Emergence of **Bias** in DNN Predictions & Its Impact on **Trainability**

E. Francazi

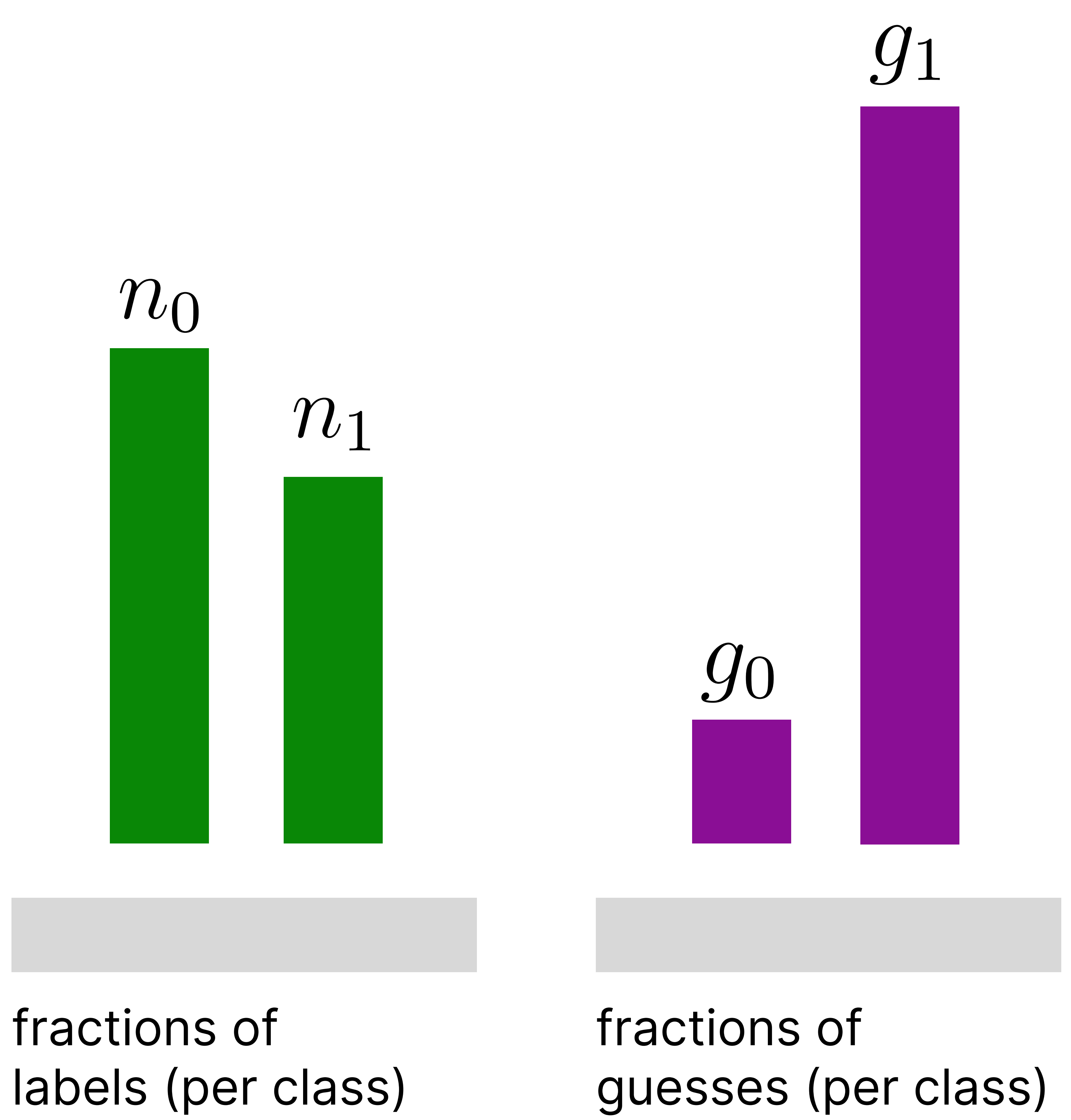
Outline

- I. **Initial Guessing Bias**
  - **Theory**: *When and why does initial bias appear?*
  - **Application**: *How can we control initial bias?*

- II. **Relevance for Learning**
  - **Implications**: *How does initial bias influence trainability?*

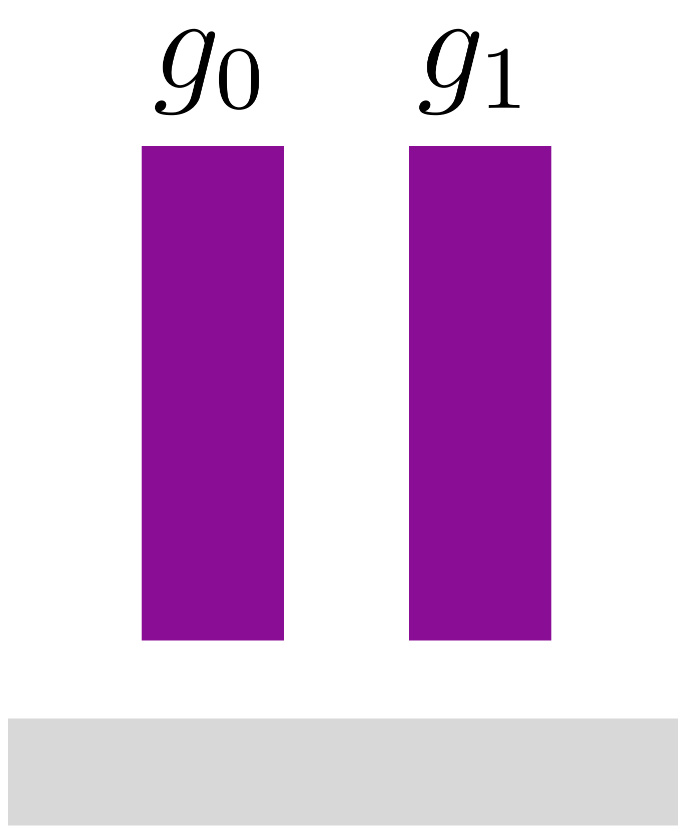
Bias In Supervised Learning

Bias: Model **predictions imbalanced** toward one of the classes



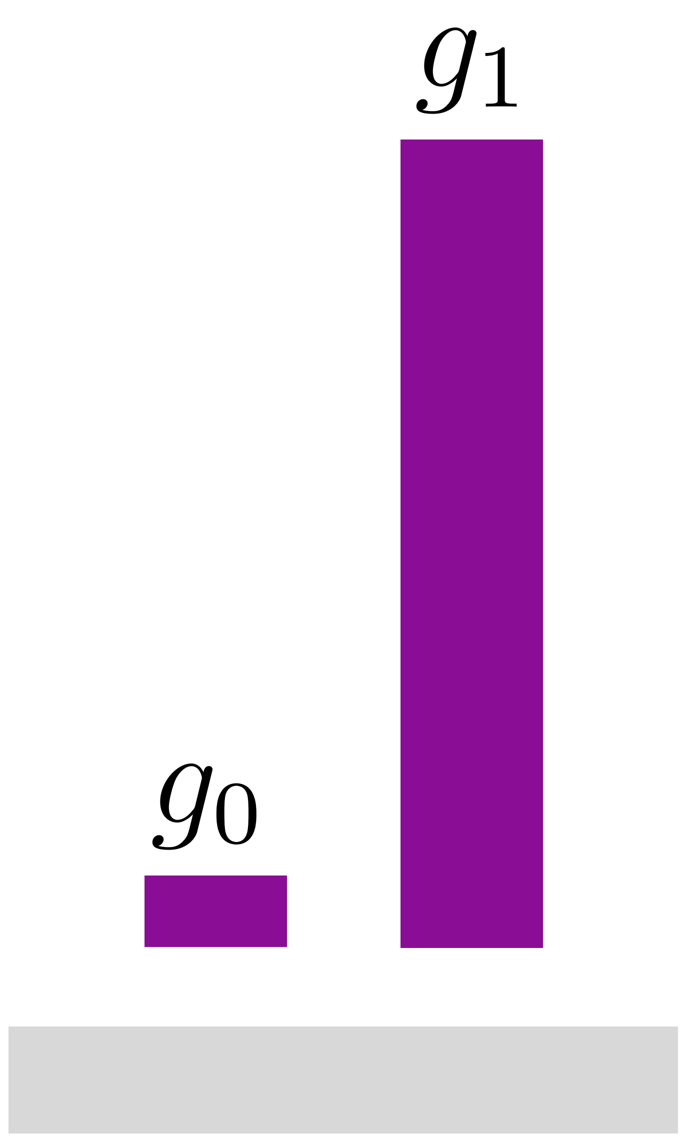
Predictive Behaviour Of Untrained Model

Neutrality



fractions of
guesses (per class)

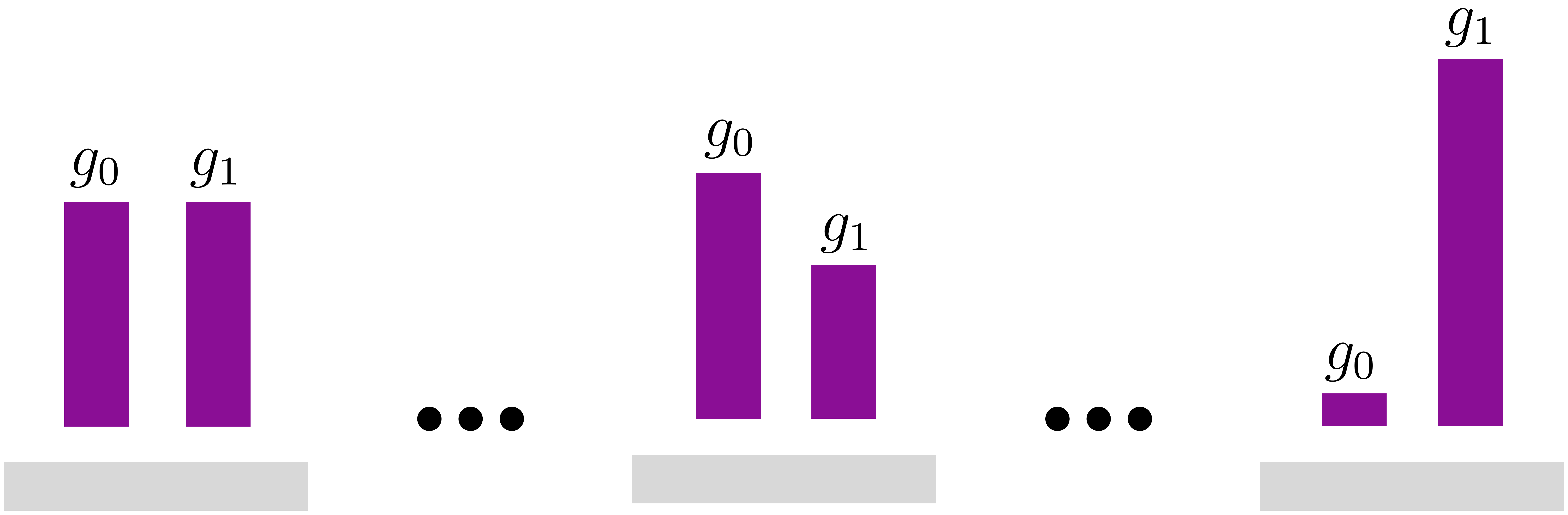
Deep Prejudice



fractions of
guesses (per class)

Predictive Behaviour Of Untrained Model

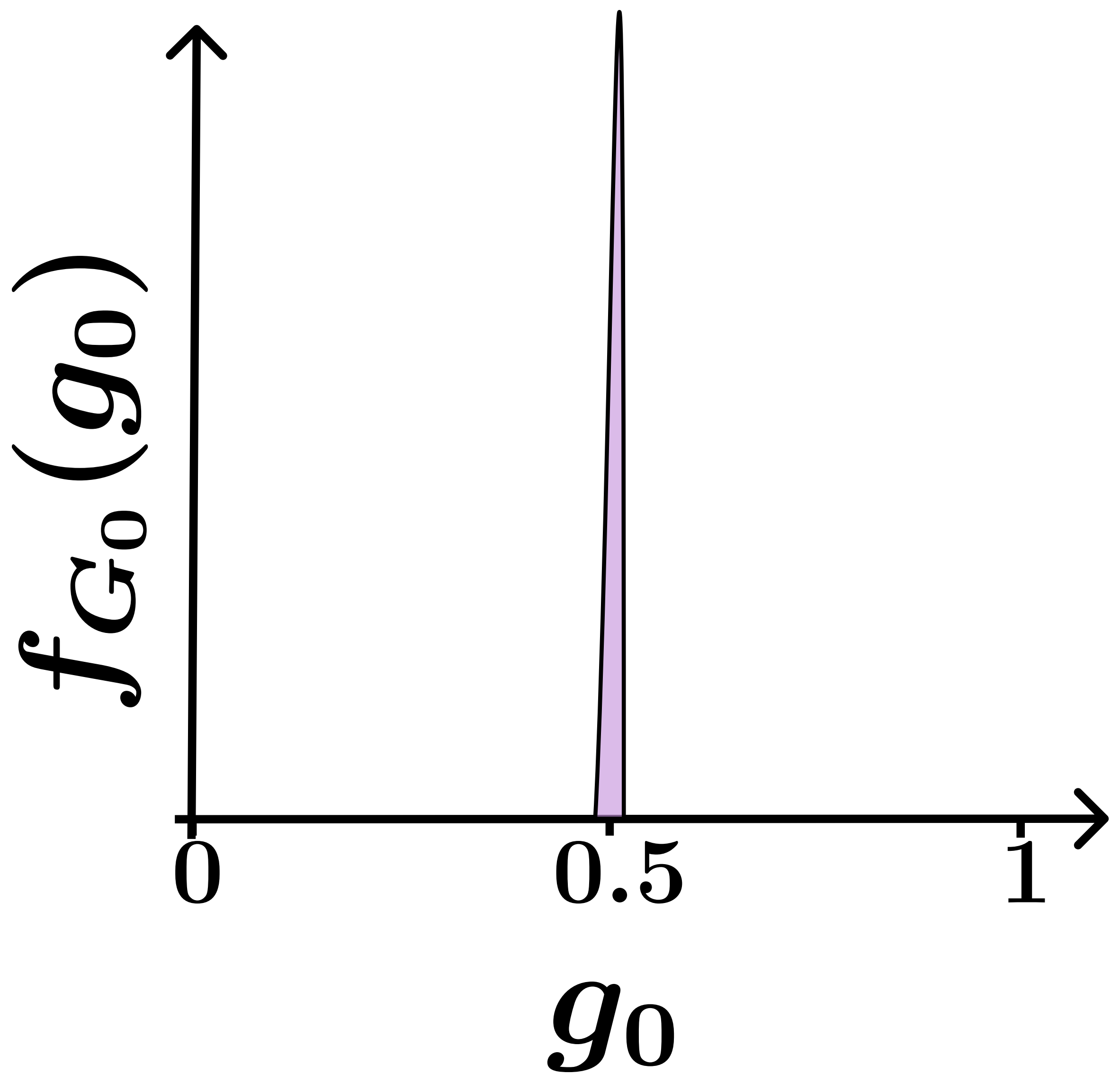
Neutrality **Weak Prejudice** **Deep Prejudice**



fractions of guesses (per class) fractions of guesses (per class) fractions of guesses (per class)

