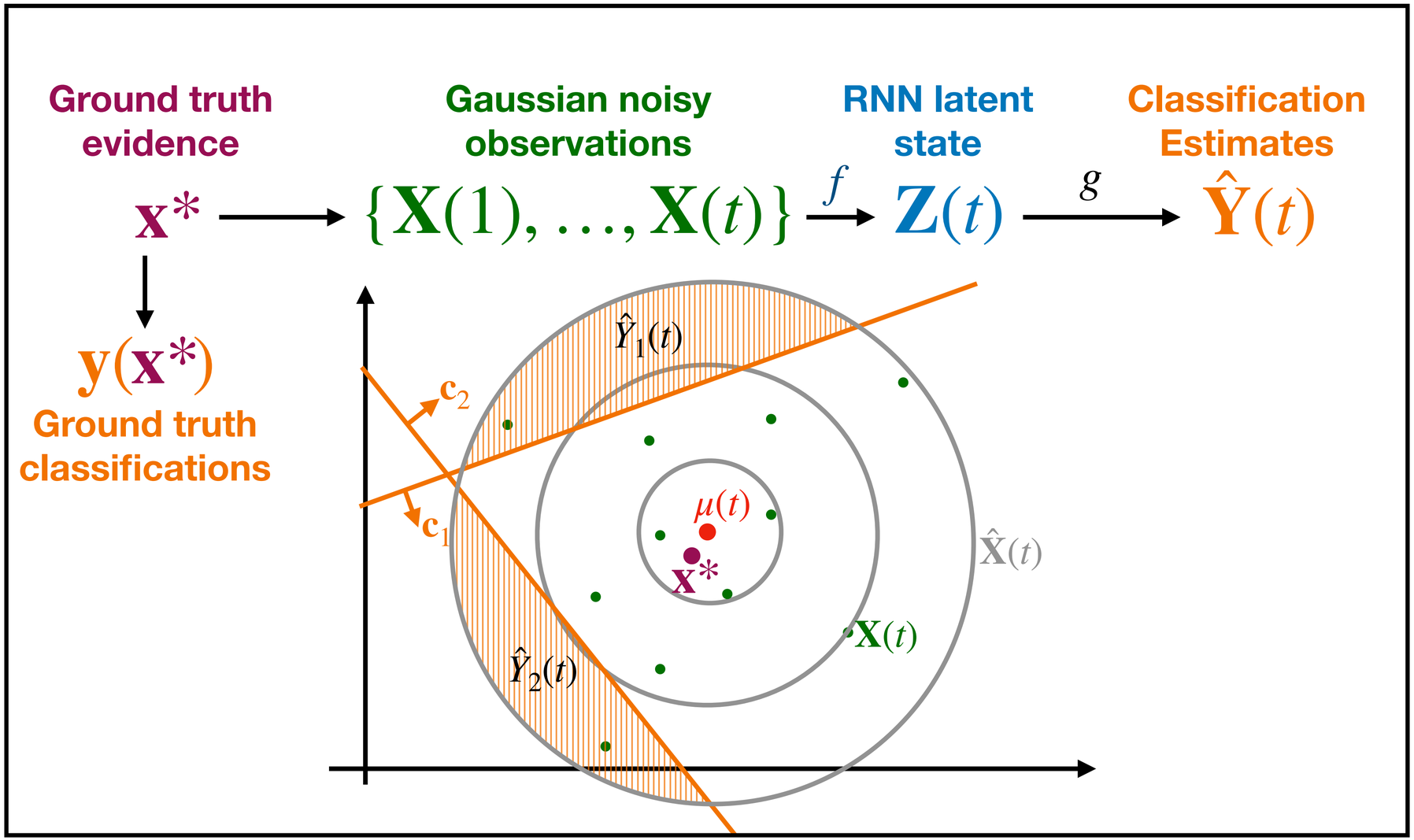
Disentangled Representation Theorem

A crash course for the CNS-minded

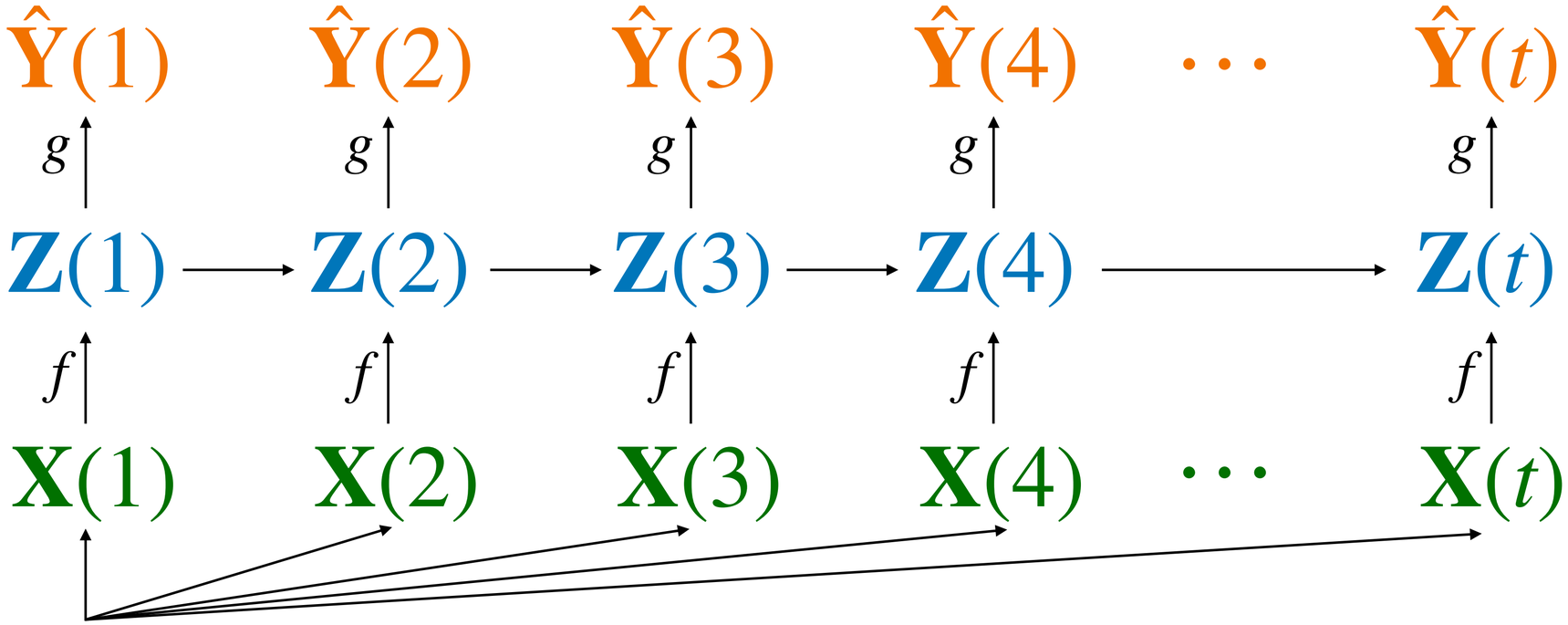
Pantelis Vafidis (Compute/Experiment), **Aman Bhargava** (Theory), **Matt Thomson** (advice), **Antonio Rangel** (PI)

Problem Setup: Multi-task evidence aggregation classification



We analyze generic filter/evidence aggregators with latent state $\mathbf{Z}(t)$.

Problem Statement • Pie Slice Intuition • Classifications are Distances • Disentangled Reps Theorem



Multi-task classification estimates

E.g. RNN latent state

+ i.i.d. isotropic Gaussian noise

$\mathbf{x}^*$ **Ground truth evidence**

$\mathbf{y}(\mathbf{x}^*)$ **Ground truth classifications**

f is an injective, potentially non-linear & high-dim observation map

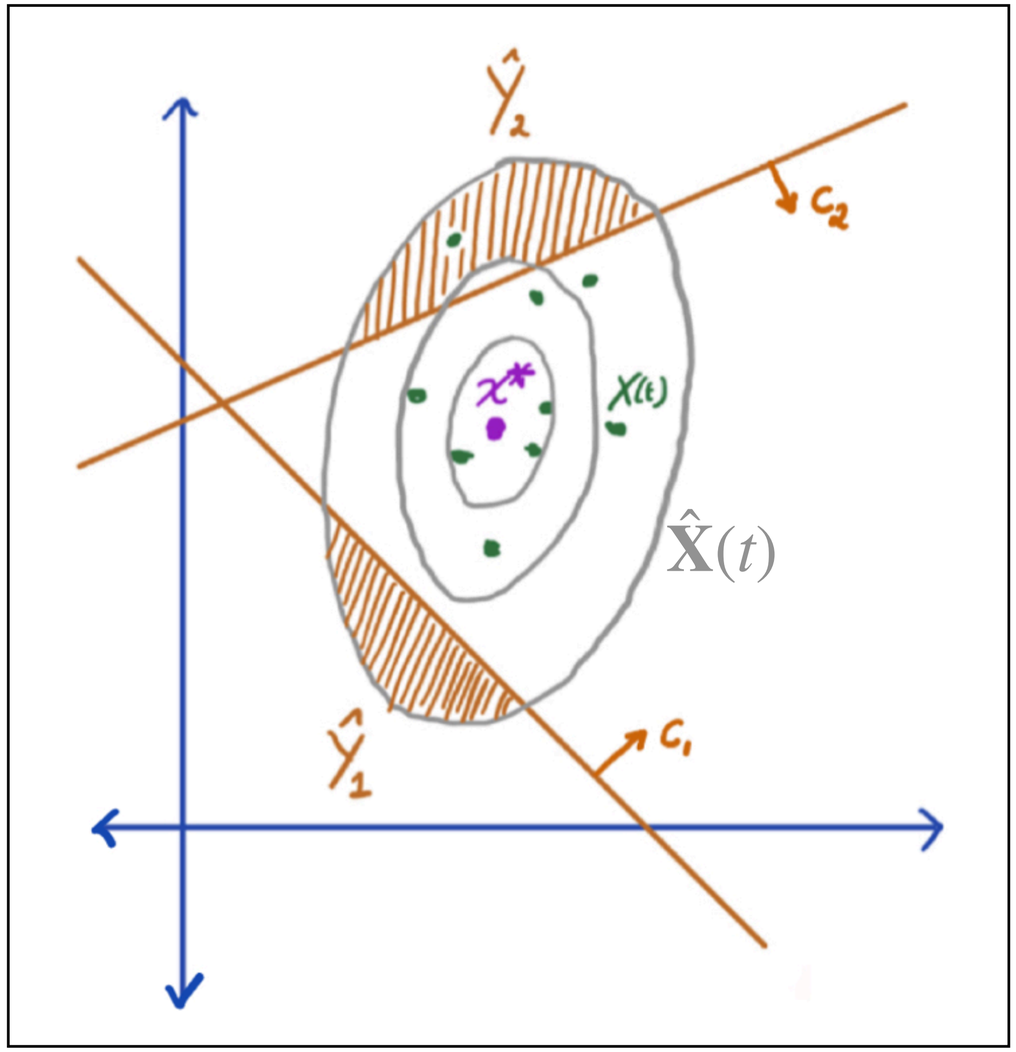
g is the classification readout (e.g., $g = \tanh$)

Classification Boundaries are Linear in $\mathbf{x}^*$.

- "Ground truth" $\mathbf{y}(\mathbf{x}) \in \{-1, +1\}^{N_{task}}$ are hyperplanes in $\mathbf{x} \in \mathbb{R}^D$

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

$$\mathbf{C} = \begin{bmatrix} \leftarrow & \mathbf{c}_1 & \rightarrow \\ \leftarrow & \mathbf{c}_2 & \rightarrow \\ \leftarrow & \vdots & \rightarrow \\ \leftarrow & \mathbf{c}_{N_{task}} & \rightarrow \end{bmatrix}$$



Optimal Estimation: MAP = ML estimation for no prior on $\mathbf{x}^*$.