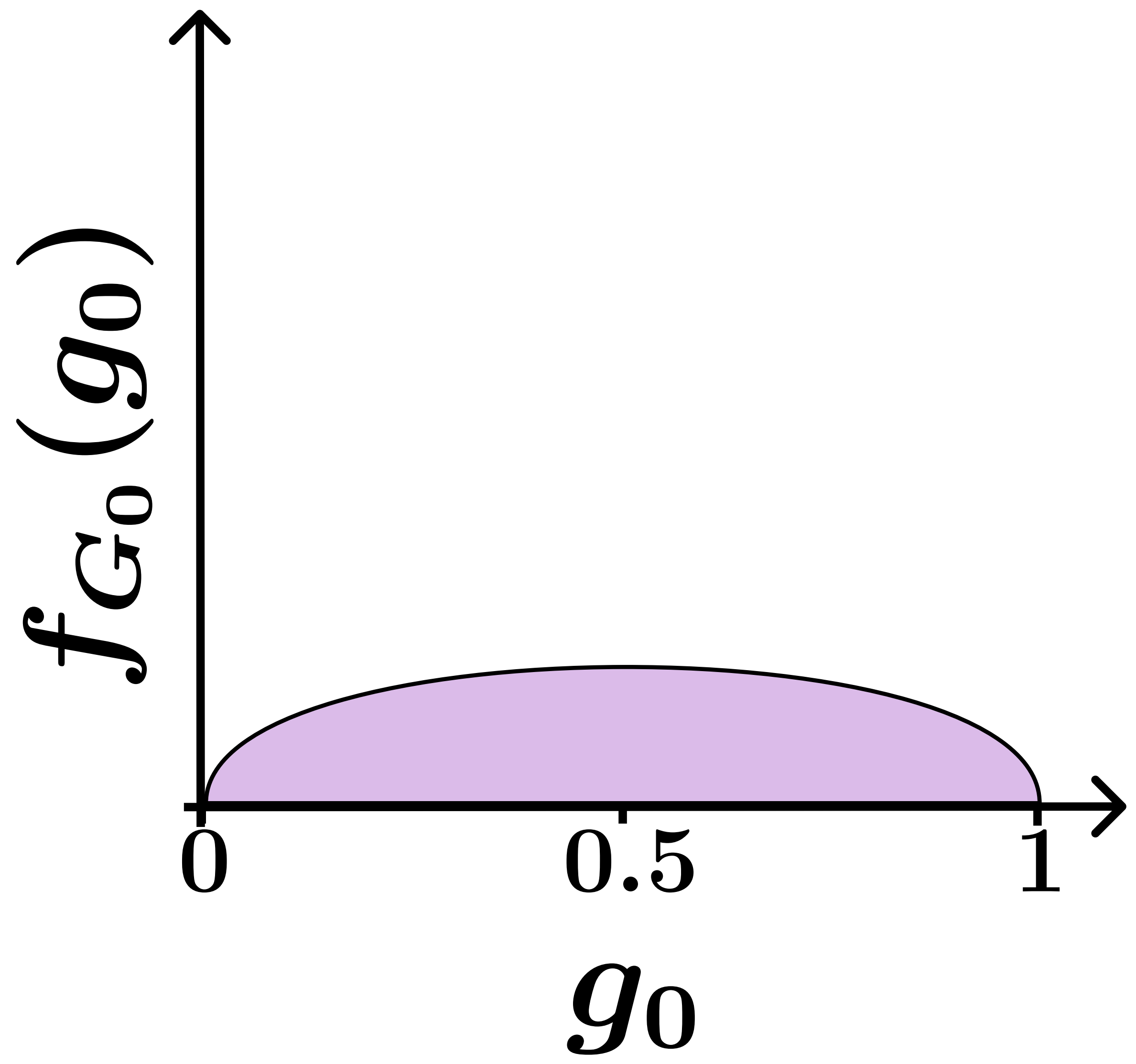
From Single Instance To Distribution

Neutrality



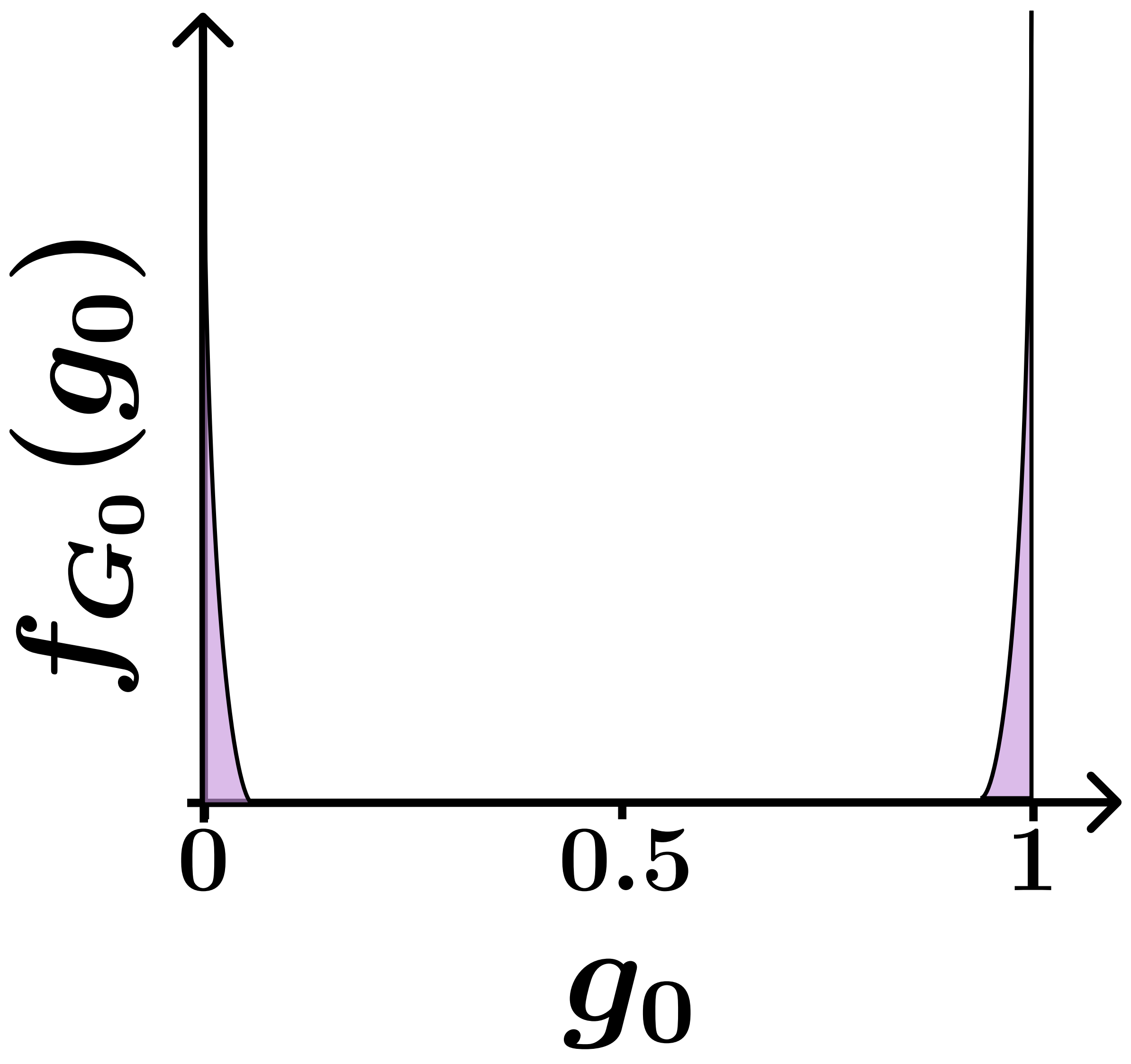
• • •

Weak Prejudice



• • •

Deep Prejudice



Initial Guessing Bias (IGB)

Untrained model on cats and dogs. Pass the whole (**balanced**) dataset through it. Is the model **neutral** at initialization?

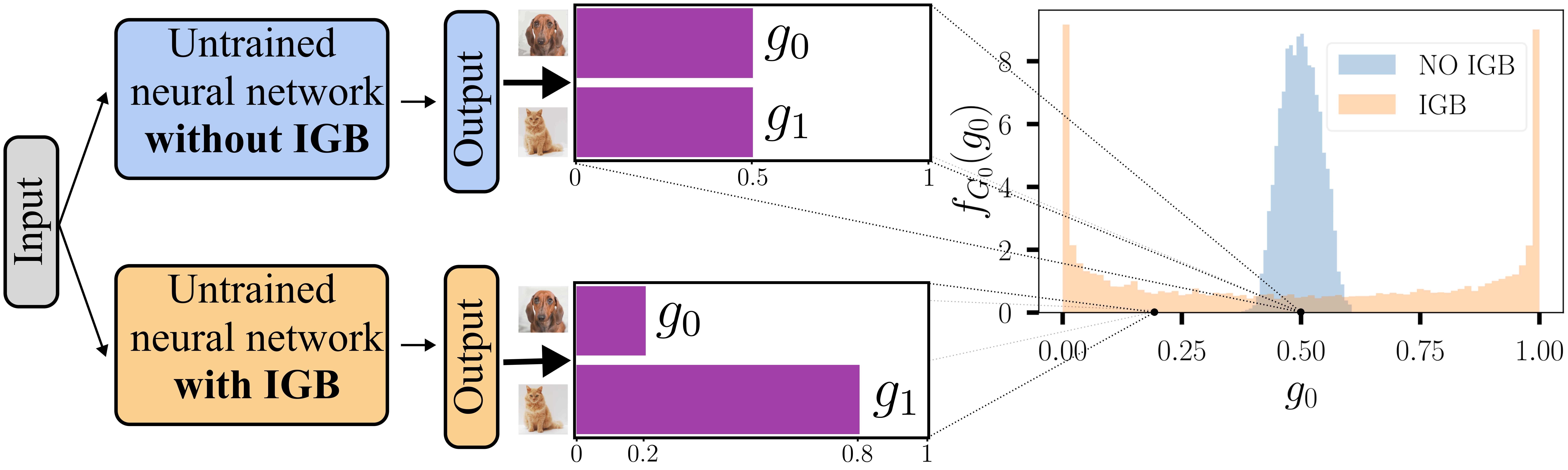
Initial Guessing Bias (IGB)

Untrained model on cats and dogs. Pass the whole (**balanced**) dataset through it.
Is the model **neutral** at initialization?

The answer **depends on the model**.

Initial Guessing Bias (IGB)

Untrained model on cats and dogs. Pass the whole (**balanced**) dataset through it. Is the model **neutral** at initialization?



The answer **depends on the model**.

IGB: Setting & Methods

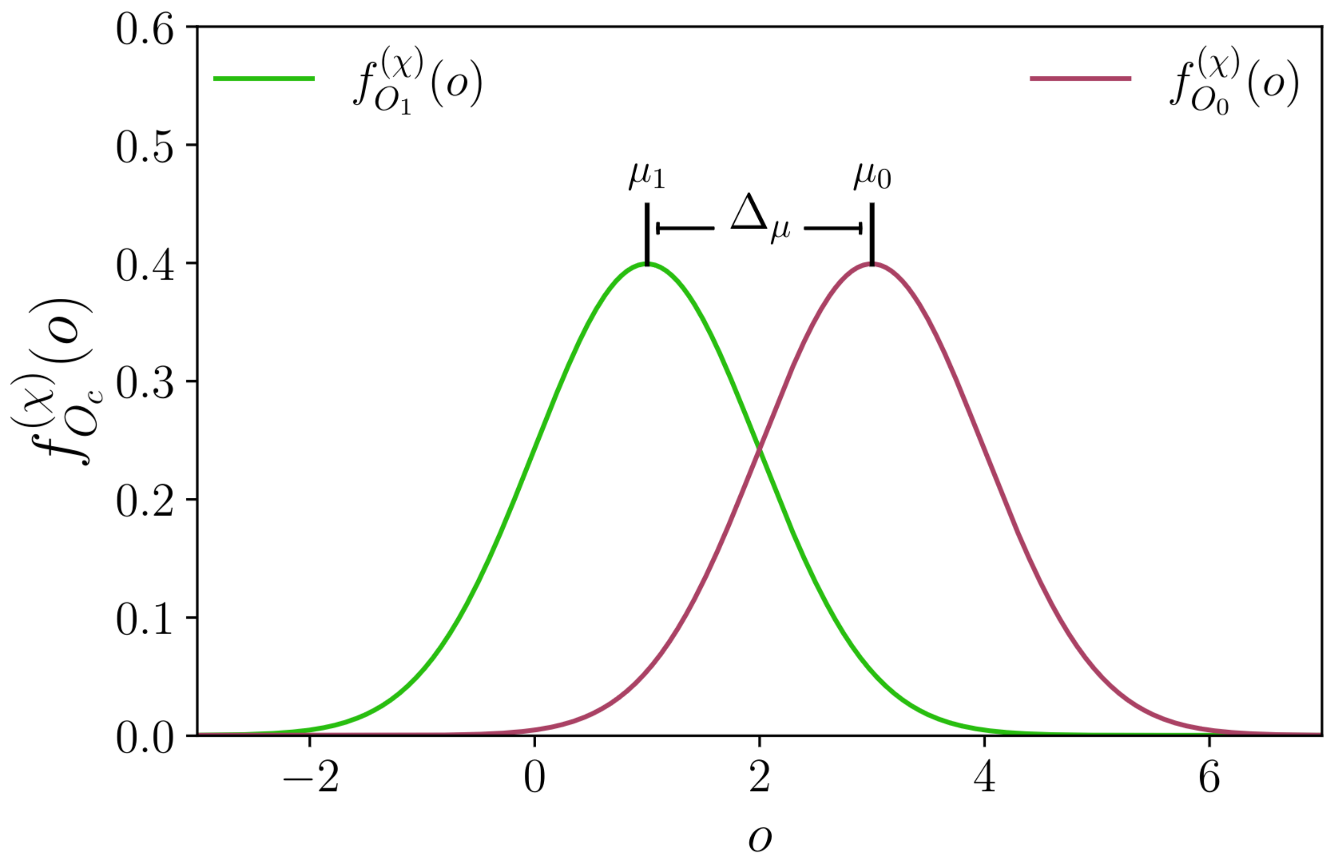
- **Data:** Dataset χ of random uncorrelated data (D datapoints)
- **Model:** Untrained with fixed weights $\mathcal{W}$
- **Process:**
  - Initialize DNN
  - Pass the whole dataset through the model (w/o changing weights)
  - Study p.d.f. of the outputs for the fixed set of weights $f_{O_c}^{(\chi)}(o)$
  - Study frequency of guesses

$$\lim_{D\to\infty} g_0(\mathcal{W}) = \mathbb{P}\left(O_0 > O_1 \mid \mathcal{W}\right)$$

Procedure

Distribution of outputs: $f_{O_c}^{(\chi)}(o) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(o; \mu_c, \mathrm{Var}_\chi(O))$

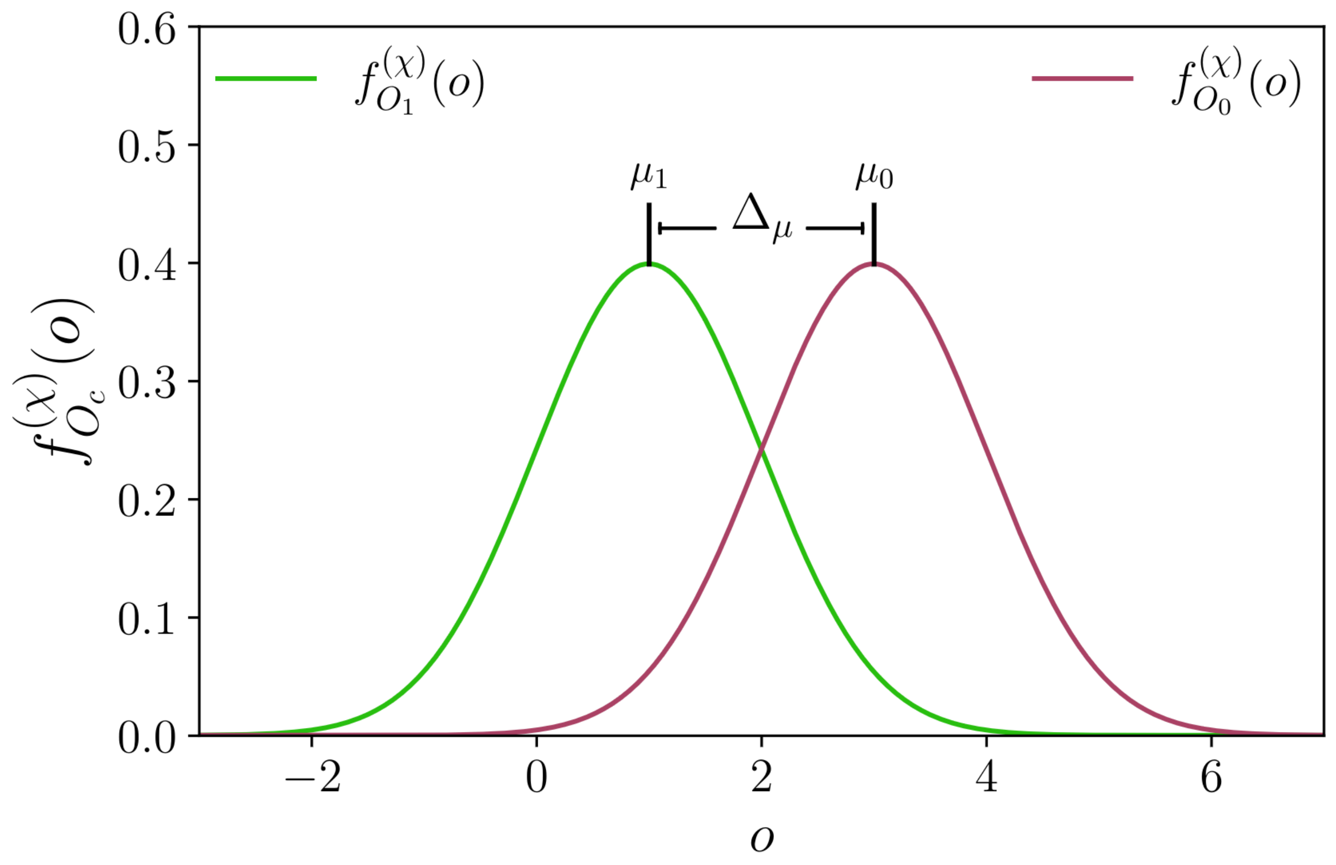
Distribution of centers: $f_{\mu_c}(m) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(m; 0, \mathrm{Var}_\mathcal{W}(\mu))$



Procedure

Distribution of outputs: $f_{O_c}^{(\chi)}(o) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(o; \mu_c, \mathrm{Var}_\chi(O))$

Distribution of centers: $f_{\mu_c}(m) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(m; 0, \mathrm{Var}_\mathcal{W}(\mu))$



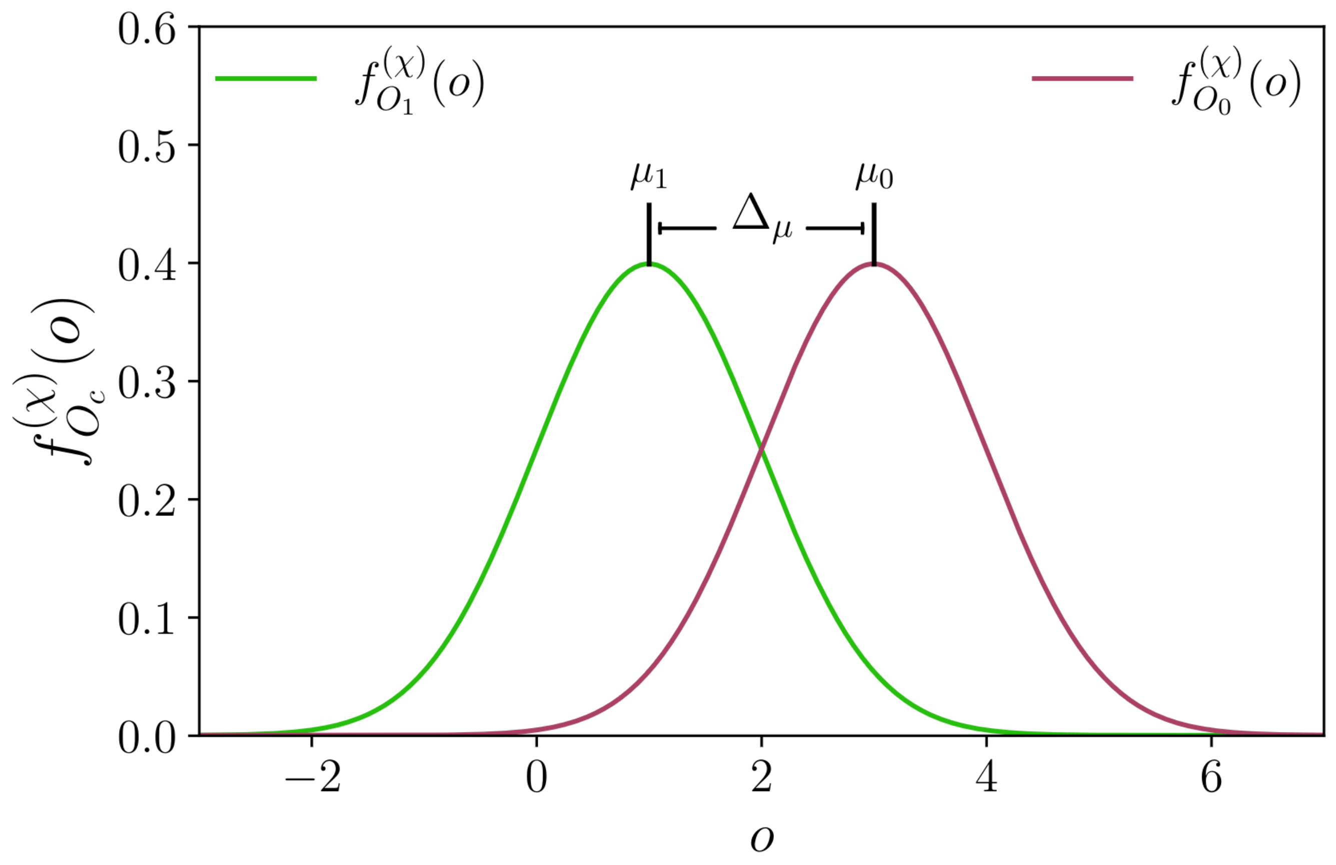
Quantify the level of IGB:

$$\gamma = \frac{\mathrm{Var}_\mathcal{W}(\mu)}{\mathrm{Var}_\chi(O)}$$

Procedure

Distribution of outputs: $f_{O_c}^{(\chi)}(o) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(o;\mu_c,\mathrm{Var}_\chi(O))$

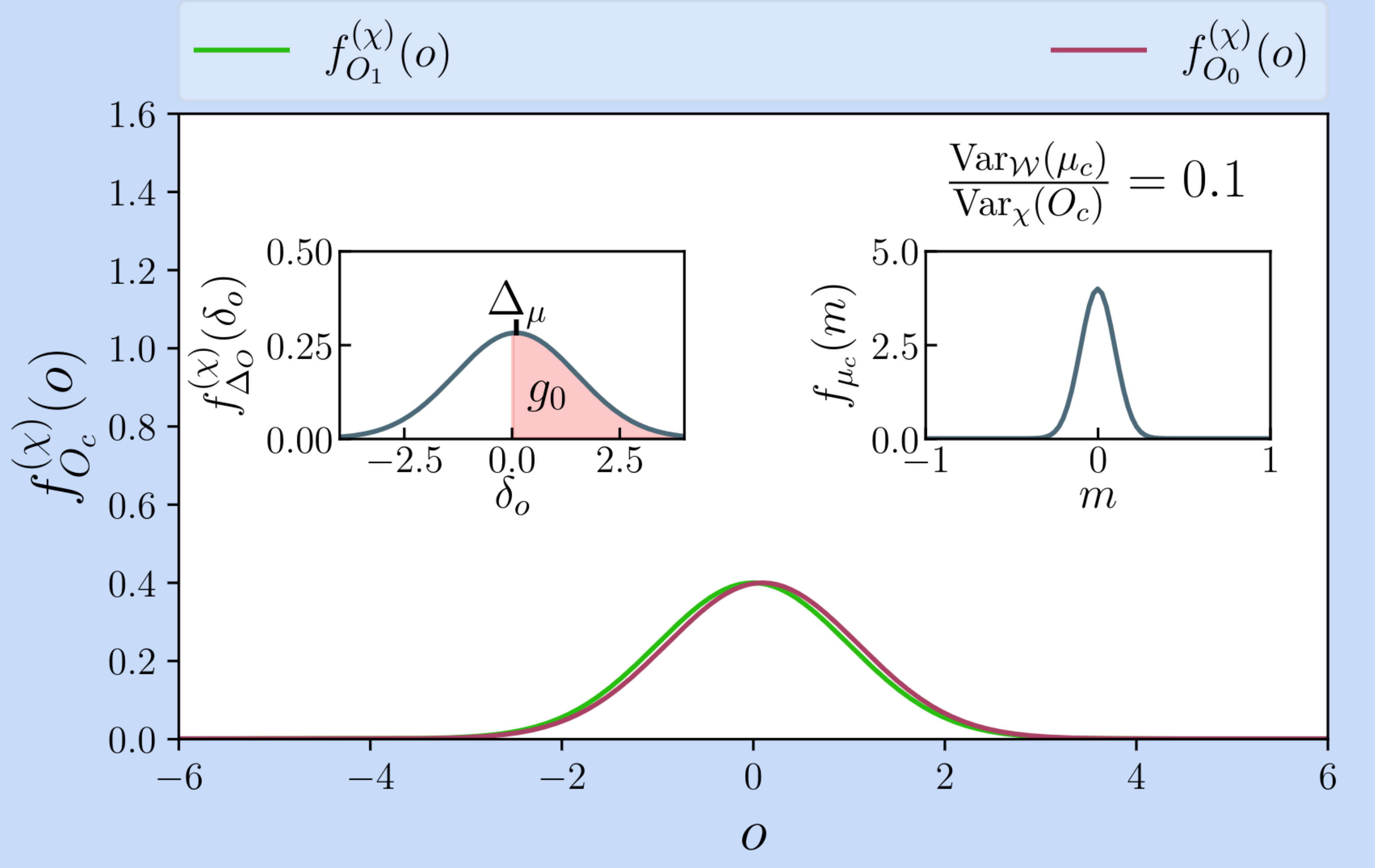
Distribution of centers: $f_{\mu_c}(m) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(m;0,\mathrm{Var}_\mathcal{W}(\mu))$



Quantify the level of IGB:

$$\gamma = \frac{\mathrm{Var}_\mathcal{W}(\mu)}{\mathrm{Var}_\chi(O)}$$

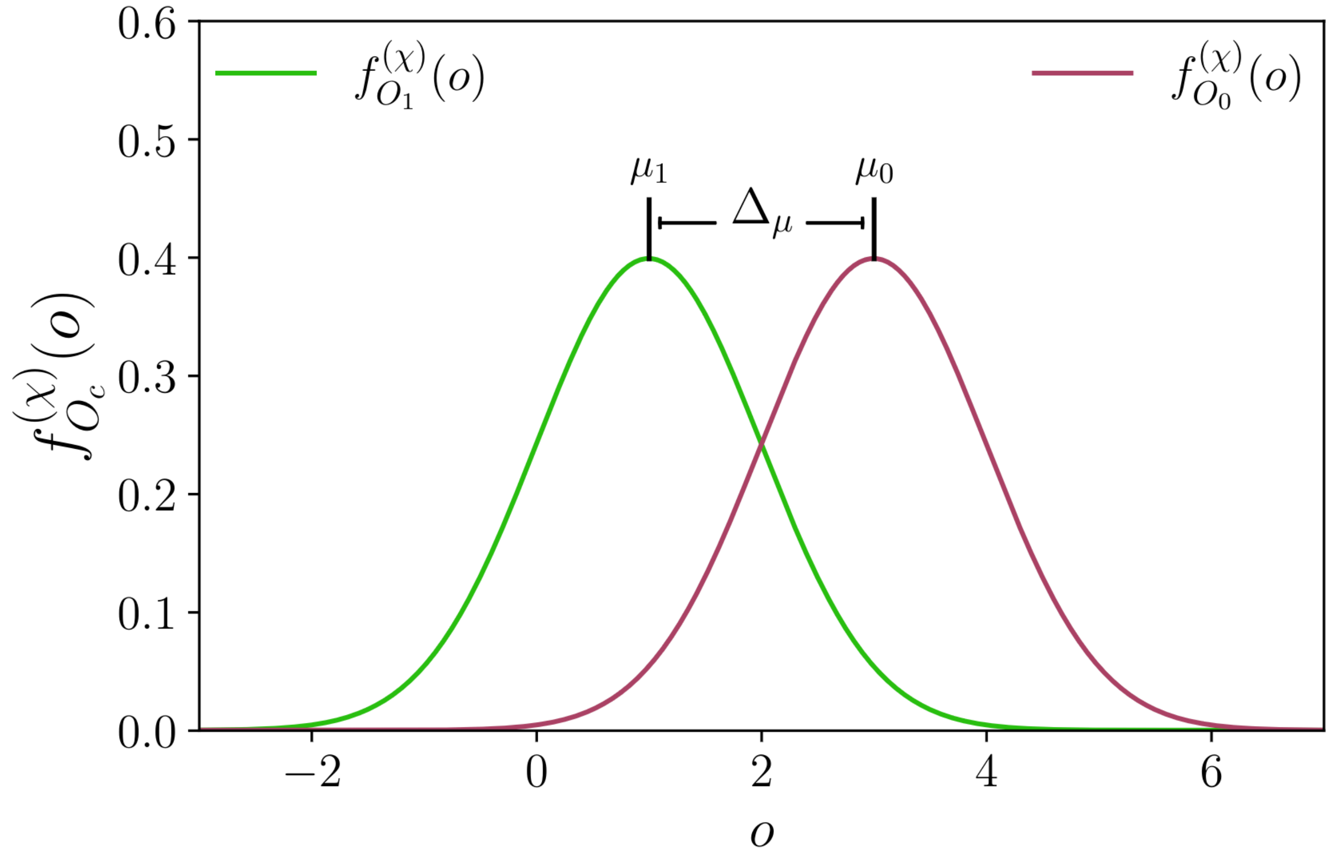
NO IGB



Procedure

Distribution of outputs: $f_{O_c}^{(\chi)}(o) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(o;\mu_c,\mathrm{Var}_\chi(O))$