Problem Statement • Pie Slice Intuition • Classifications are Distances • Disentangled Reps Theorem

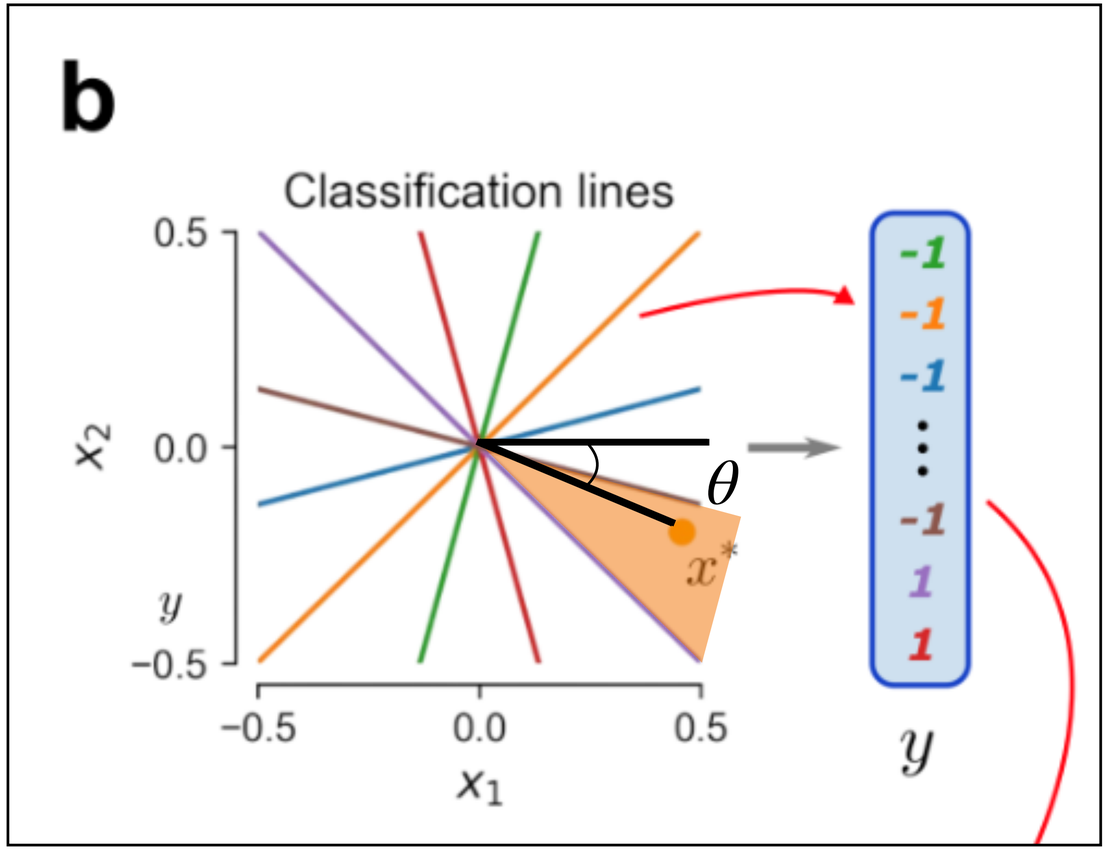
- Estimator $\hat{\mathbf{Y}}(t)$ is a maximum a posteriori (MAP) estimator of $\mathbf{y}(\mathbf{x}^*)$ if

$$\hat{Y}_i(t) = \Pr\left\{ y_i(\mathbf{x}^*) = 1 \,\middle|\, \mathbf{X}(1), \ldots, \mathbf{X}(t) \right\}$$

- Which is the same as the maximum likelihood (ML) estimator if we assume no prior in $\mathbf{x}^*$
- *we will derive the optimal estimator later…

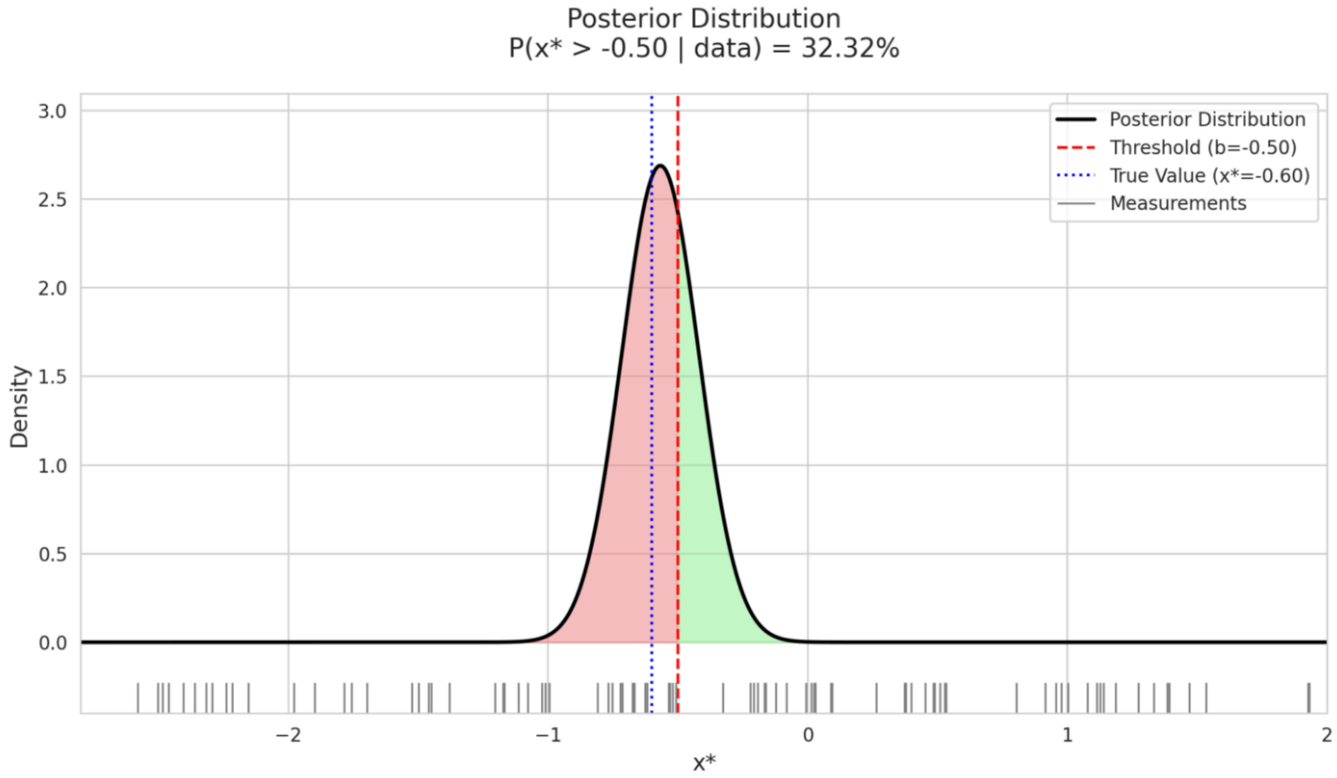
$\hat{\mathbf{Y}}(t)$ specifies the "pie slice" where $\mathbf{x}^*$ resides.

- **Concept 1**: Each correct classification narrows $\mathbf{x}^*$ to a half space.
- **Concept 2**: Closer $\mathbf{x}^*$ to decision boundary $i \implies$ less certainty in $Y_i(t)$.



$\hat{Y}_i(t)$ represent distance k_i between decision boundary i and estimate of $\mathbf{x}^*$.

$$\hat{Y}_i(t) = \Pr\left\{ y_i(\mathbf{x}^*) = 1 \,\middle|\, \mathbf{X}(1), \ldots, \mathbf{X}(t) \right\}$$



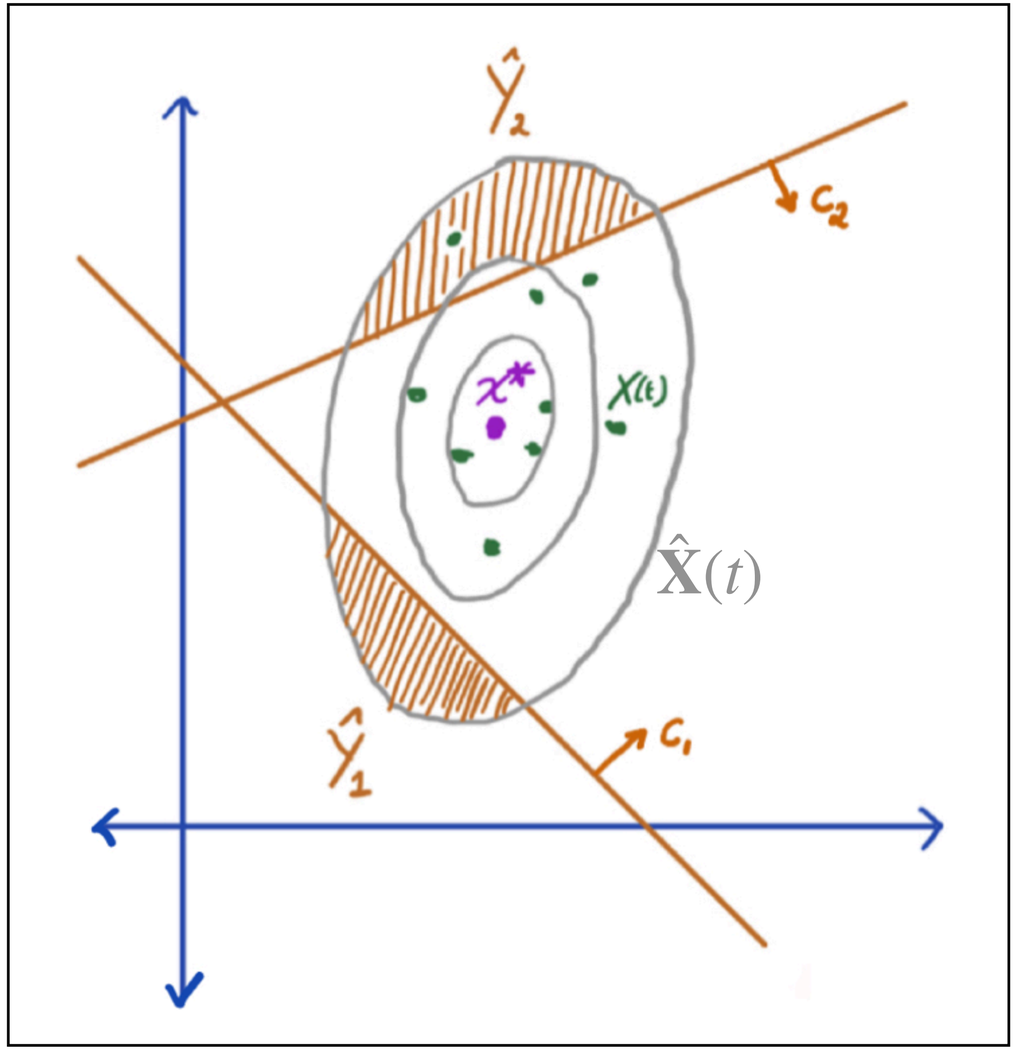
$\hat{Y}_i(t)$ represent distance k_i between decision boundary i and estimate of $\mathbf{x}^*$.