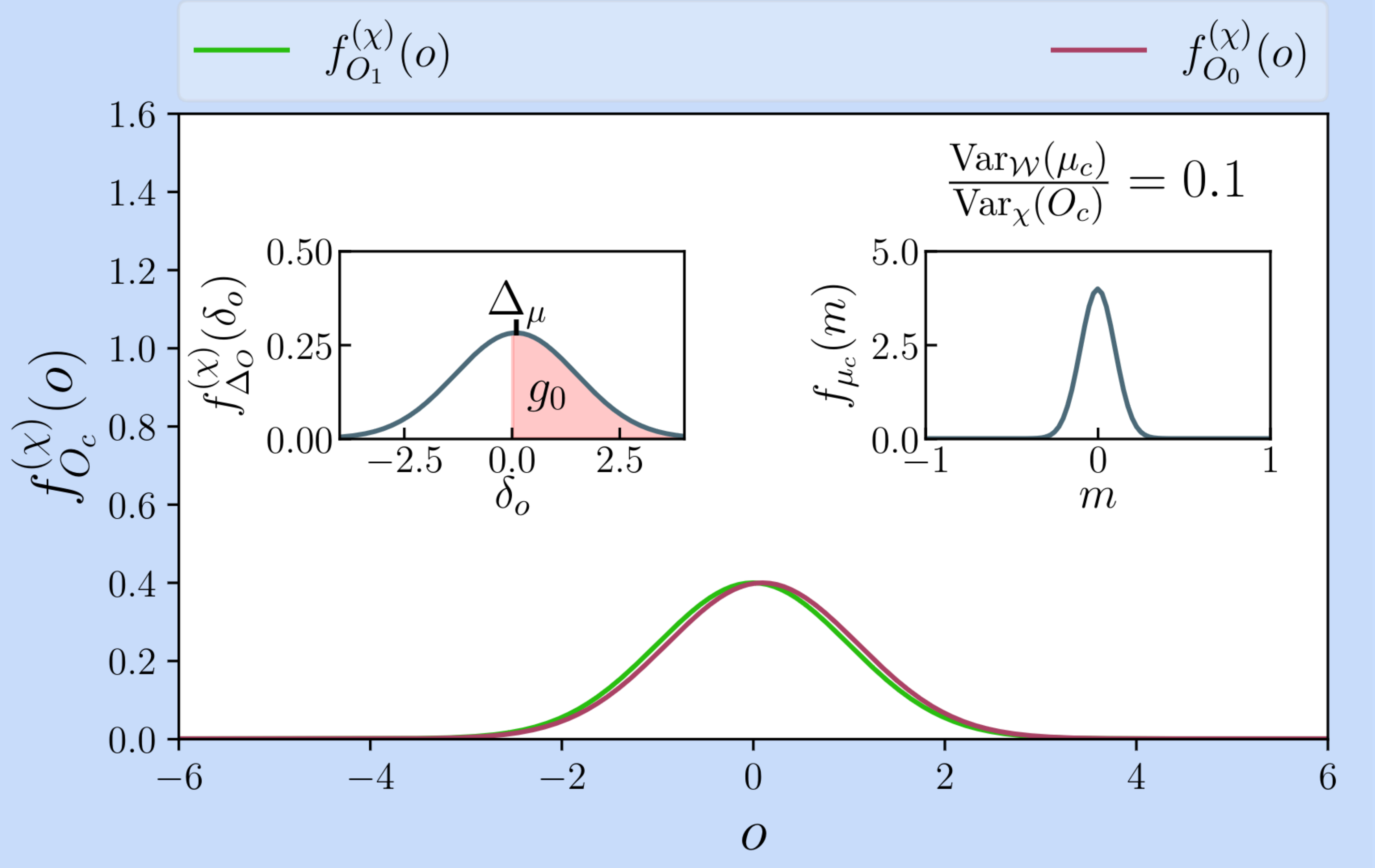
Distribution of centers: $f_{\mu_c}(m) \xrightarrow{|\mathcal{W}|\to\infty} \mathcal{N}(m;0,\mathrm{Var}_\mathcal{W}(\mu))$



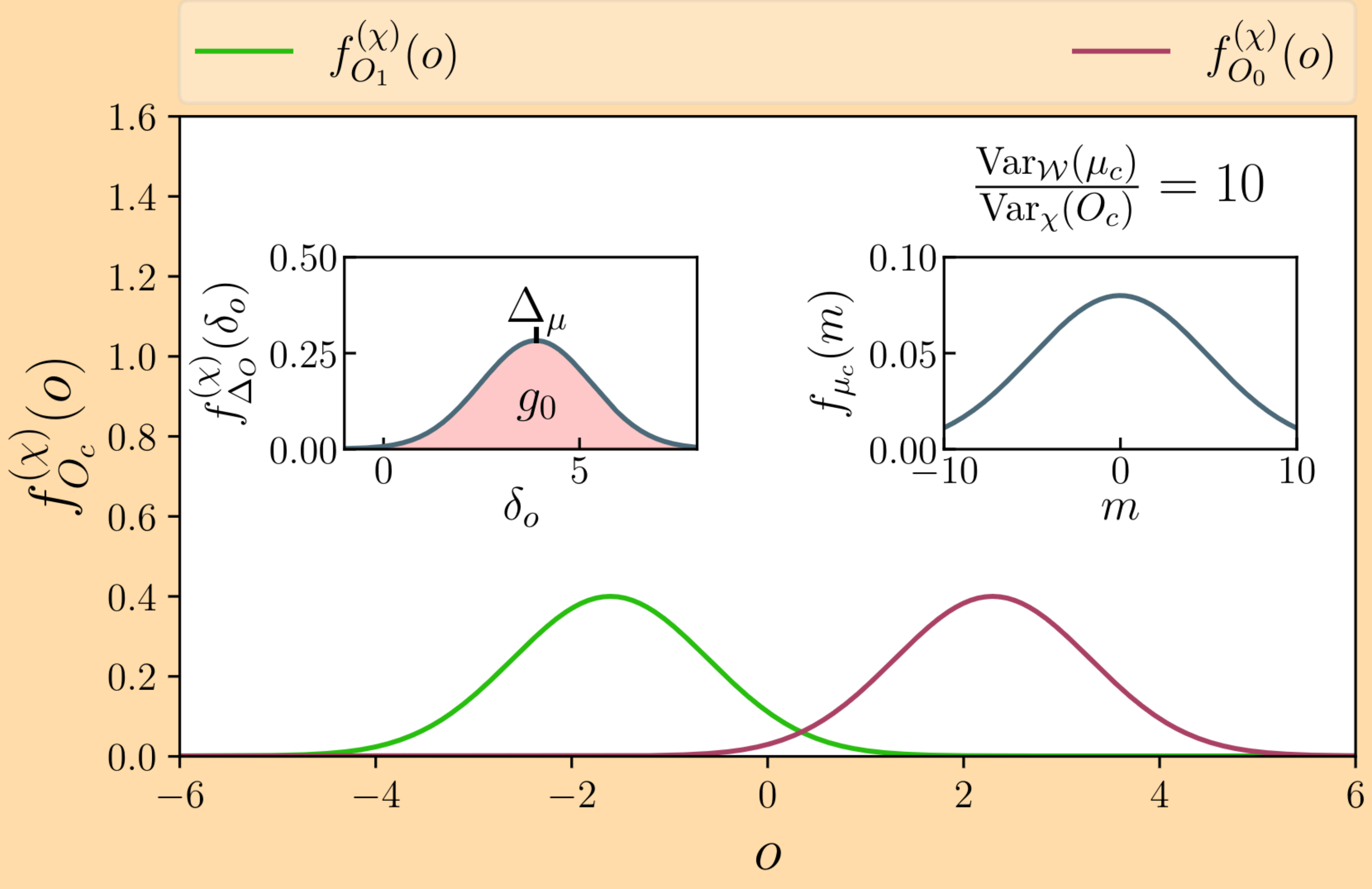
Quantify the level of IGB:

$$\gamma = \frac{\mathrm{Var}_\mathcal{W}(\mu)}{\mathrm{Var}_\chi(O)}$$

NO IGB

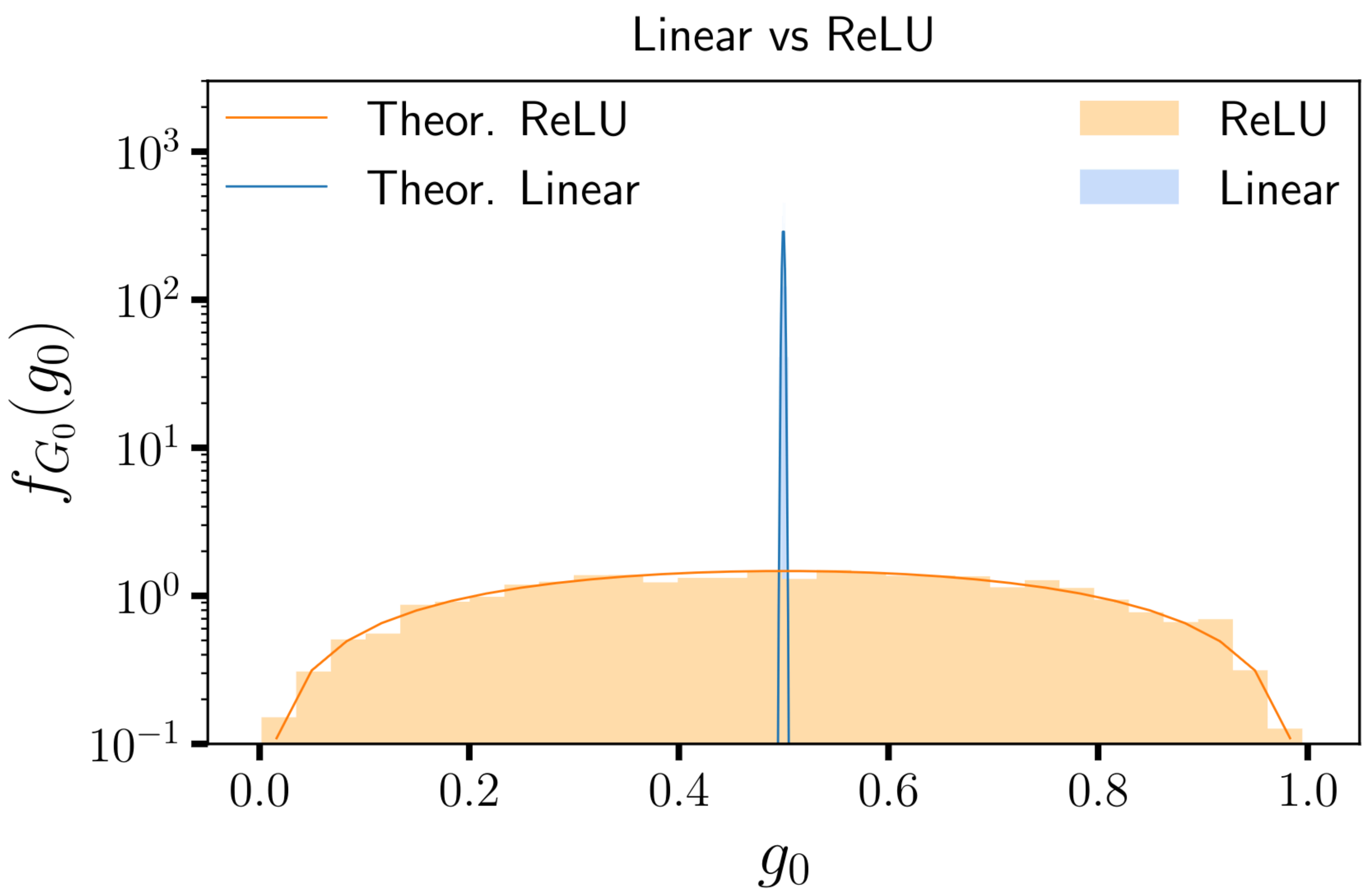


IGB



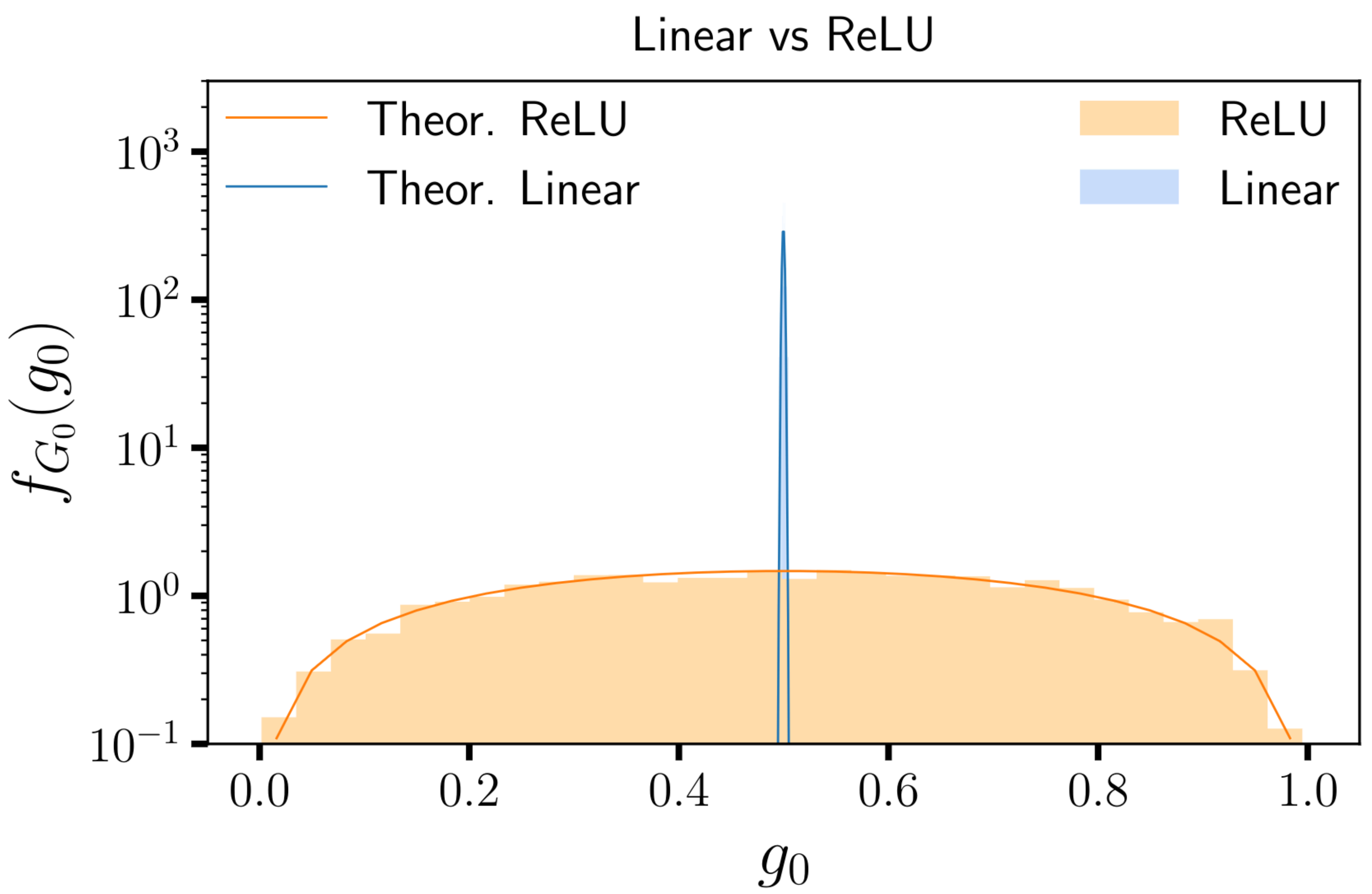
When Does IGB Appear?

- ReLU causes IGB, tanh does not
  - generic rule: activation has no IGB iff average over data of its output =0

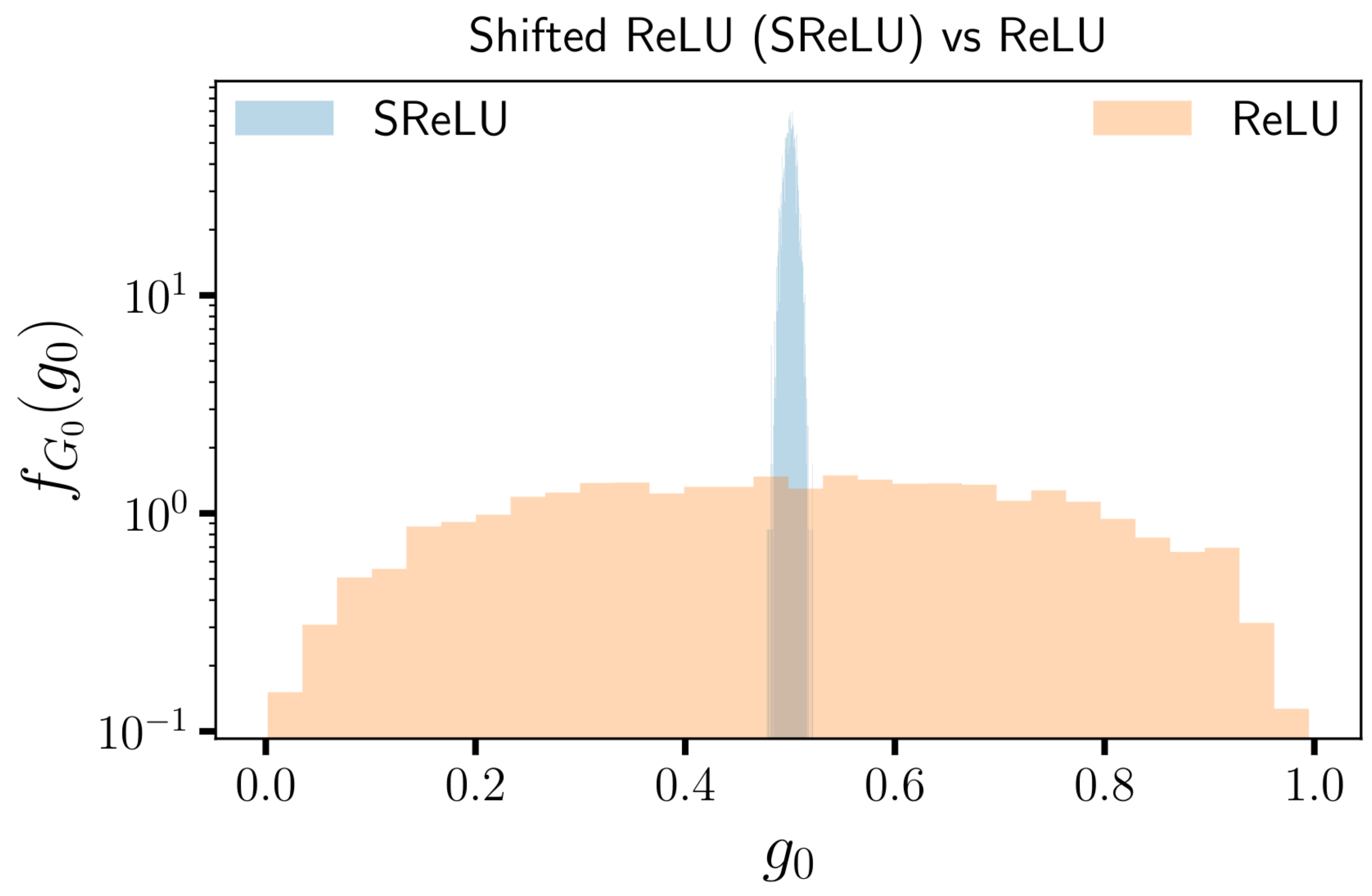


When Does IGB Appear?

- ReLU causes IGB, tanh does not
  - generic rule: activation has no IGB iff average over data of its output =0



- slightly modifying an activation function (e.g. by a shift) we can eliminate/trigger IGB

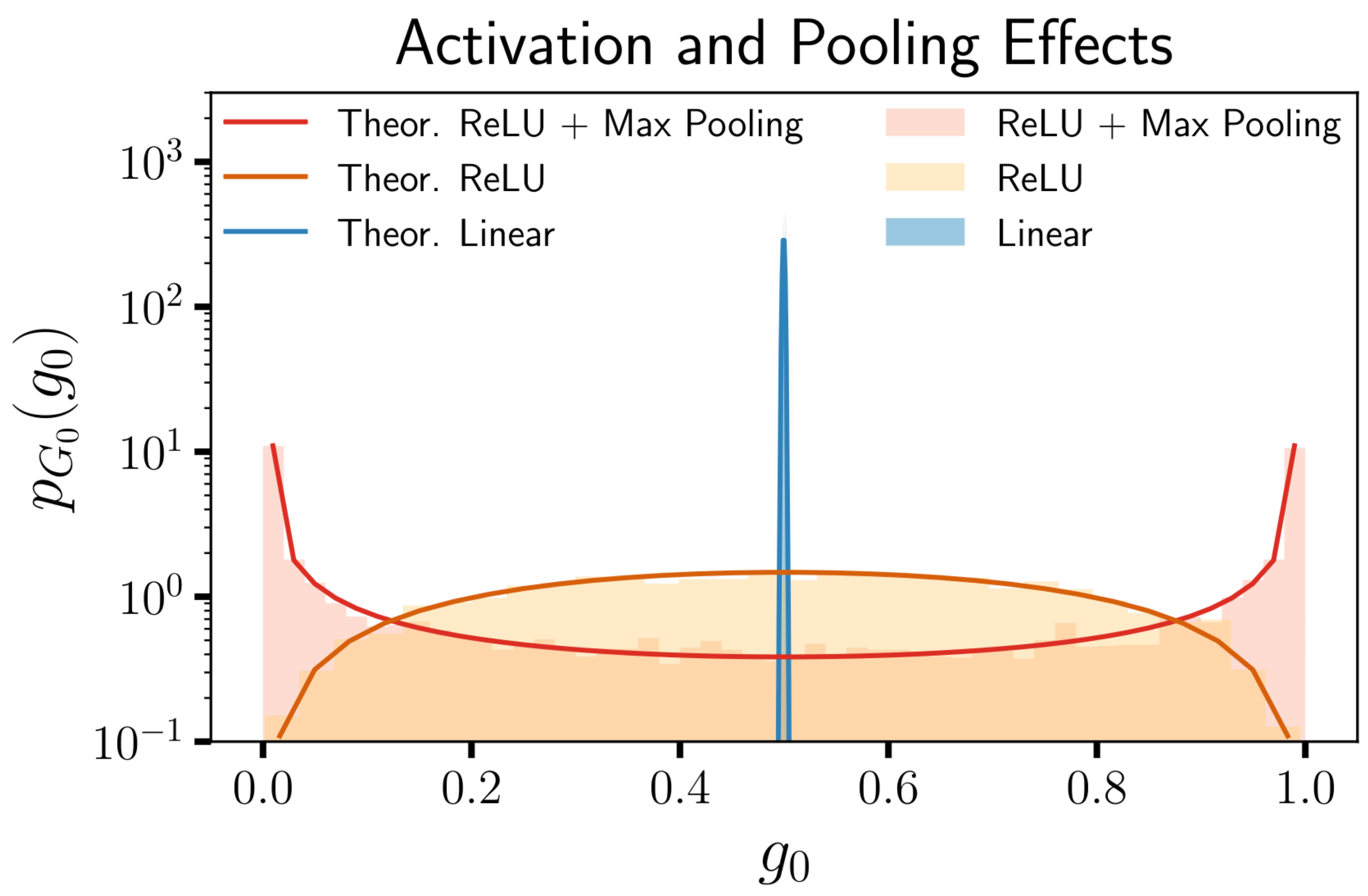


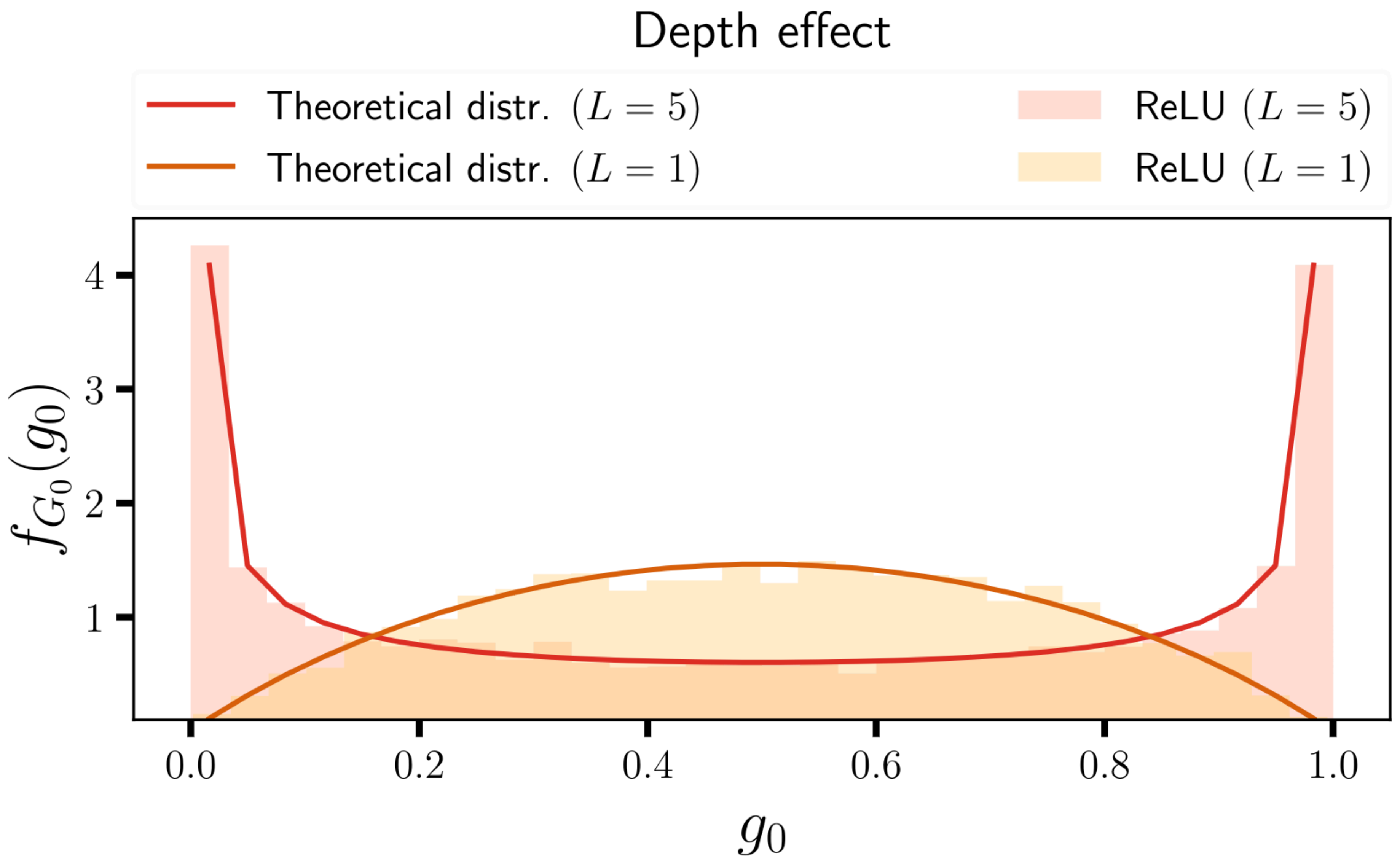
When Does IGB Appear?

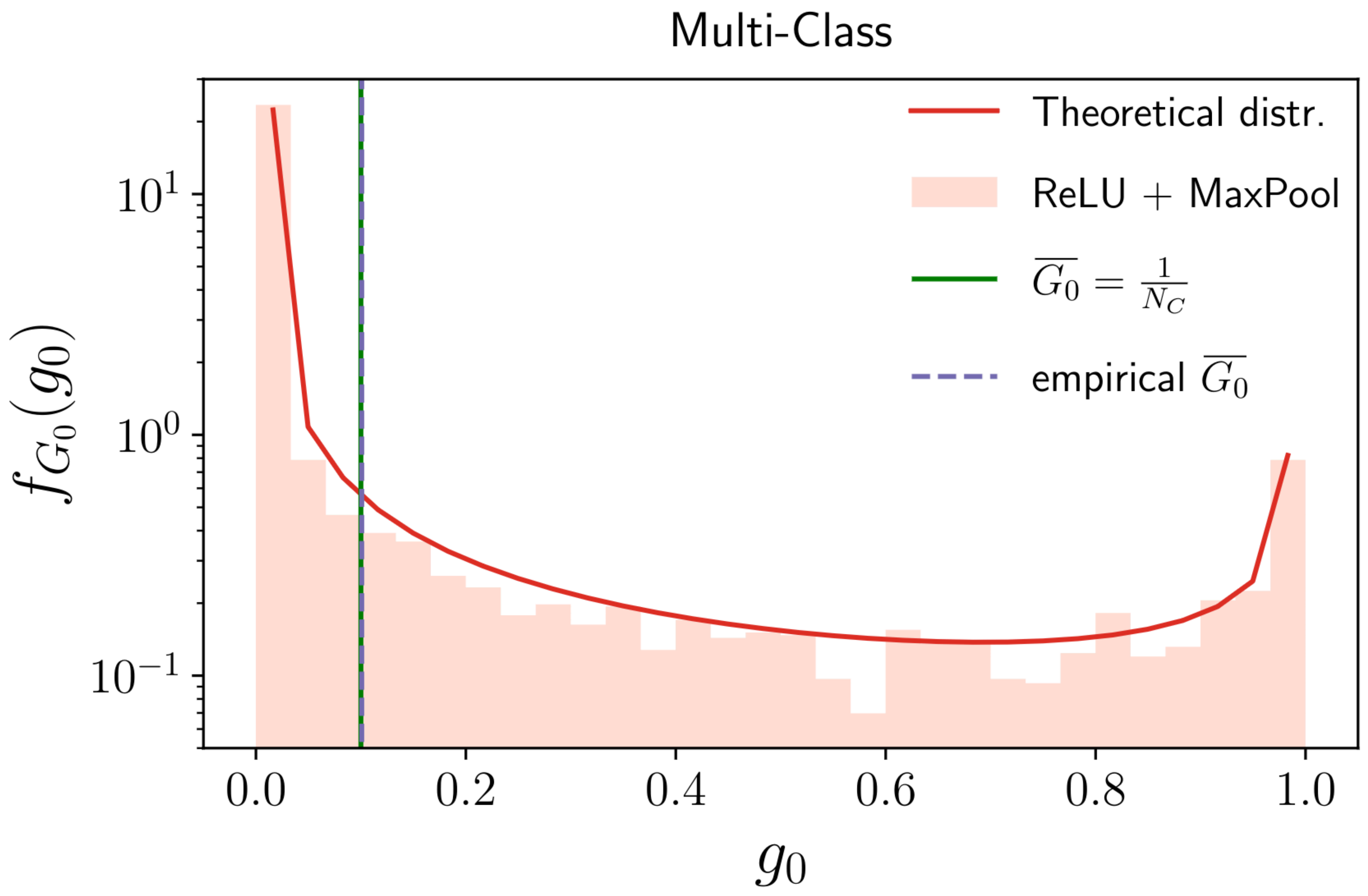
- generic rule: activation has no IGB iff average over data of its output =0