- **Optimal $\hat{\mathbf{Y}}(t)$:** Compute $\hat{\mathbf{X}}(t) = P\left(\mathbf{x}^* \,|\, \mathbf{X}(1)\ldots\mathbf{X}(t)\right)$. Integrate probability mass on each side of boundaries $(\mathbf{c}_i, b_i)$, $i \in [N_{task}]$.

$$\hat{Y}(t) \triangleq \Pr\{\mathbf{c}^\top \mathbf{x}^* > b \,|\, \mathbf{X}(1)\ldots\mathbf{X}(t)\}$$

$$= \Phi(k\sqrt{t}/\sigma)$$

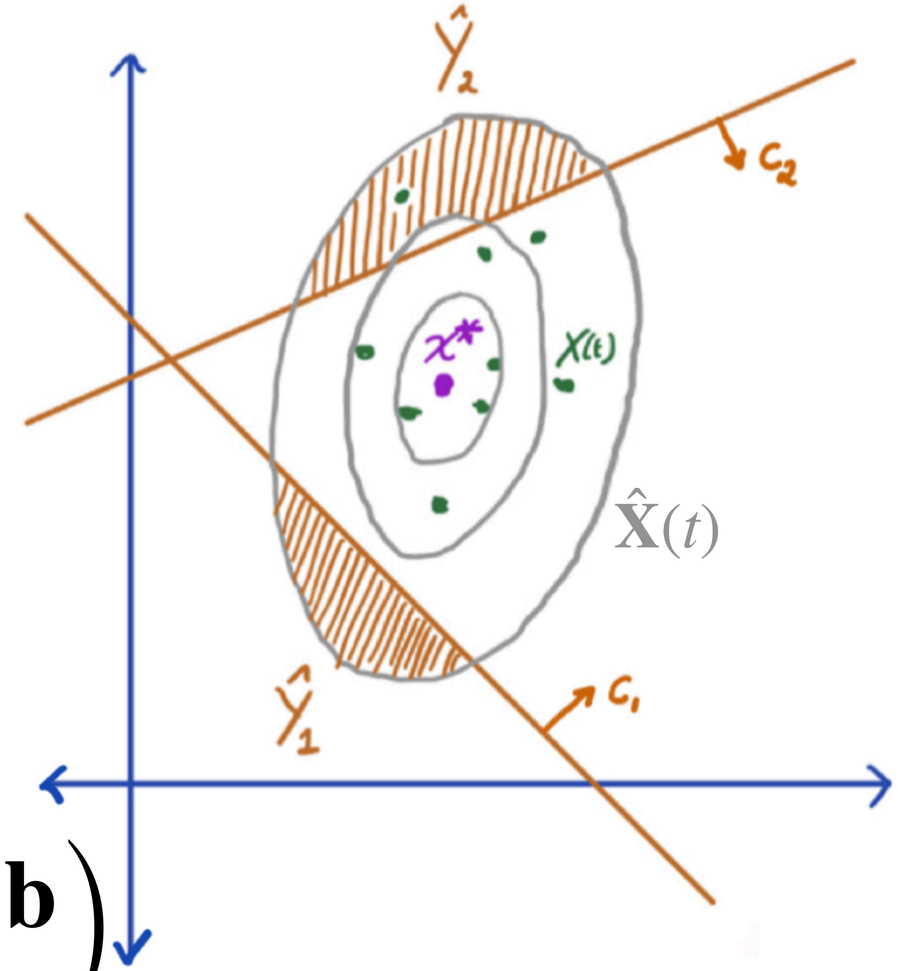
$$\implies k = \frac{\sigma}{\sqrt{t}}\Phi^{-1}(\hat{Y}(t))$$



Multiple $\hat{Y}_i(t)$ trilaterate the position of $\mathbf{x}^*$ estimate.

Problem Statement • Pie Slice Intuition • **Classifications are Distances** • Disentangled Reps Theorem

- $\hat{\mathbf{X}}(t) = P\left(\mathbf{x}^* \mid \mathbf{X}(1)\ldots\mathbf{X}(t)\right)$
  - $= \mathcal{N}\left(\mu(t), \Sigma(t)\right)$ where
  - $\mu(t) = \text{mean}\left(\mathbf{X}(1)\ldots\mathbf{X}(t)\right)$
- **Theorem**: If decision boundary matrix $\mathbf{C}$ is full-rank and $N_{task} \geq D$, then then $\left(\hat{\mathbf{Y}}(t), t, \mathbf{b}, \mathbf{C}, \sigma\right)$ are sufficient to reconstruct the exact value of $\mu(t)$, the optimal estimator for $\mathbf{x}^*$.

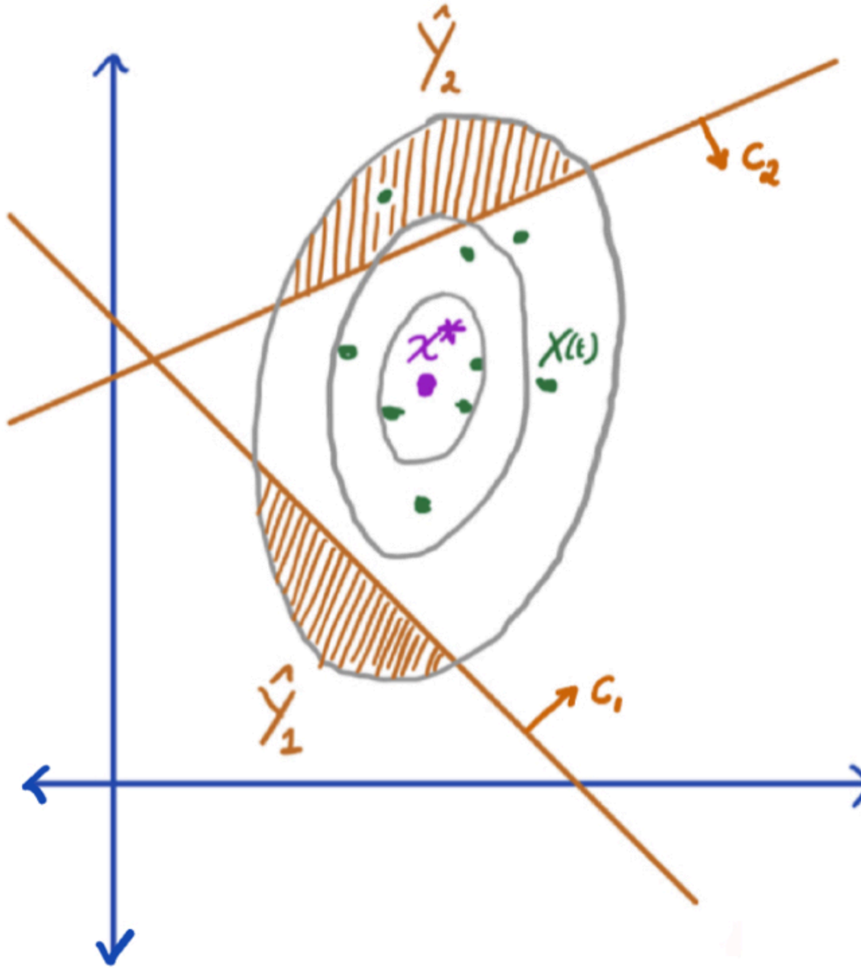


$$\mu(t) = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\left(\frac{\sigma}{\sqrt{t}}\Phi^{-1}\left(g(\mathbf{Z}(t))\right) + \mathbf{b}\right)$$

$\mathbf{x}^*$ estimate is approx. linearly decodable from $\mathbf{Z}(t)$ if $g = \tanh$.

$$\mu(t) = (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \left(\frac{\sigma}{\sqrt{t}} \Phi^{-1}\big(g(\mathbf{Z}(t))\big) + \mathbf{b} \right)$$

$$\mu(t) \approx \underbrace{\frac{2\sqrt{3}\sigma}{\pi\sqrt{t}} (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top}_{\text{linear}} \underbrace{(\mathbf{Z}(t) + \mathbf{b})}_{\text{affine}}$$



Final Result

Theorem 3.1 (Disentangled Representation Theorem). *If* $\mathbf{C} \in \mathbb{R}^{N_{task} \times D}$ *is a full-rank matrix and* $N_{task} \geq D$ *and noise* $\sigma > 0$*, then*

1. *Any optimal estimator of* $\mathbf{y}(\mathbf{x}^*)$ ***must encode a finite-sample, maximum likelihood estimate*** $\mu(t)$ *of the ground truth evidence variable* $\mathbf{x}^*$ *in its latent state* $\mathbf{Z}(t)$.

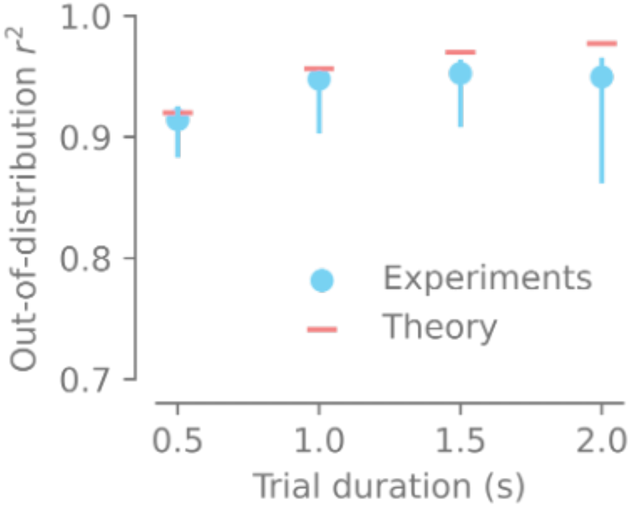
2. *If the activation function is sigmoid-like,* $\mu(t)$ *will be* ***linearly decodable from*** $\mathbf{Z}(t)$*, thus implying that* $\mathbf{Z}(t)$ *contains an abstract representation of* $\mu(t)$ *(Ostojic & Fusi, 2024).*

3. *The representation is guaranteed to be disentangled (orthogonal) as* $N_{task} \gg D$ *for random decision boundaries.*

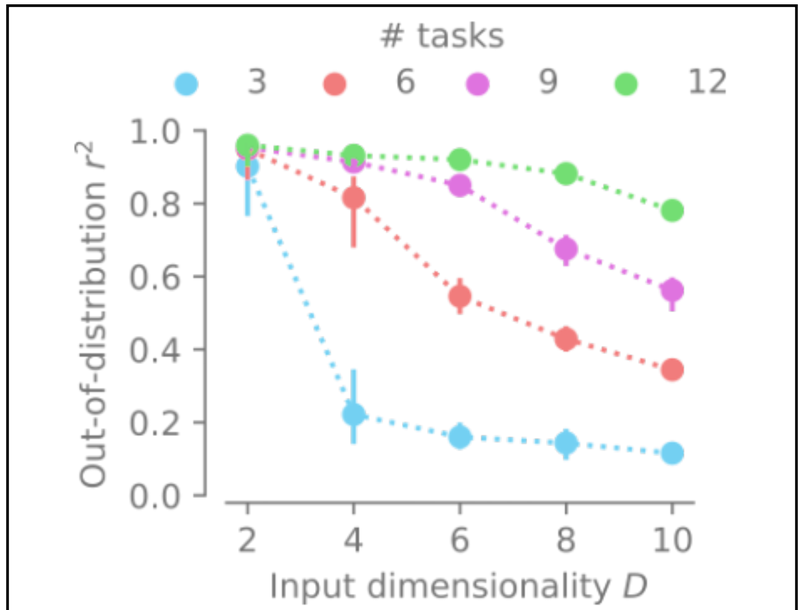
Theory Generates Testable Hypotheses!

Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**

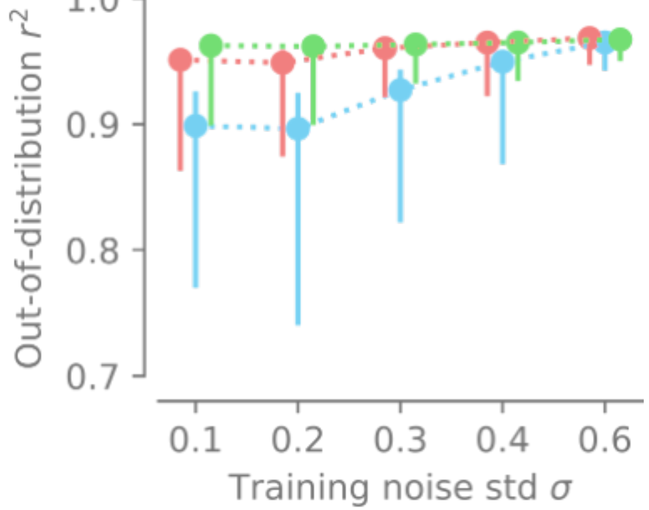
Generalization
r^2 w.r.t. trial duration t



Generalization
r^2 w.r.t. N_{task}, D



Generalization
r^2 w.r.t. noise σ



$$MSE(x_i, \mu_i) \qquad \mathbb{E}[(x_i - x_i + \mathcal{N}(0, t^{-1}\sigma^2))^2] \qquad \sigma^2$$

Theory Generates Testable Hypotheses!

Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**

Generalization

r^2 **w.r.t.** $\mathbf{x}^*$ **correlation, suboptimality**

***Predicts excellent performance even with non-linear/high-dim**

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 $\hat{\mathbf{Y}}$ (Equation 13) in the form of the Moore-Penrose pseudoinverse:

$$\tilde{\mu} = (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^T (\mathbf{k} + \mathbf{b}) \tag{16}$$

Proof. We use proof by contradiction to extend the linear case of the general representation theorem to account for injective observation maps f that map $\mathbf{X}(t)$ before they are input to $\mathbf{Z}(t)$. Assume toward a contradiction that there exists a superior way of computing $\hat{Y}$ based on injectively mapped $f(\mathbf{X}(t))$ other than learning f^{-1} and following the same procedure as when $\mathbf{X}(t)$ was fed in directly (which we derived the optimal estimator for in Lemma B.1 and Lemma B.2). This assumption implies there is some additional information in $f(\mathbf{X}(t))$ that is not in $\mathbf{X}(t)$, violating the data processing inequality.

*all experiments use a randomly initialized 100-dim MLP to encode input, still follows theory + gets excellent OOD results on disentangled reps.

Bonus: Conditions on Orthogonal Representations

The orthogonality of the representations in $\mathbf{Z}(t)$ is therefore governed by the orthogonality of the rows in the matrix $\mathbf{A} := (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top \in \mathbb{R}^{D\times N_{task}}$. If $\mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{D\times D}$ is diagonal, then the rows of $\mathbf{A}$ are orthogonal.

$$\mathbf{A}\mathbf{A}^\top = \left((\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\right)\left((\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\right)^\top$$

If the singular values are approximately uniform $\sigma_1 \dots \sigma_D \approx \sigma$ then

$$\mathbf{A}\mathbf{A}^\top \approx \mathbf{V}(\frac{1}{\sigma^2}\mathbf{I}_D)\mathbf{V}^T$$

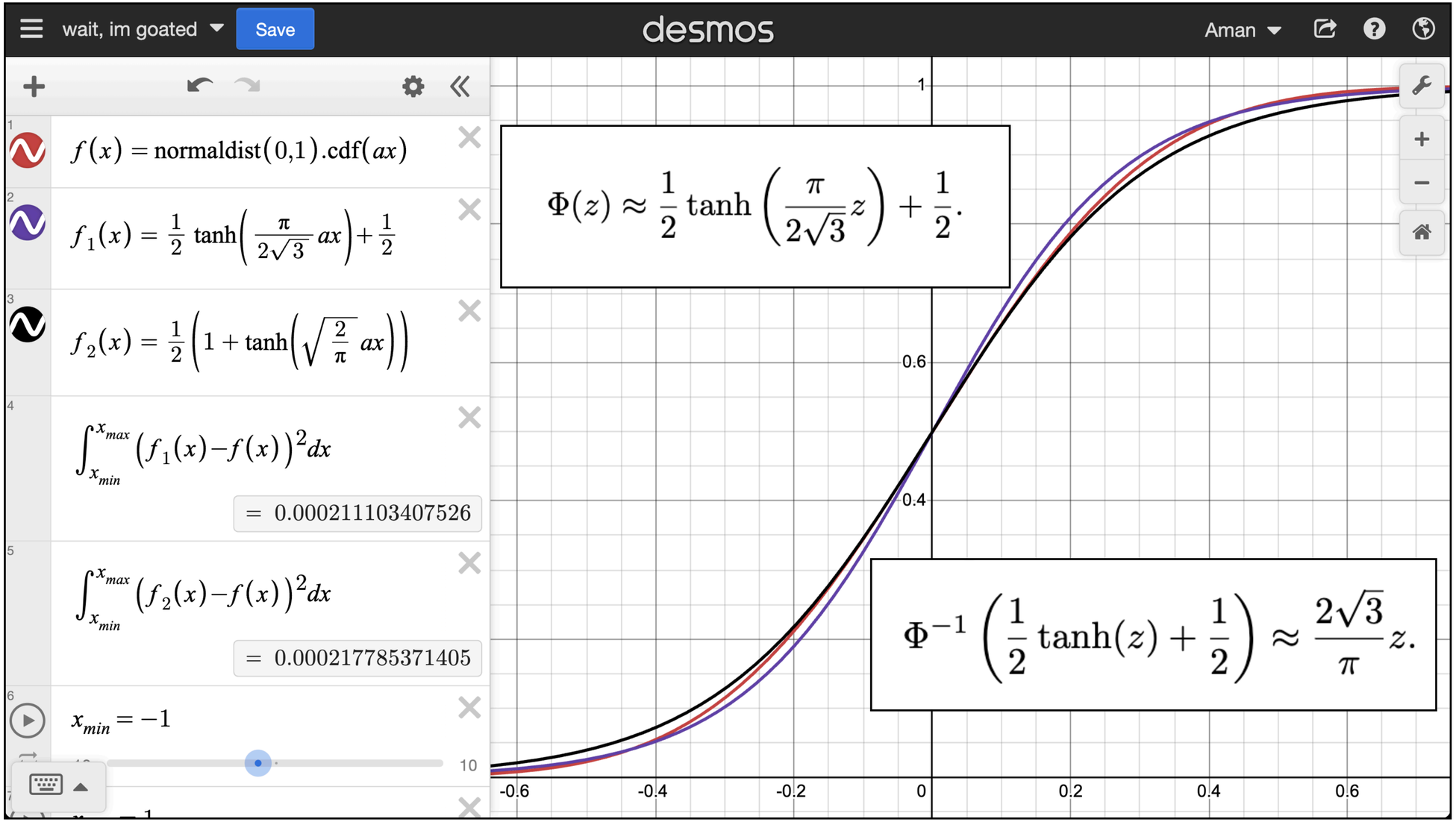
$$\mathbf{A}\mathbf{A}^\top \approx \frac{1}{\sigma^2}\mathbf{V}\mathbf{V}^T = \frac{1}{\sigma^2}\mathbf{I}_D$$

Therefore, uniform singular values in $\mathbf{C}$ is a sufficient condition to guarantee an orthogonal, disentangled representation of $\mu(t)$ in $\mathbf{Z}(t)$[5]. □

Bonus: Novel Φ Approximation?

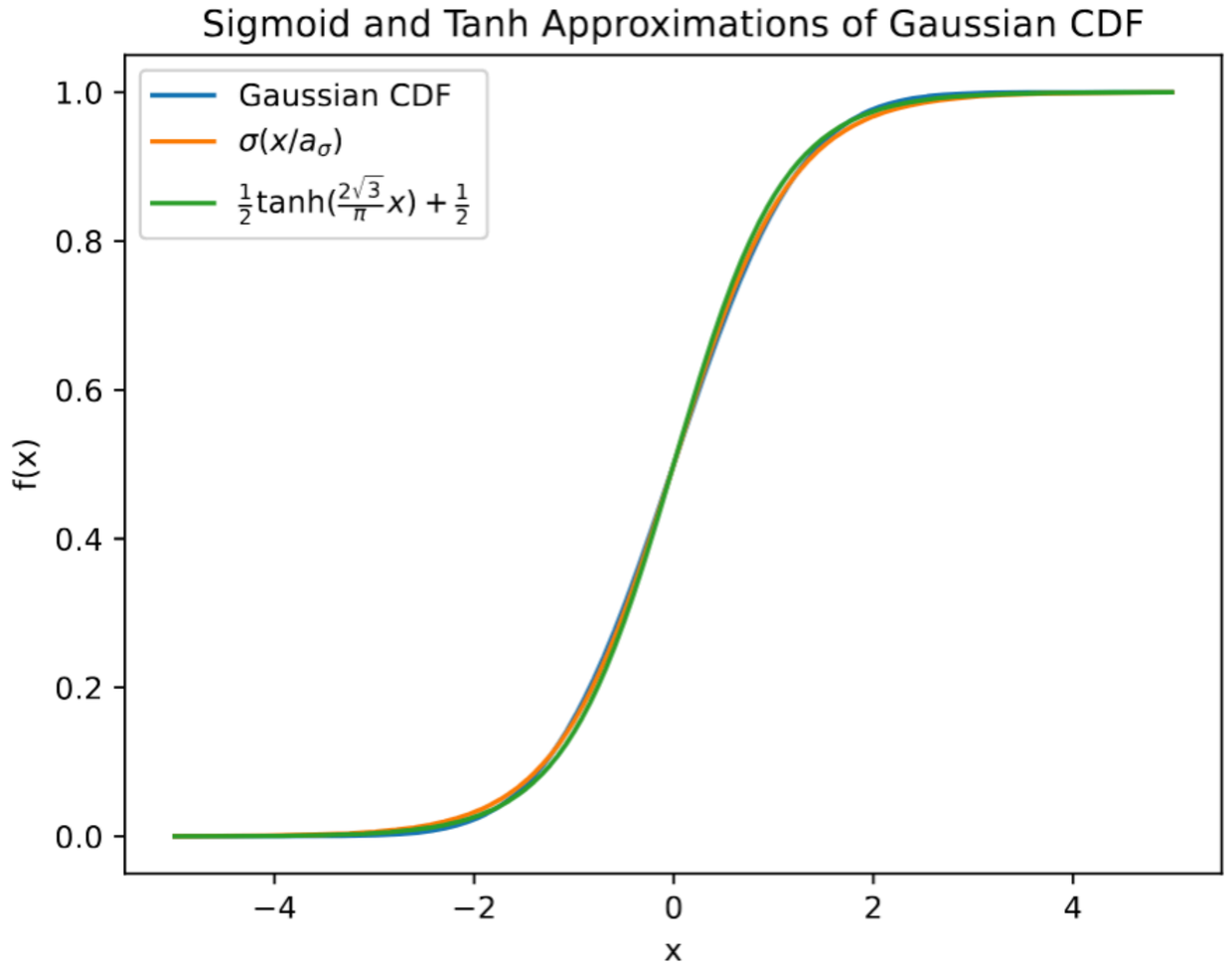
Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**

Novel Page (1977)



Bonus: Novel Φ Approximation?

Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**



Bonus: Non-Gaussian, Non-Isotropic Noise

- **Additive noise**: $\mathbf{X}(t) = \mathbf{x}^* + \epsilon(t)$
- **Tractable, invertible posterior**: $P(\mathbf{x}^* \mid \{\mathbf{X}(t)\}_{i=1}^{T})$
- **Support over** $\mathbf{x}^* \in \mathcal{X}^* \subseteq \mathbb{R}^D$

- Exponential dist.
- Multivariate T-dist.
- Anisotropic Gaussian
- Elliptical distribution

$$\hat{Y}_i(t) = \Pr\{\mathbf{c}_i^\top \mathbf{x}^* > b_i \mid [\mathbf{X}(s)]_{s=1}^{T}\}$$

$$\Longrightarrow \quad k_i = F_i^{-1}(1 - \hat{Y}_i(t))$$

Projection distance from $\mu(t)$ to boundary i

Bonus: Manifold Observation Maps f

Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**

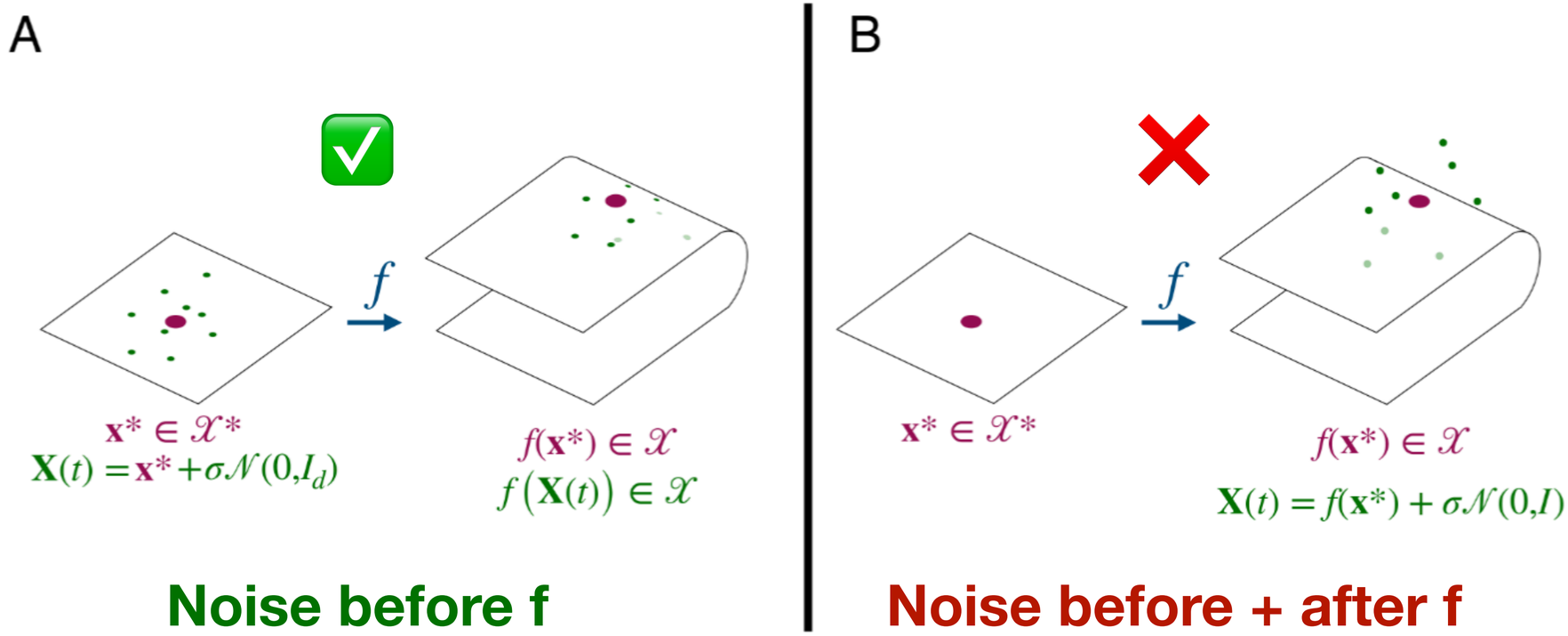
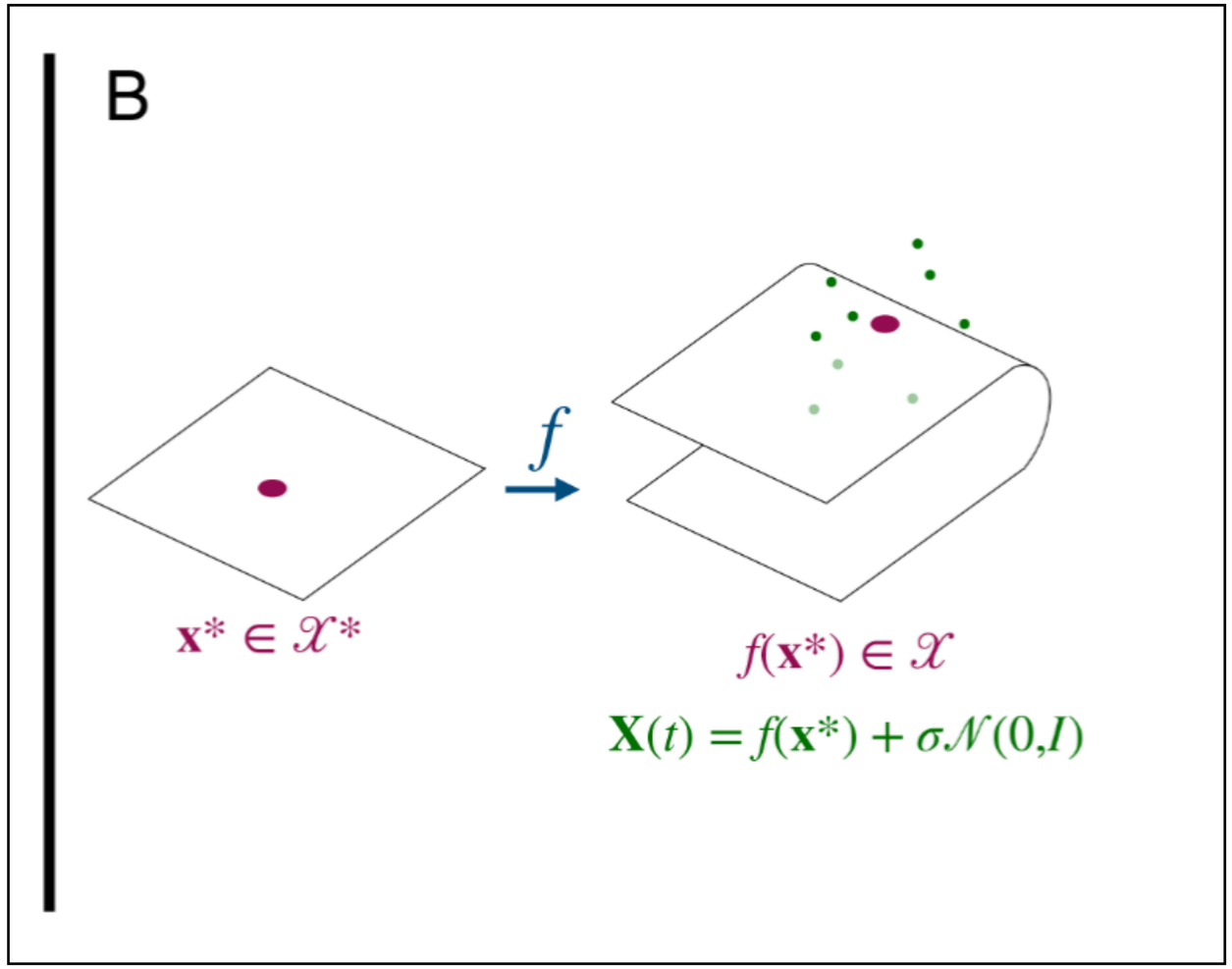


Figure S10: (A) $\mathbf{x}^*$ is noised before being transformed by injective observation map f, resulting in observations $f(\mathbf{X}(t))$ lying on the image of f (here f is a 2D folded surface). (B) $\mathbf{x}^*$ is first transformed by injective observation map f and noise is added afterward, resulting in observations $f(\mathbf{x}^*) + \sigma\mathcal{N}(0, I)$ that do not lie on the image of f.

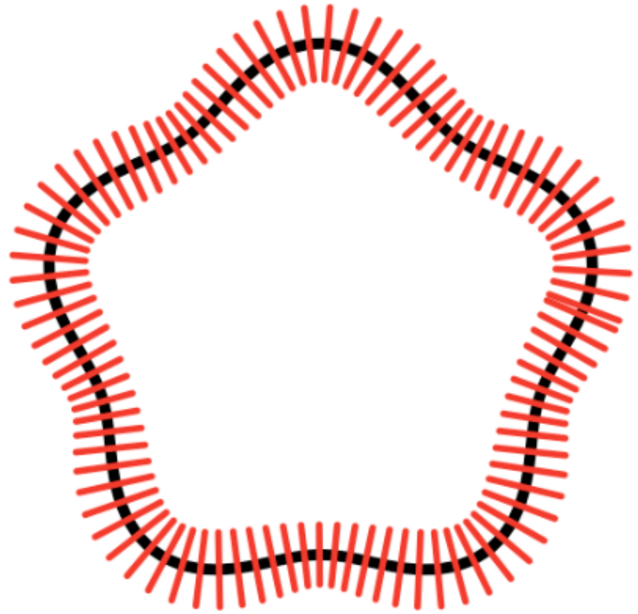
Bonus: Manifold Observation Maps f

Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**



Post-observation map noise is overcome if

$$\tau \gg \sigma$$

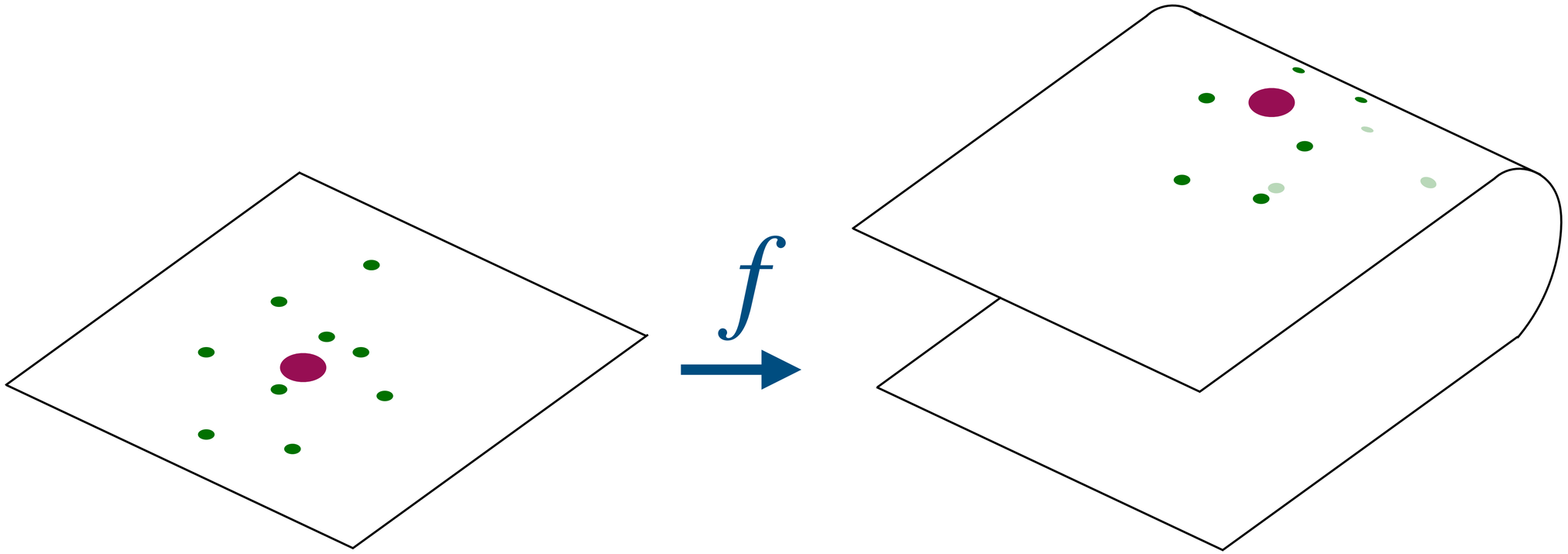


τ = reach of manifold

Linear Decisions in Latents = Local Linearization of Manifold Decisions

Problem Statement • Pie Slice Intuition • Classifications are Distances • **Disentangled Reps Theorem**

Exercise for the next paper: Prove the manifold hypothesis.