- Max pooling causes and exacerbates IGB

- Depth increases IGB



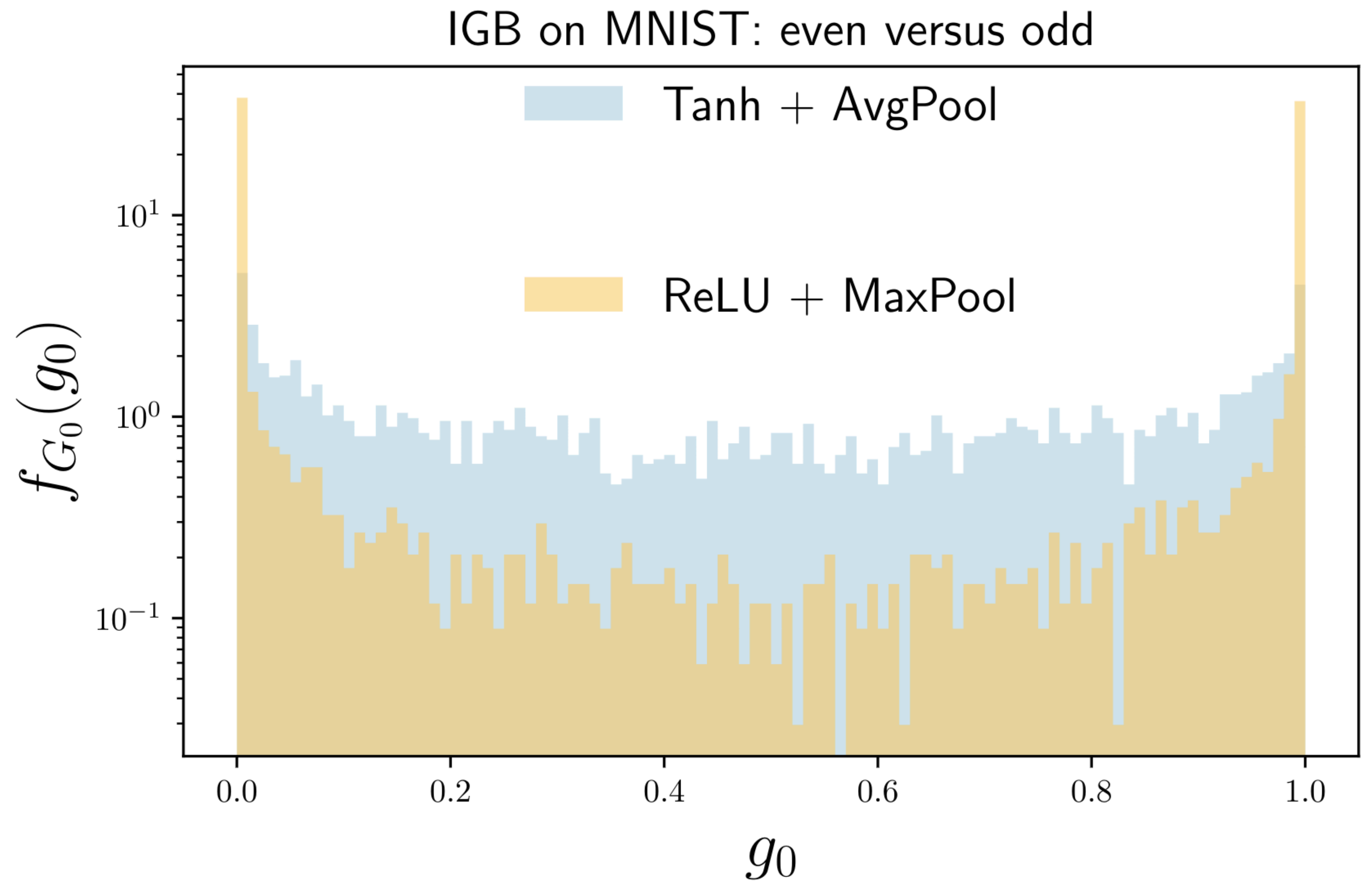




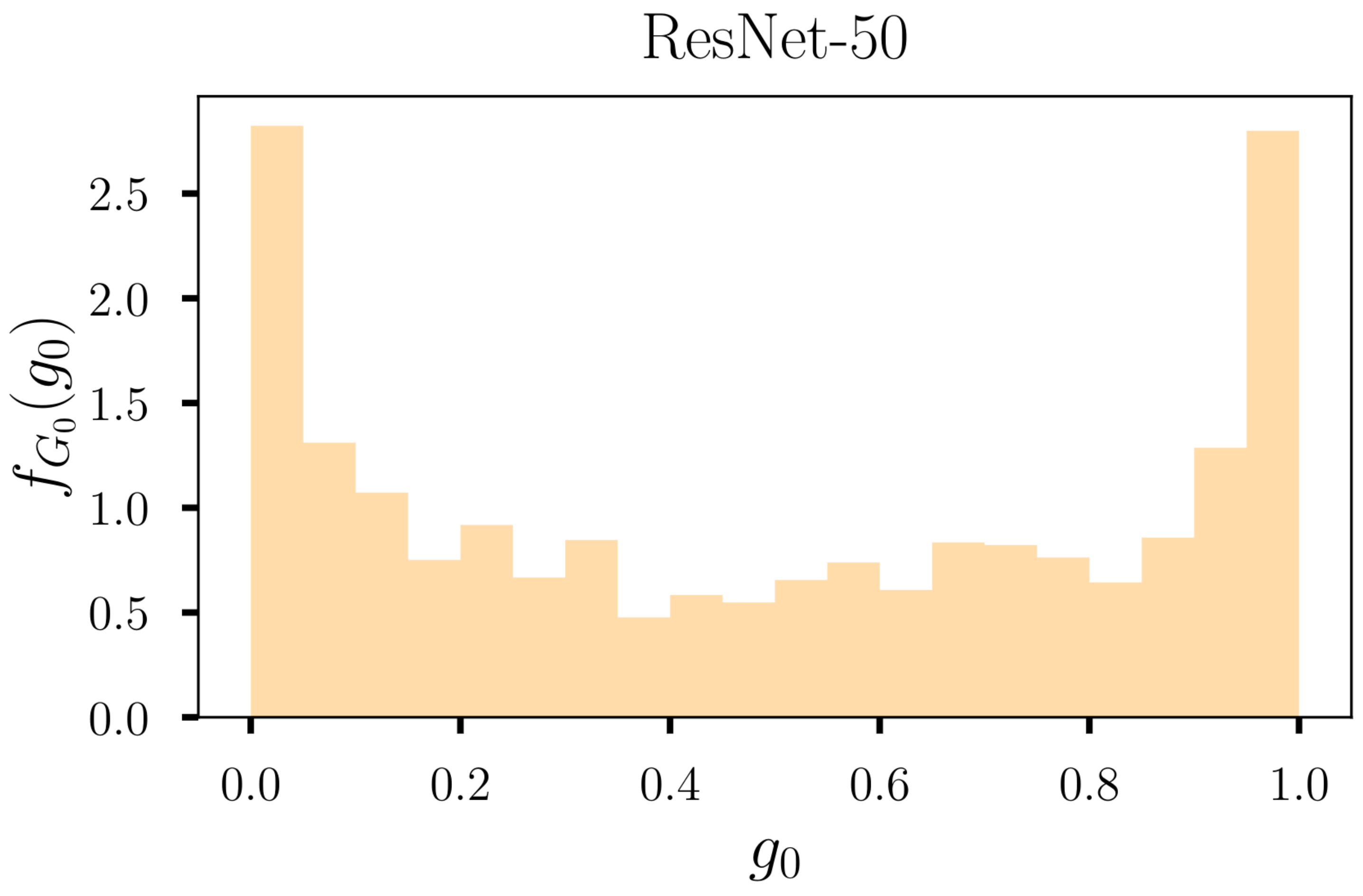
Real Settings

We Place Ourselves In A Setting Where The Effect Of IGB Is Minimal

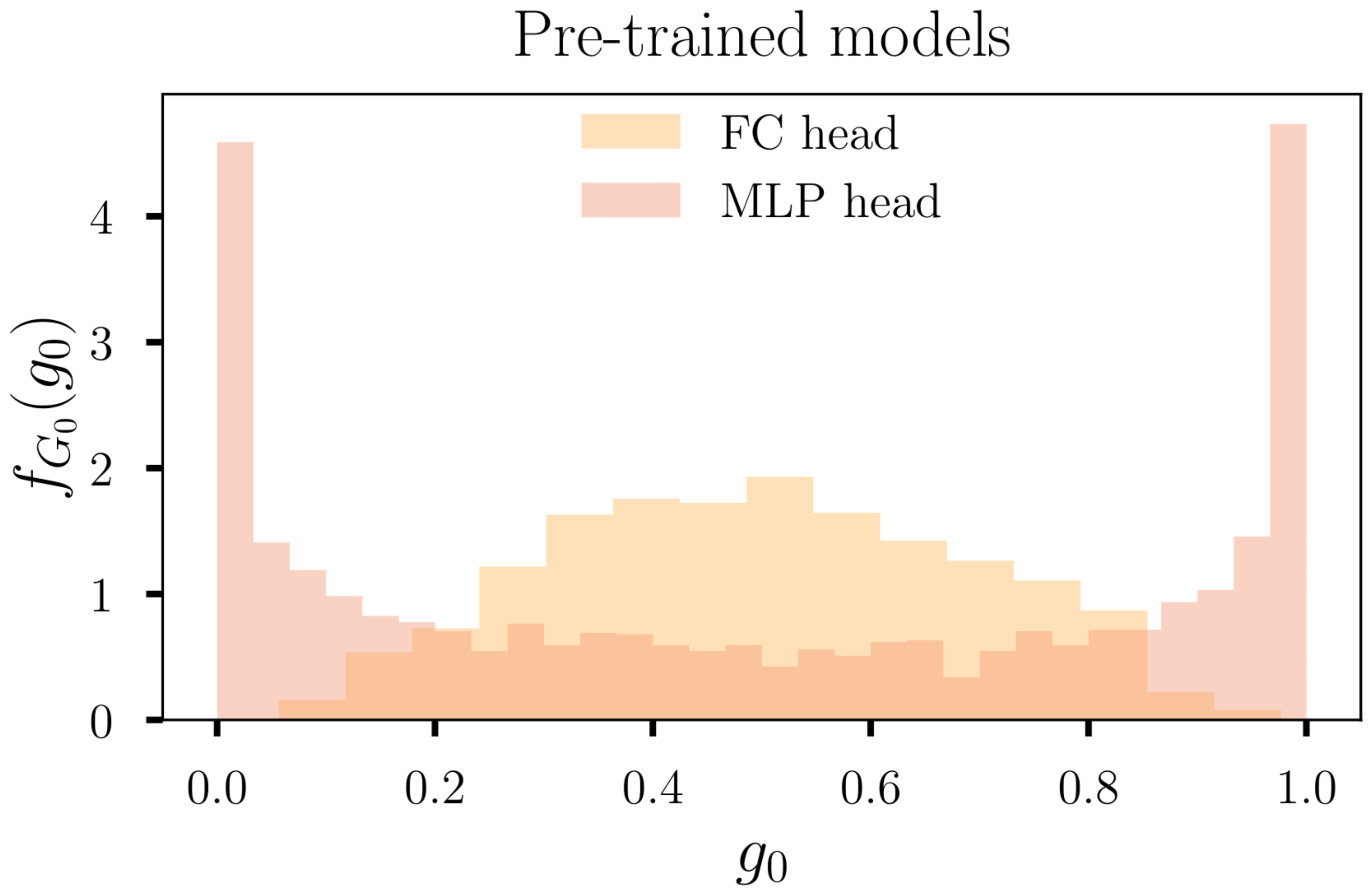
Empirical On Real Data: Even Stronger IGB



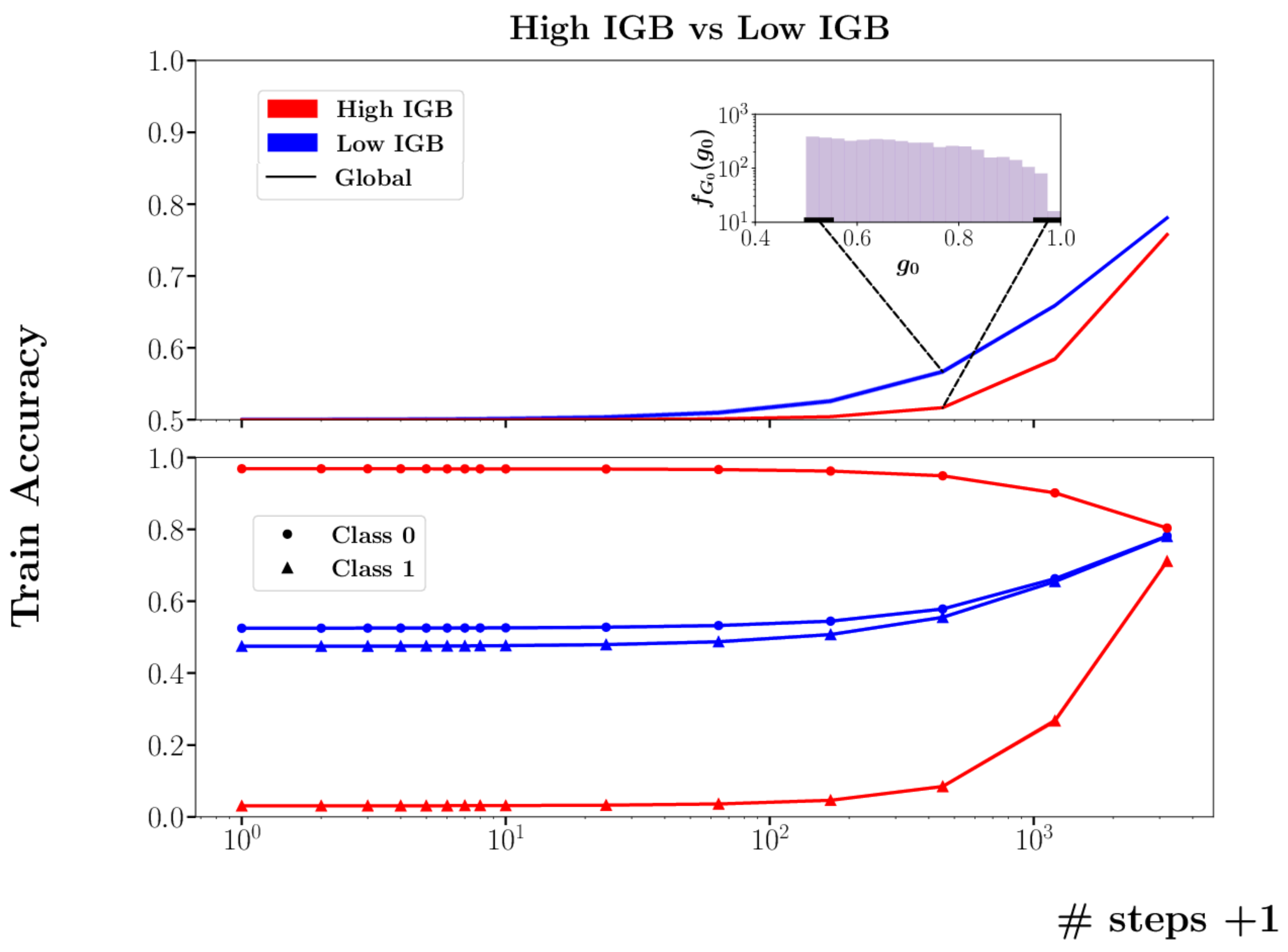
IGB Appears Broad Range Of Architectures...



...Including
Pre-Trained Models

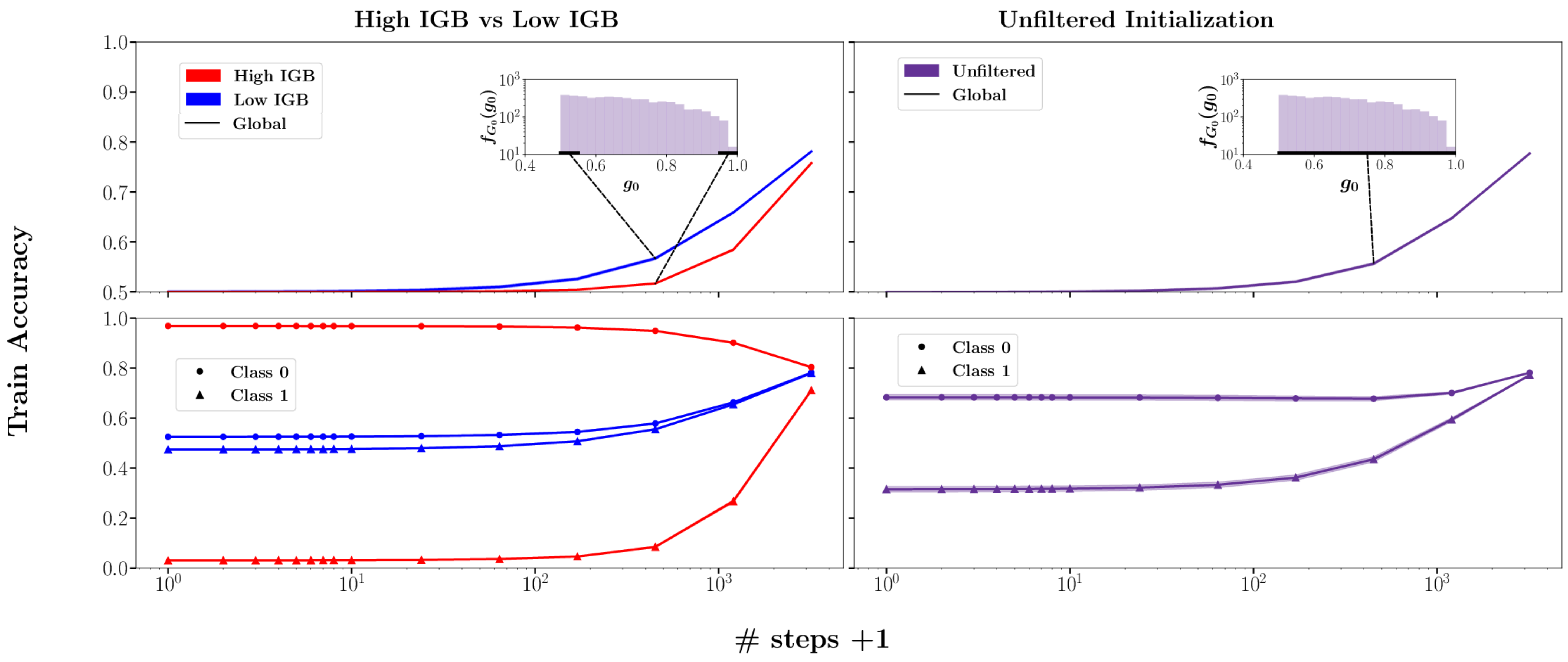


Impact On Dynamics: Preliminary



Grouping initializations by predictive behavior (neutral vs. prejudiced) reveals distinct training dynamics (left).