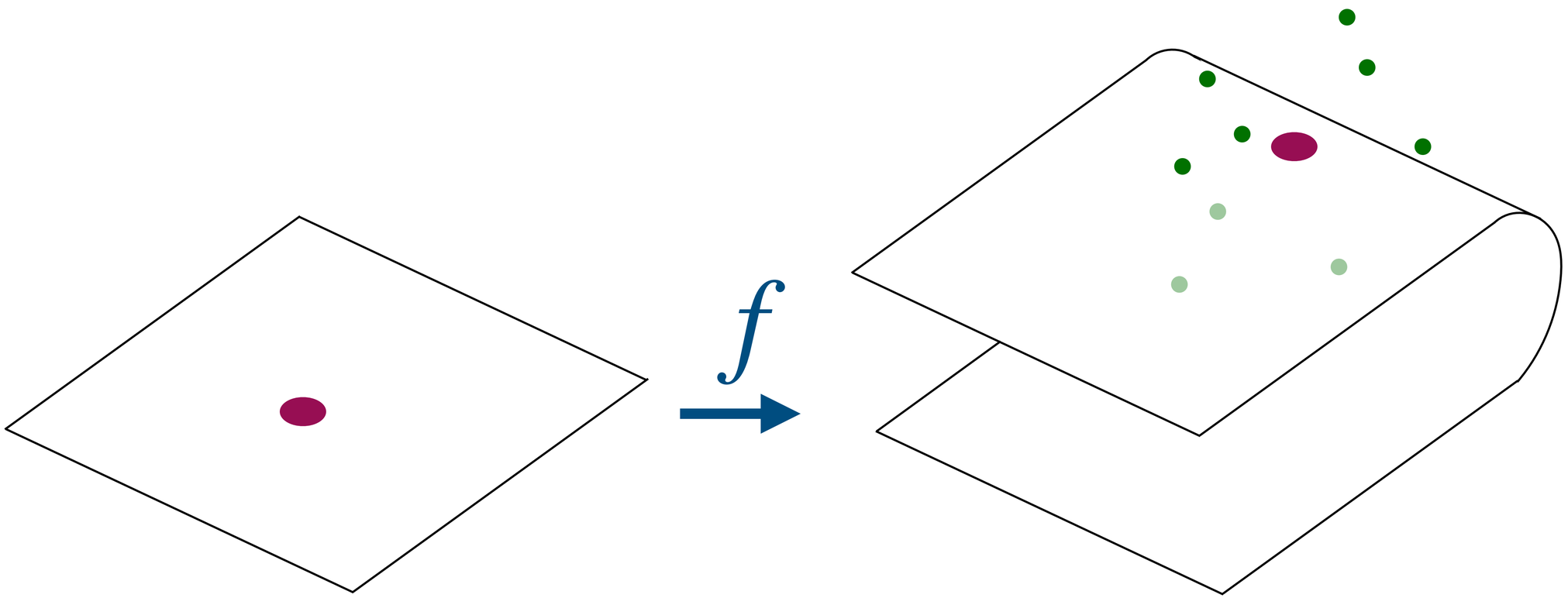


$\mathbf{x}^* \in \mathcal{X}^*$

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

$f(\mathbf{x}^*) \in \mathcal{X}$

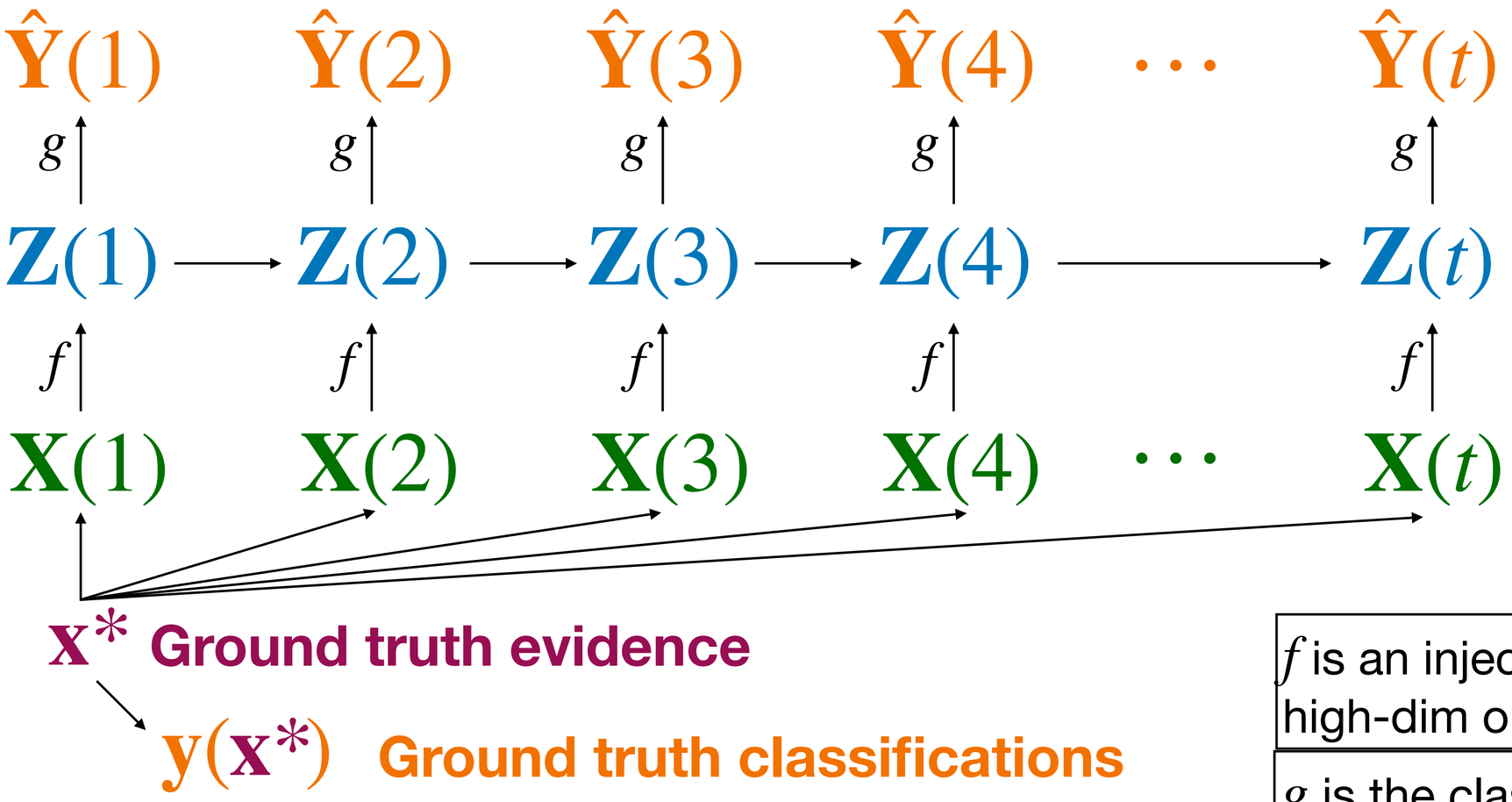
$f\big(\mathbf{X}(t)\big) \in \mathcal{X}$



$\mathbf{x}^* \in \mathcal{X}^*$

$f(\mathbf{x}^*) \in \mathcal{X}$

$\mathbf{X}(t) = f(\mathbf{x}^*) + \sigma\mathcal{N}(0,I)$



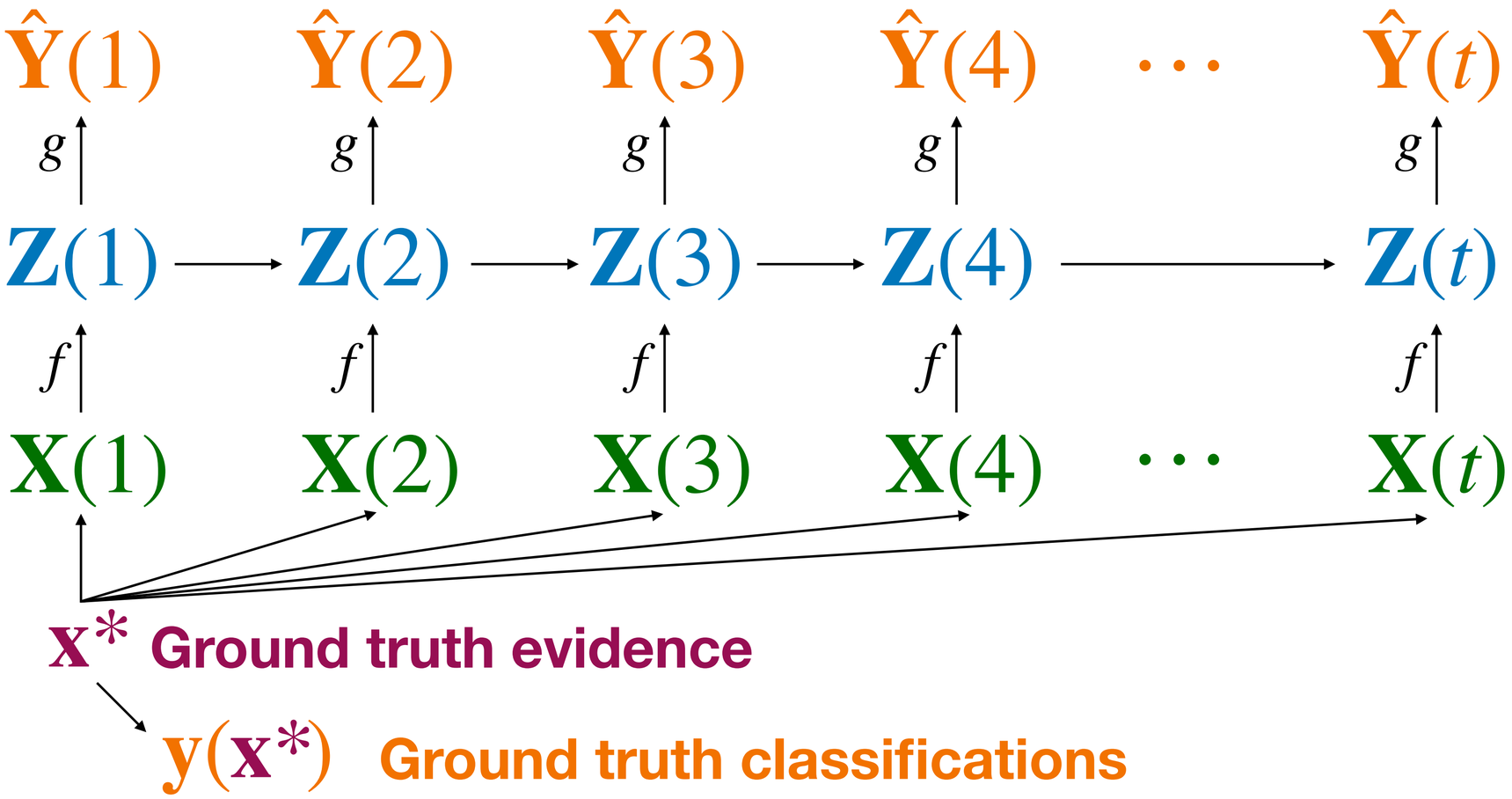
Multi-task classification estimates

E.g. RNN latent state

+ i.i.d. isotropic Gaussian noise

f is an injective, potentially non-linear & high-dim observation map

g is the classification readout (e.g., $g = \tanh$)



Multi-task classification estimates

E.g. RNN latent state

+ i.i.d. isotropic Gaussian noise

Ground truth evidence

Gaussian noisy observations

RNN latent state

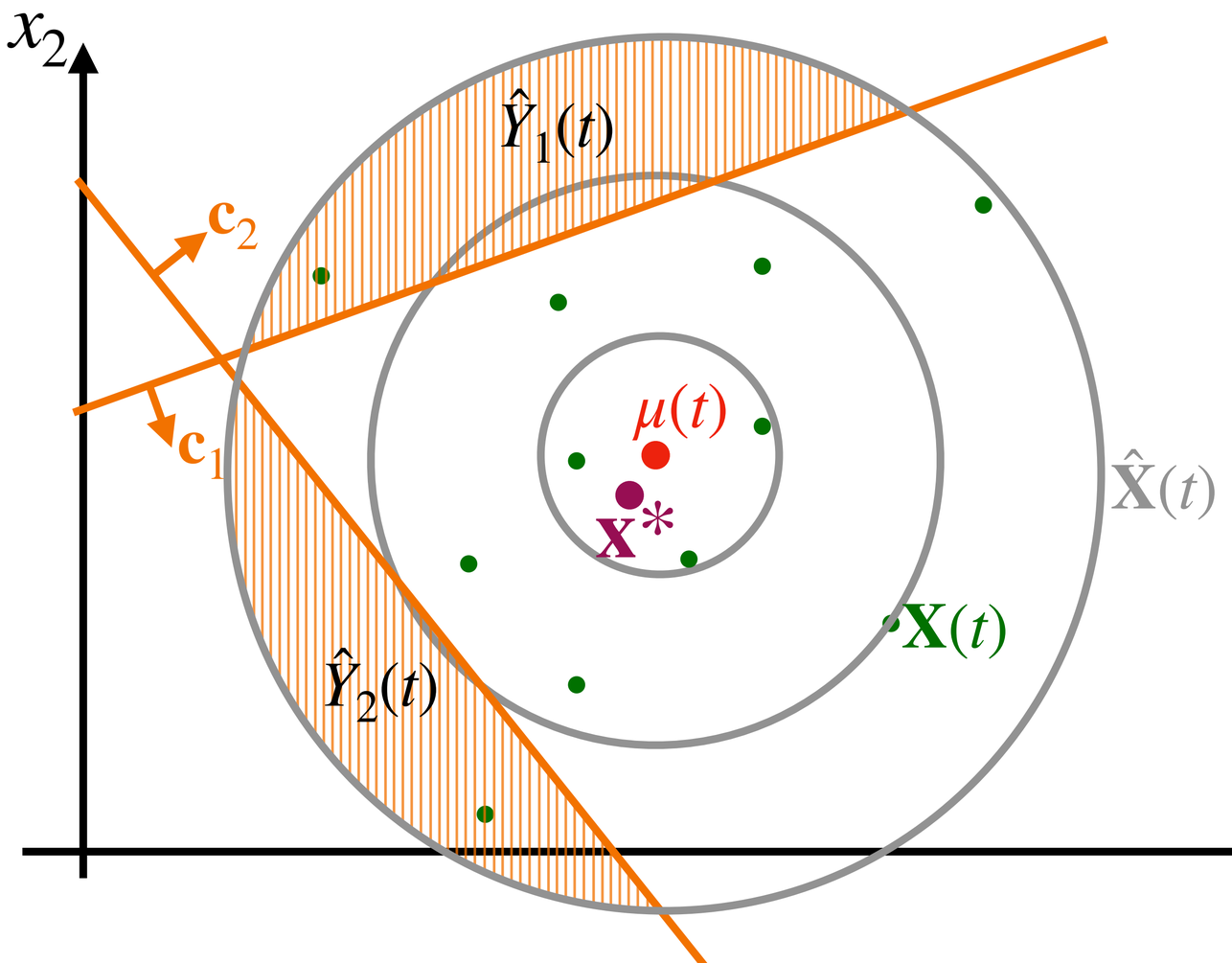
Classification Estimates

$\mathbf{x}^* \longrightarrow \{\mathbf{X}(1), \ldots, \mathbf{X}(t)\} \xrightarrow{f} \mathbf{Z}(t) \xrightarrow{g} \hat{\mathbf{Y}}(t)$