Impact On Dynamics: Preliminary

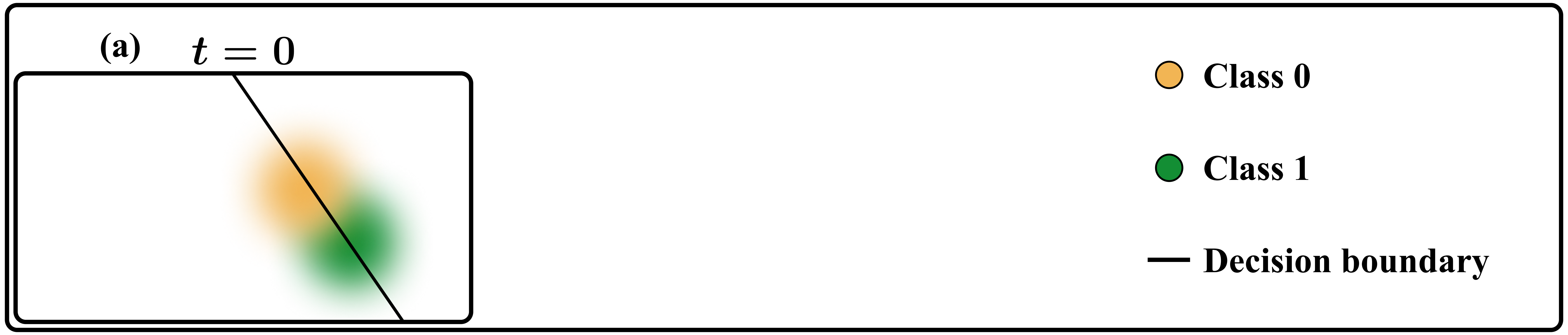


Grouping initializations by predictive behavior (neutral vs. prejudiced) reveals distinct training dynamics (left).

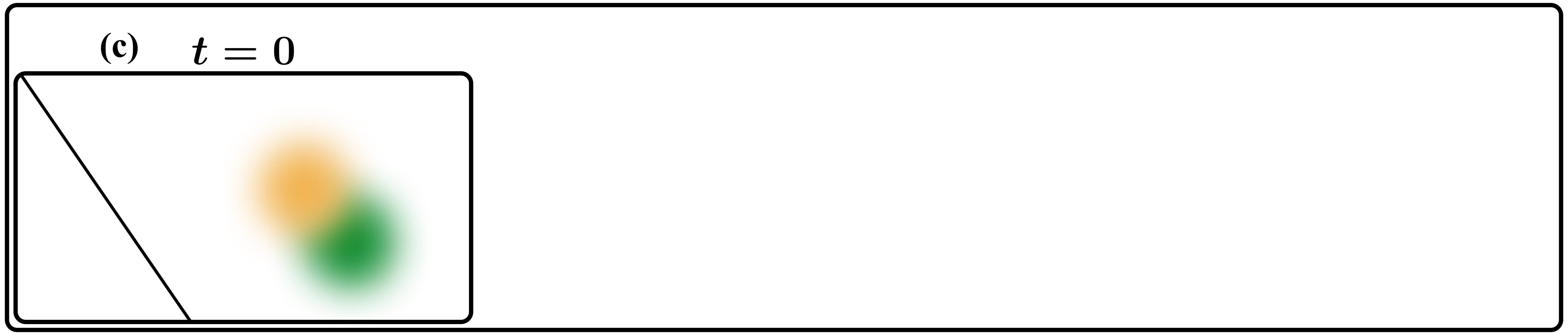
The average behavior across random initializations reflects a mixture of both regimes (right).