$\mathbf{x}^* \downarrow$

$\mathbf{y}(\mathbf{x}^*)$

Ground truth classifications



Ground truth evidence

Gaussian noisy observations

RNN latent state

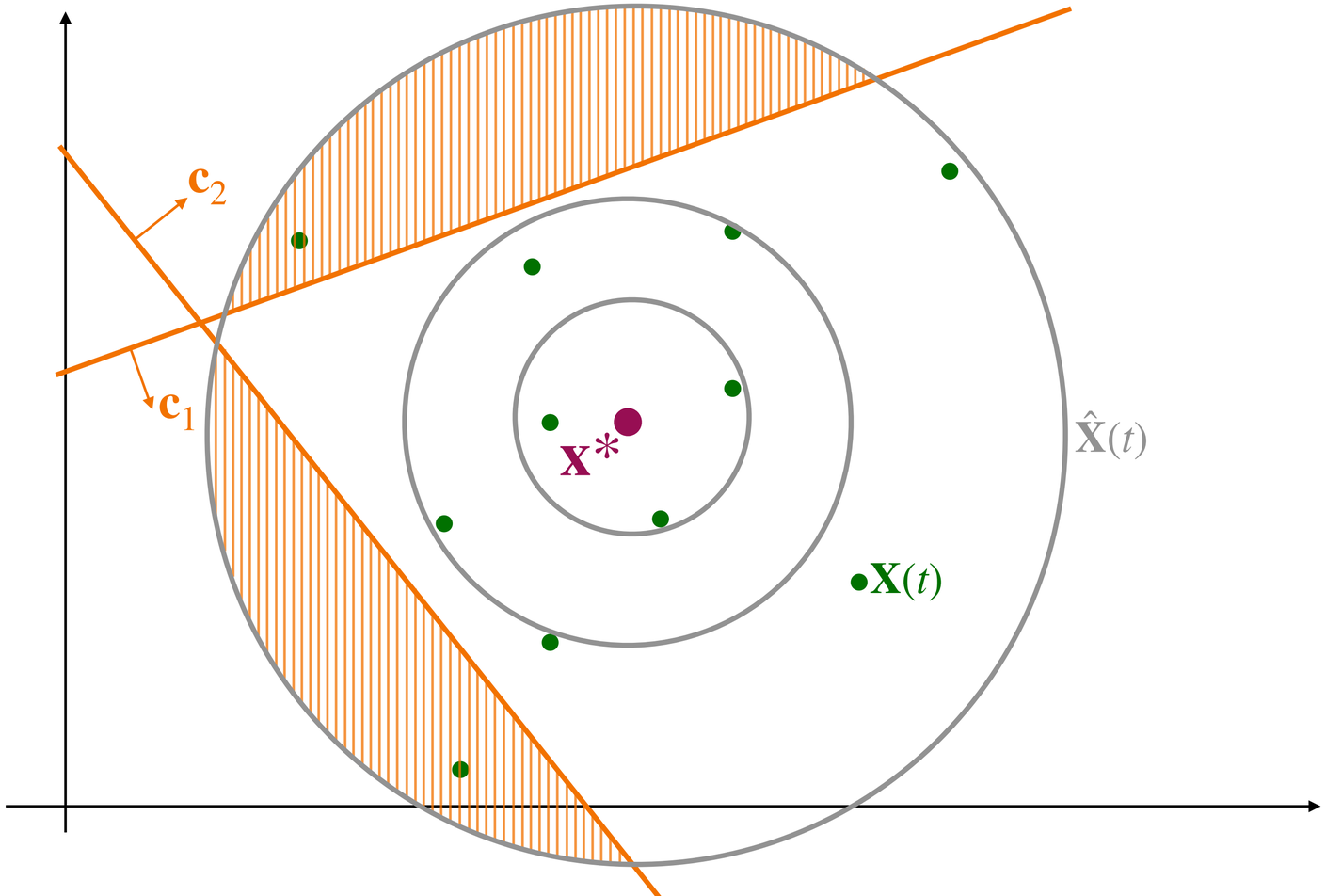
Classification Estimates

$\mathbf{x}^* \longrightarrow \{\mathbf{X}(1), \ldots, \mathbf{X}(t)\} \xrightarrow{f} \mathbf{Z}(t) \xrightarrow{g} \hat{\mathbf{Y}}(t)$

$\mathbf{x}^* \downarrow$

$\mathbf{y}(\mathbf{x}^*)$

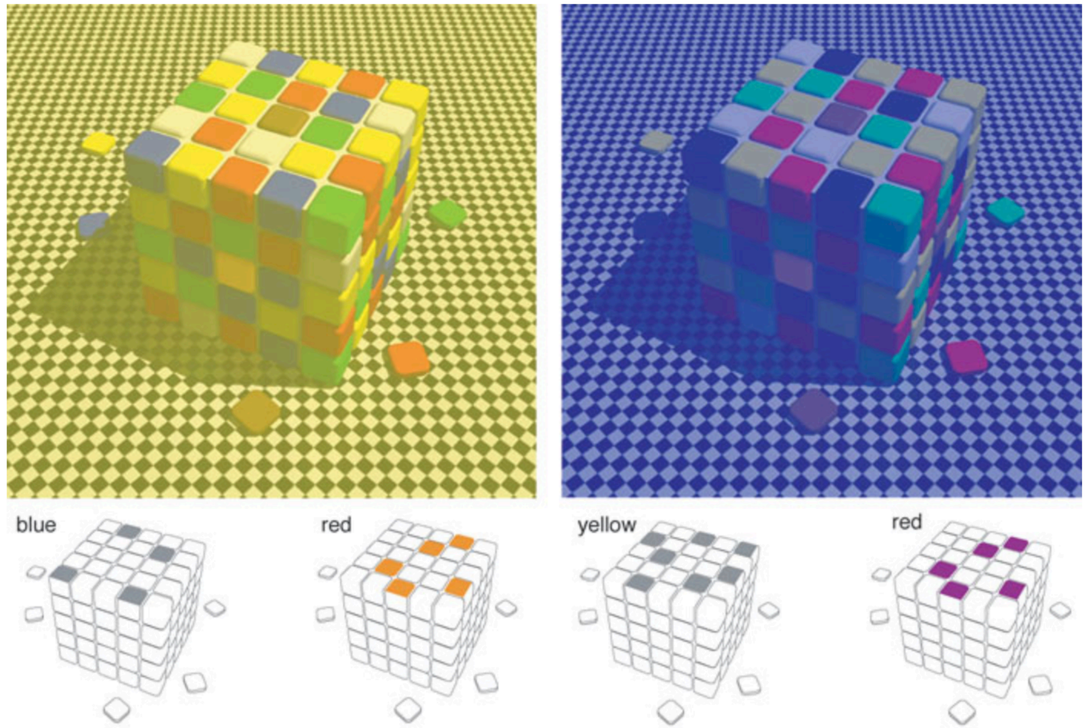
Ground truth classifications



CNS154 Problem Set → Disentangled Representations

3. **Color constancy** Inspect the two images at `https://engineering.purdue.edu/~bouman/ece637/notes/ColorConstancy/color/`. Notice that (a) tiles that you would name with the same color on the two sides actually emit a very different spectrum; and (b) tiles with the same spectrum seem to have very different color. The physical stimulus seems to be divorced from the color sensation. What are the possible benefits of this? Suggest two possible neural mechanisms linking stimulus and sensation, one involving adaptation, the other not. **6 pts**

Reprinted from Dale Purves, R. Beau Lotto, Surajit Nundy, "Why We See What We Do,", American Scientist, vol. 90, no. 3, page 236. www.americanscientist.org/template/AssetDetail/assetid/14755.



From CNS154 Homework 4 Question 3 (Meister et al, 2024)

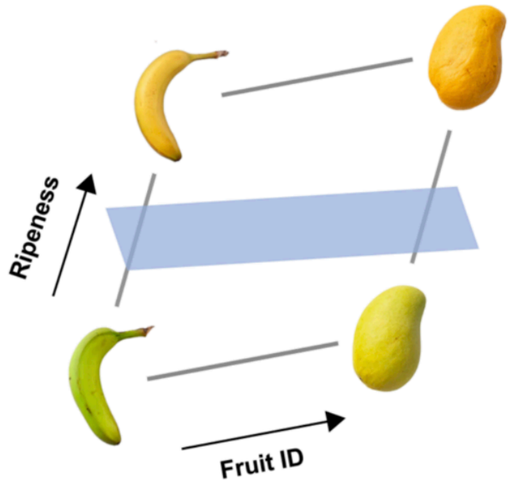
CNS154 Problem Set → Disentangled Representations

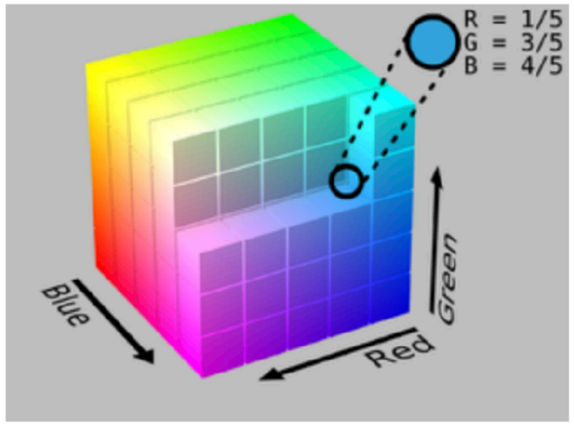
Disentangling Representations in RNNs through Multi-task Learning" by Pantelis Vafidis, Aman Bhargava, Antonio Rangel (coming to a journal near you)

Q3: Adaptation without Adaptation

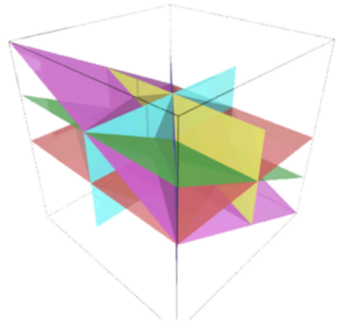
Any optimal decision making system will appear to "adapt"

- **Setup**: You're a beast in the wild making decisions about what to eat.





Color space (w/ noisy observations)



Decision boundaries in color space

From CNS154 Homework 4 Question 3 (Meister et al, 2024)

CNS154 Problem Set → Disentangled Representations

Disentangling Representations in RNNs through Multi-task Learning" by Pantelis Vafidis, Aman Bhargava, Antonio Rangel (coming to a journal near you)

Q3: Adaptation without Adaptation

Any optimal decision making system will appear to "adapt"

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,

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

From CNS154 Homework 4 Question 3 (Meister et al, 2024)

Roadmap

Proving Disentangled Representation Theorem

- **Problem statement**: Optimal multi-classifications $\hat{\mathbf{Y}}(t)$ from Gaussian-noised observations ($\mathbf{X}(t) \sim \mathbf{x}^* + \mathcal{N}(0, \sigma I_D)$).
  - Model has latent variable $\mathbf{Z}(t)$. Does it represent an $\mathbf{x}^*$ estimate? **(YES)**
- **Motivation/simple example**: "Pie slice" classifications in $\hat{\mathbf{Y}}(t) = g(\mathbf{Z}(t))$.
- **Conceptual insight**: Each classification $\hat{Y}_i(t)$ represents a **distance** between $\mathbf{x}^*$ and decision boundary i.
- **Final theorem**: Closed-form solution for recovering $\mathbf{x}^*$ estimate from $\mathbf{Z}(t)$.

+ testable hypotheses on disentangled representations with noise σ, N_{task}, D, nonlinearity f.

Roadmap

Proving Disentangled Representation Theorem

- **Problem statement**: Optimal multi-classifications $\hat{\mathbf{Y}}(t)$ from Gaussian-noised observations ($\mathbf{X}(t) \sim \mathbf{x}^* + \mathcal{N}(0, \sigma I_D)$).
  - Model has latent variable $\mathbf{Z}(t)$. Does it represent an $\mathbf{x}^*$ estimate? **(YES)**
- **Motivation/simple example**: "Pie slice" classifications in $\hat{\mathbf{Y}}(t) = g(\mathbf{Z}(t))$.
- **Conceptual insight**: Each classification $\hat{Y}_i(t)$ represents a **distance** between $\mathbf{x}^*$ and decision boundary i.
- **Final theorem**: Closed-form solution for recovering $\mathbf{x}^*$ estimate from $\mathbf{Z}(t)$.

+ testable hypotheses on disentangled representations with noise σ, N_{task}, D, nonlinearity f.

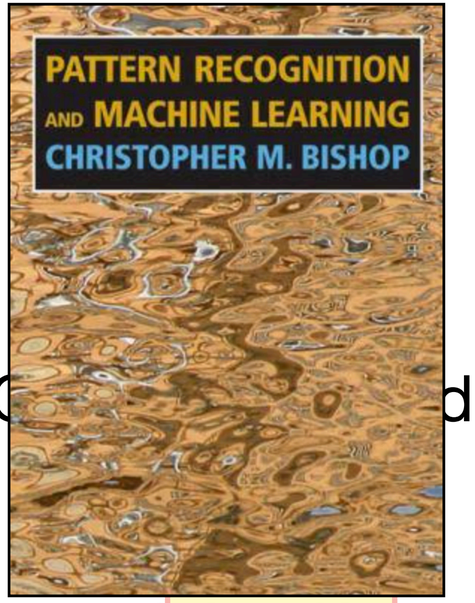
Roadmap

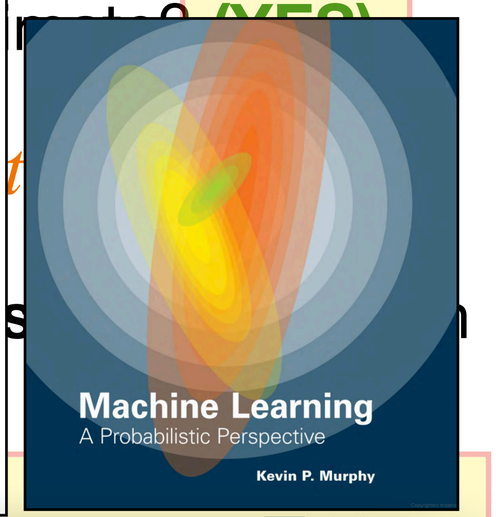
Proving Disentangled Representation Theorem

- **Problem statement**: Optimal multi-classifications $\hat{\mathbf{Y}}(t)$ from Gaussian-noised observations ($\mathbf{X}(t) \sim \mathbf{x}^* + \mathcal{N}(0, \sigma I_D)$).
  - Model has latent variable $\mathbf{Z}(t)$. Does it represent an $\mathbf{x}^*$ estimate? **(YES)**
- **Motivation/simple example**: "Pie slice" classifications in $\hat{\mathbf{Y}}(t) = g(\mathbf{Z}(t))$.
- **Conceptual insight**: Each classification $\hat{Y}_i(t)$ represents a **distance** between $\mathbf{x}^*$ and decision boundary i.
- **Final theorem**: Closed-form solution for recovering $\mathbf{x}^*$ estimate from $\mathbf{Z}(t)$.

  + testable hypotheses on disentangled representations with noise σ, N_{task}, D, nonlinearity f.

Contents



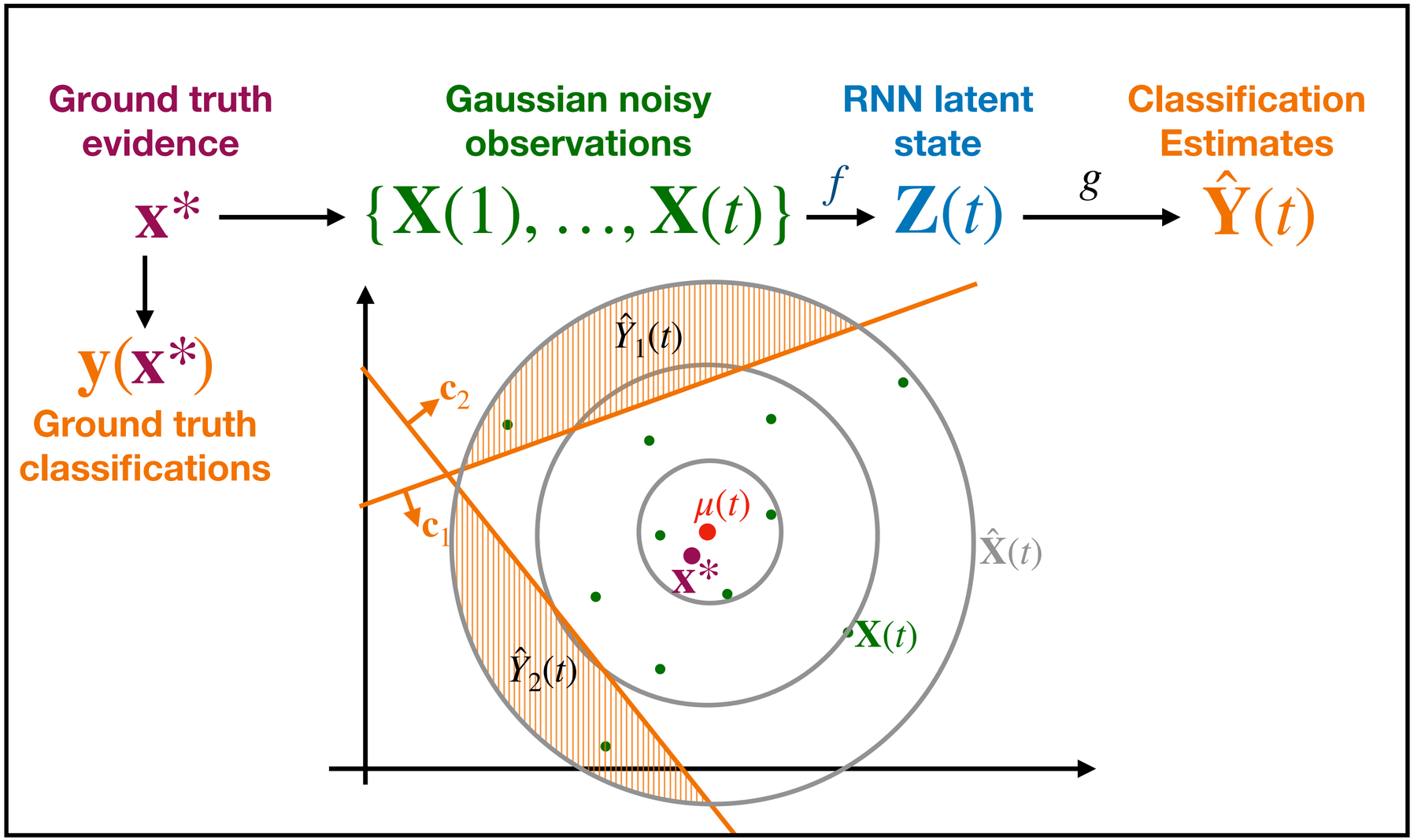


- **Final theorem**: Closed-form solution for recovering $\mathbf{x}^*$ estimate from $\mathbf{Z}(t)$.

https://github.com/amanb2000/EngSci_Abridged/blob/main/pdf/ECE368.pdf

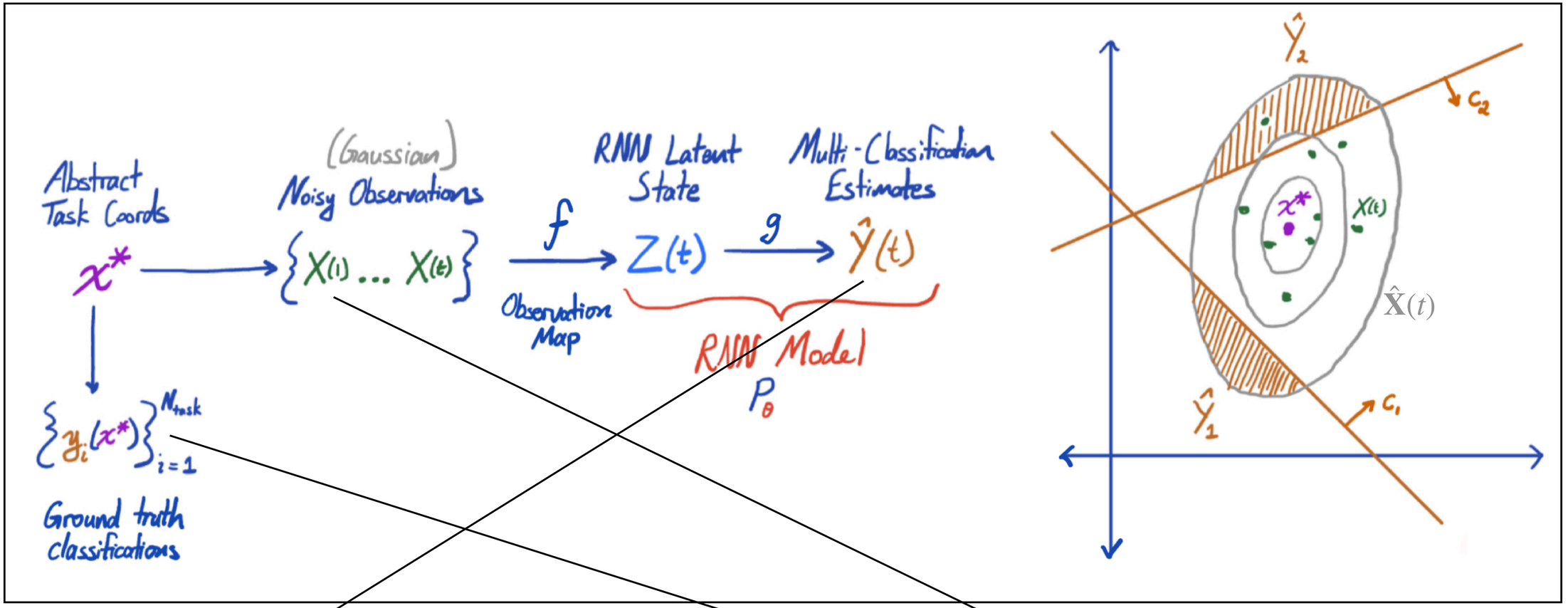
+ testable hypotheses on disentangled representations with noise σ, N_{task}, D, nonlinearity f.

Problem Setup: Multi-task evidence aggregation classification



Optimal Estimation: MAP = ML estimation for no prior on $\mathbf{x}^*$.

Problem Statement • Pie Slice Intuition • Classifications are Distances • Disentangled Reps Theorem



$$\hat{\mathbf{Y}}(t) = \underset{\mathbf{y}(\mathbf{x}^*)}{\arg\max}\, P\left(\mathbf{y}(\mathbf{x}^*) \mid \mathbf{X}(1), \ldots, \mathbf{X}(t)\right)$$