Insight Behind IGB

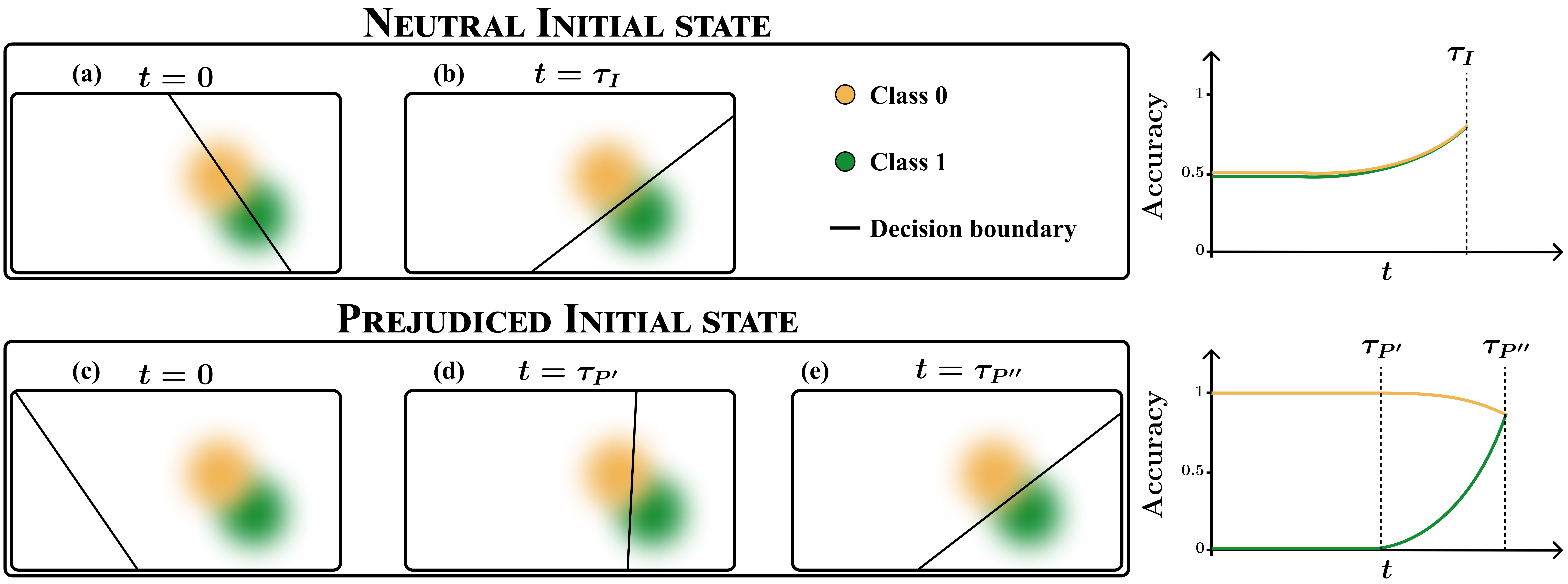
Neutral Initial state



Prejudiced Initial state

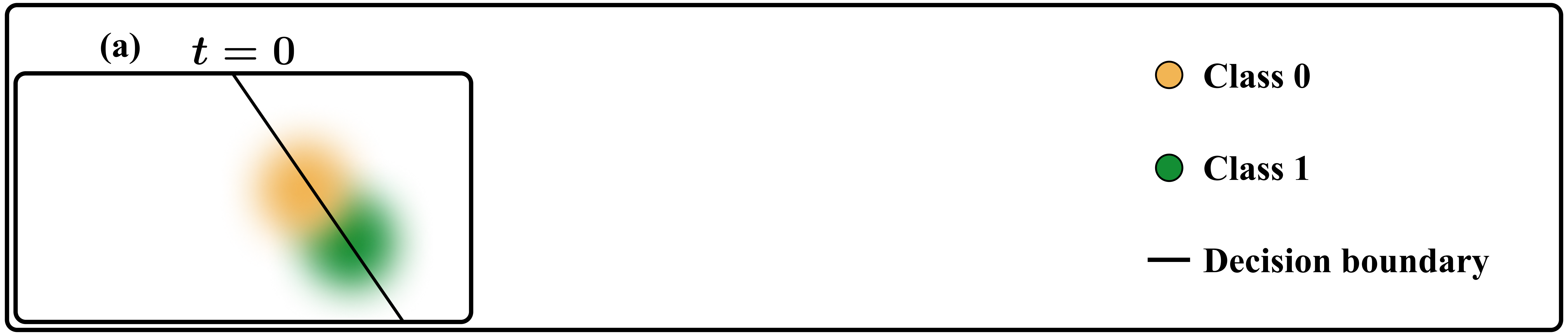


Insight Behind IGB

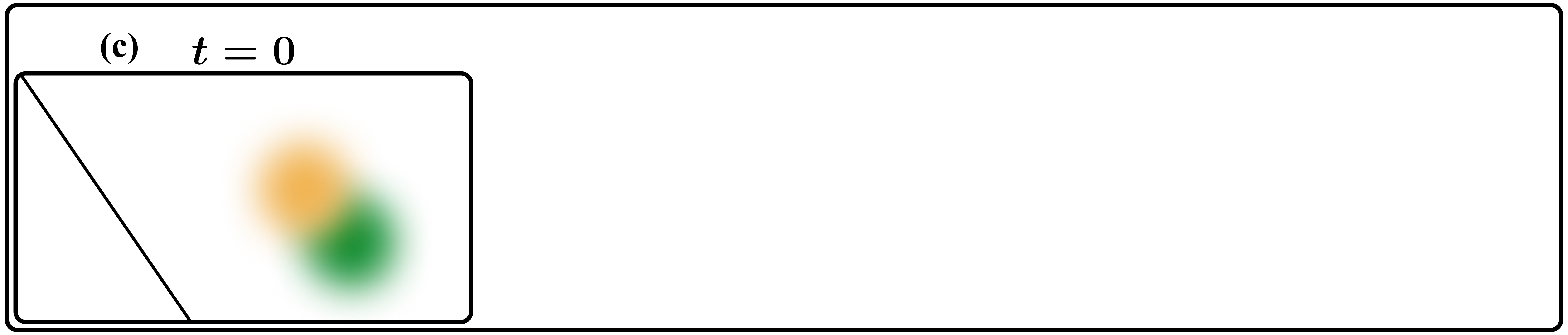


Insight Behind IGB

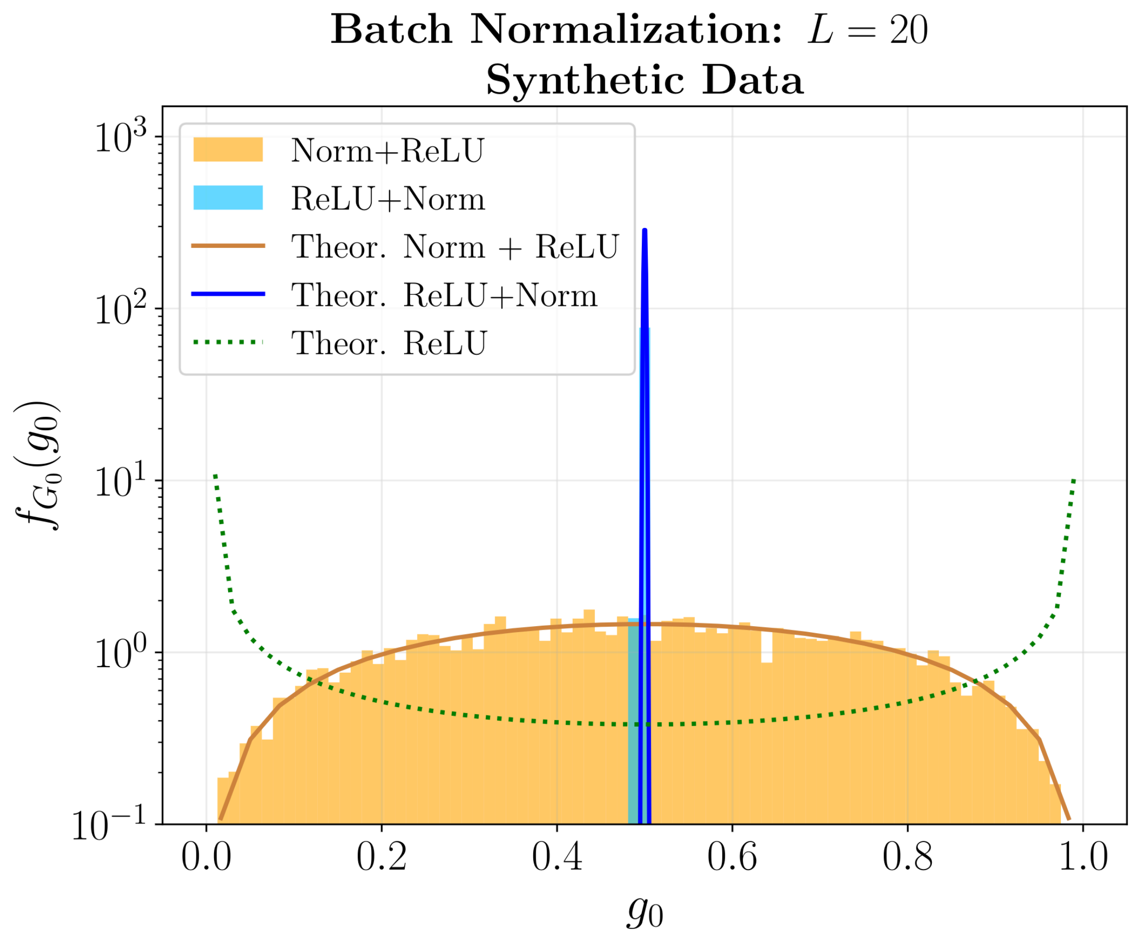
NEUTRAL INITIAL STATE



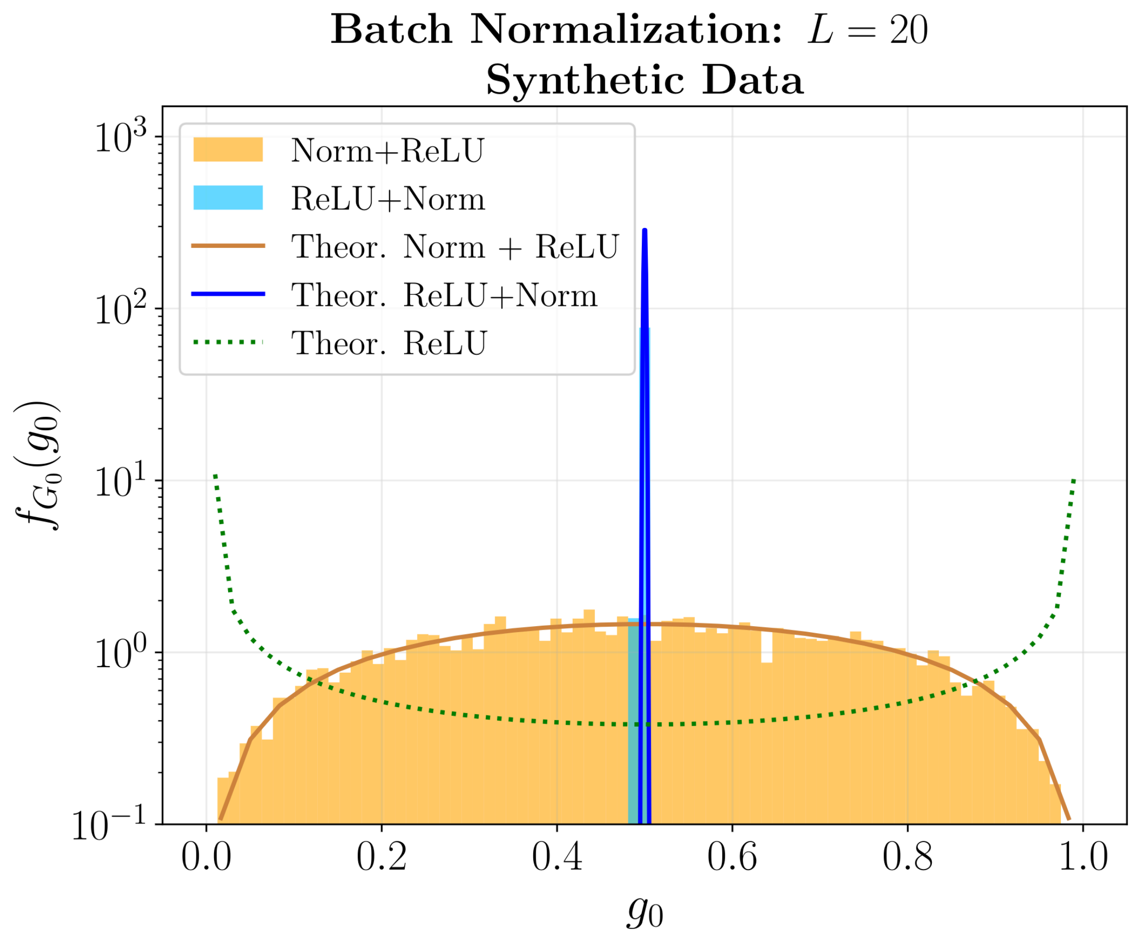
PREJUDICED INITIAL STATE

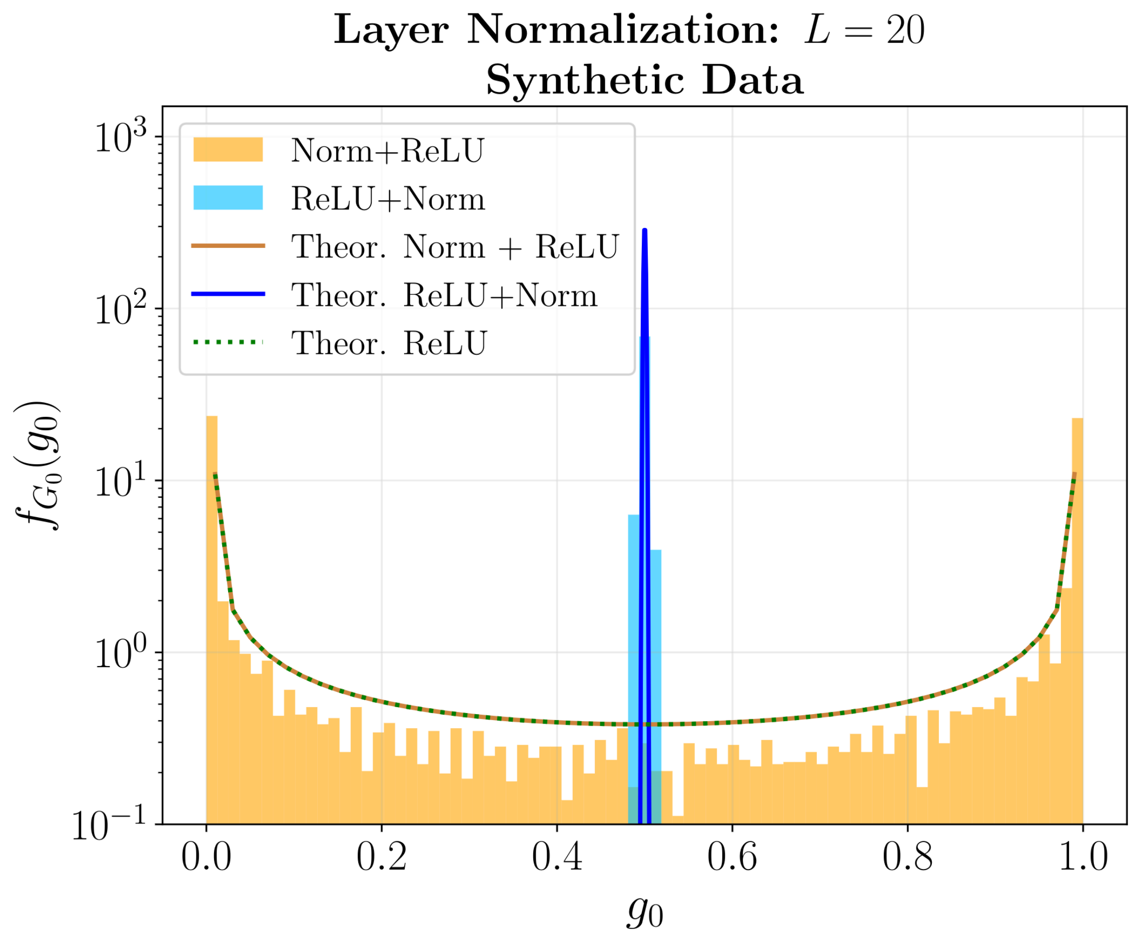


IGB And Normalization

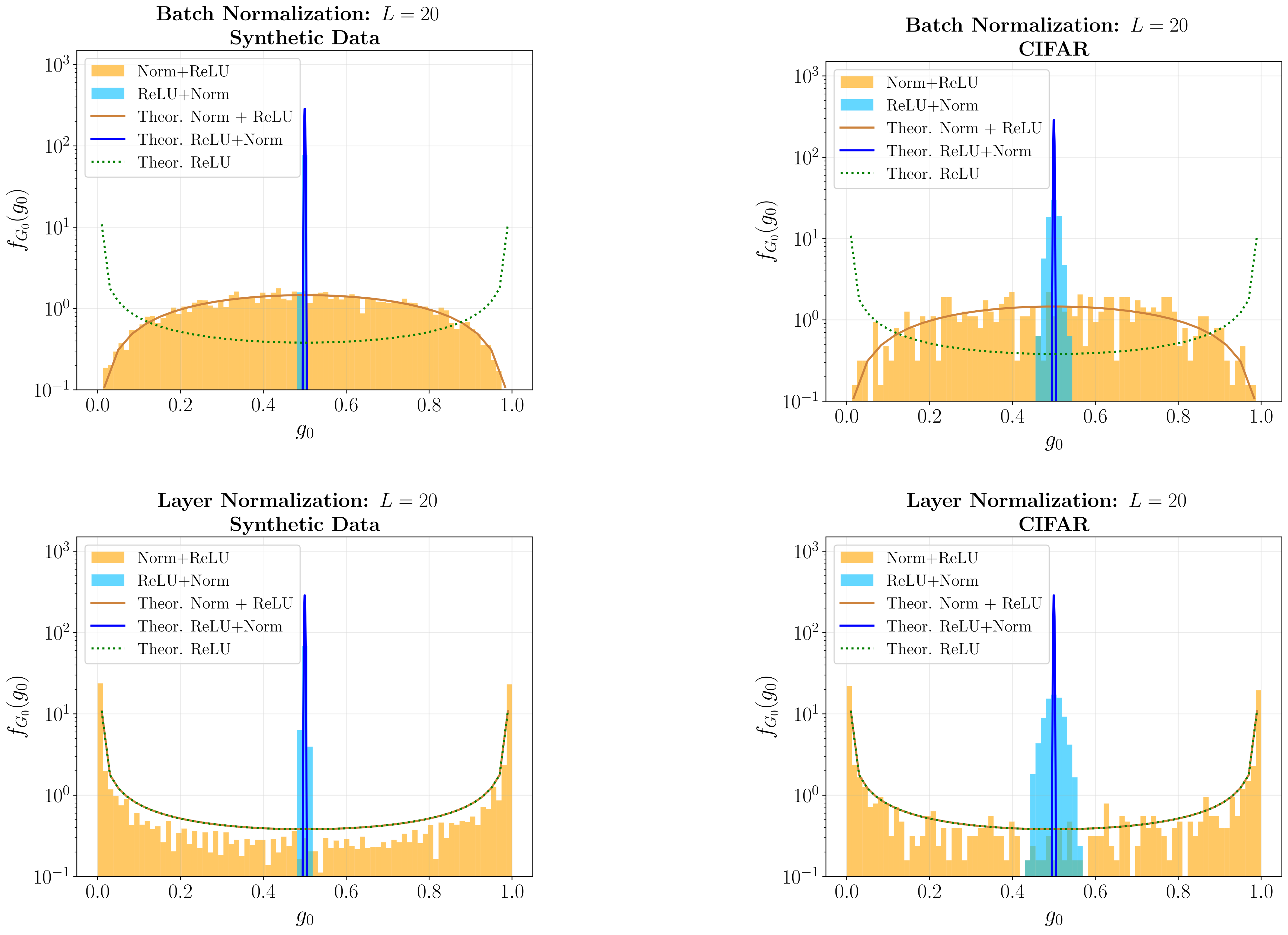


IGB And Normalization



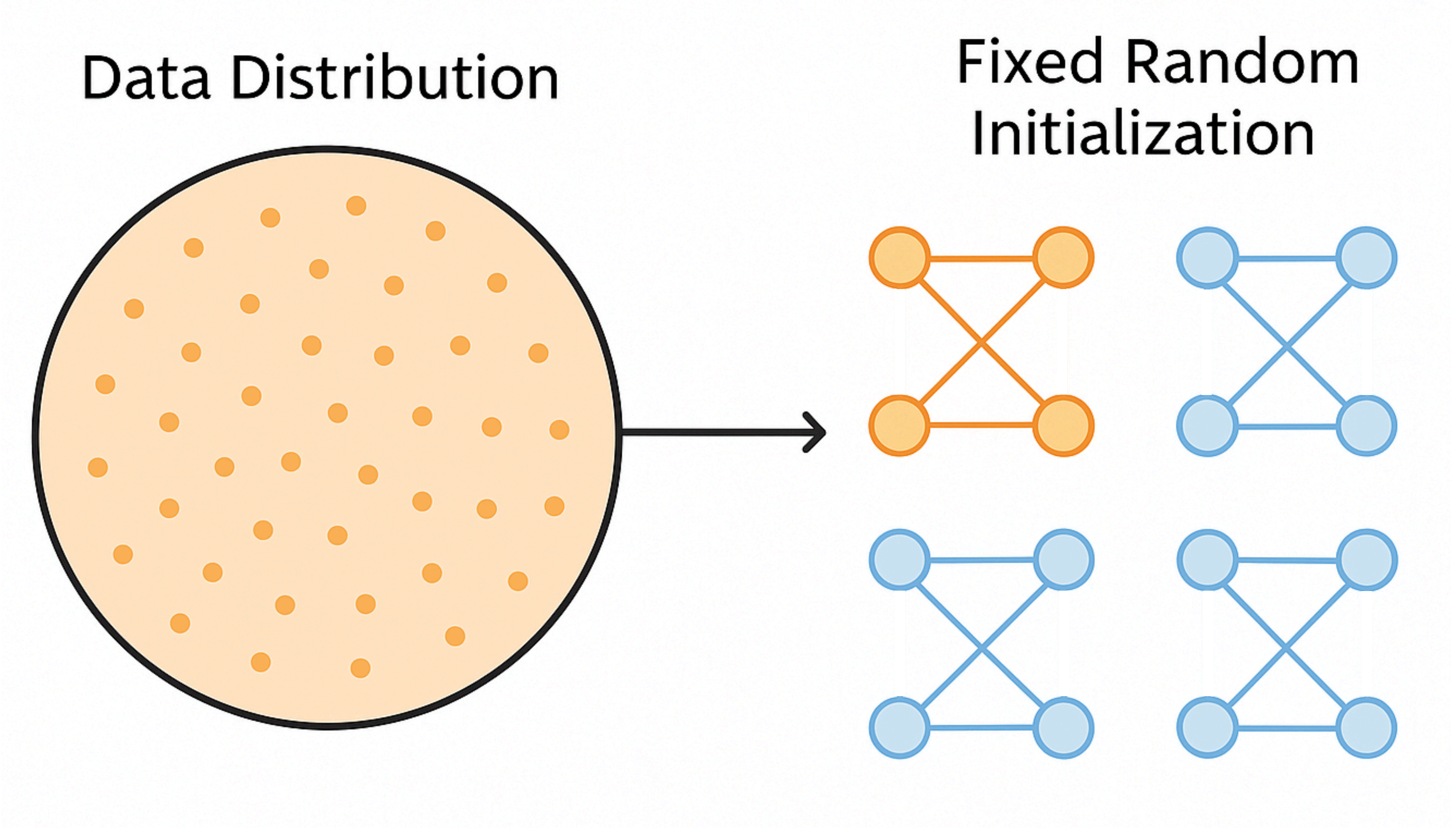


IGB And Normalization



Mean Field (MF) Approach

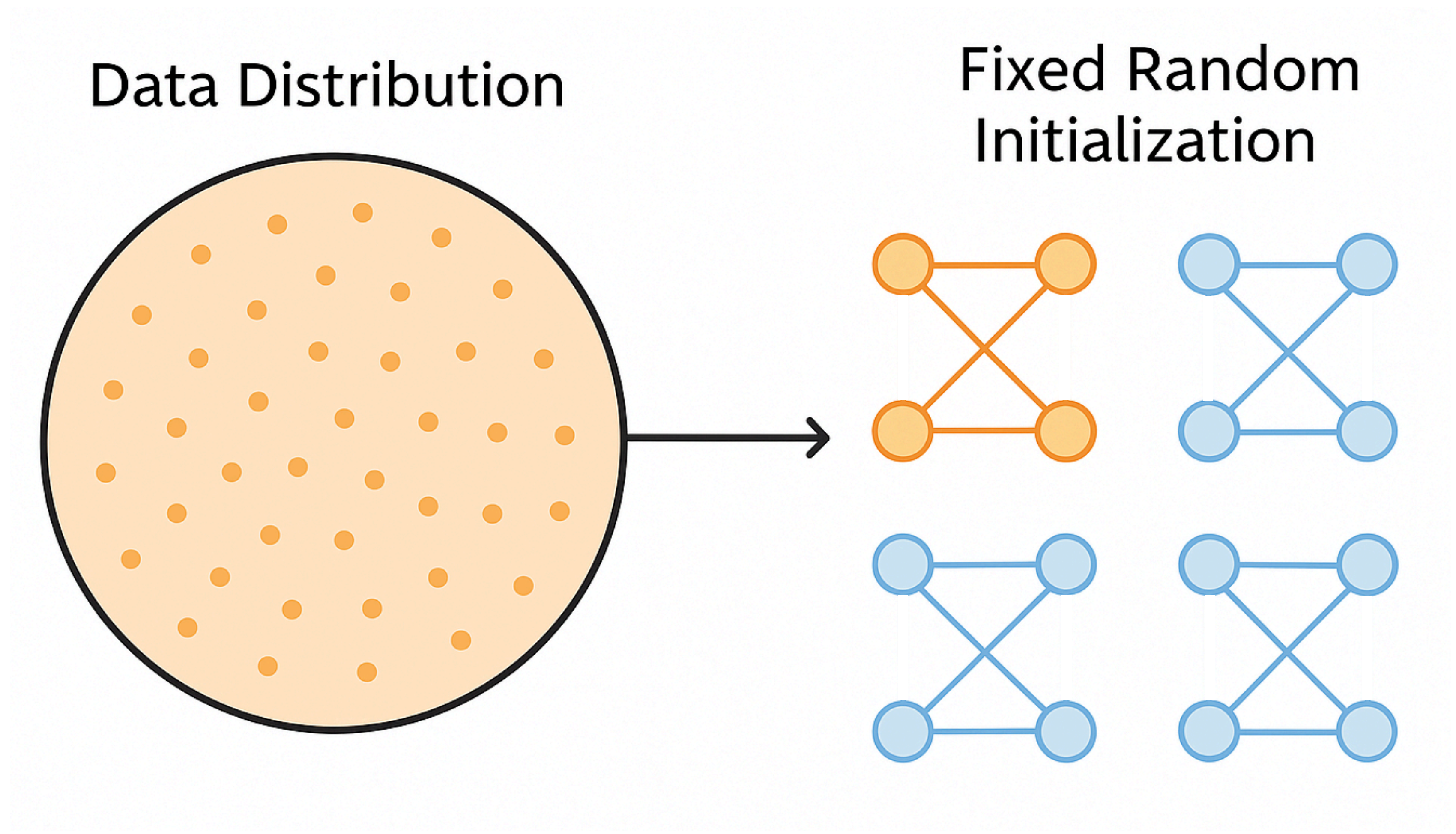
- **IGB:**



- Fix **a random initialization**
- Forward the **entire dataset** through it
- Key quantity G_0 : averaged over inputs, not weights

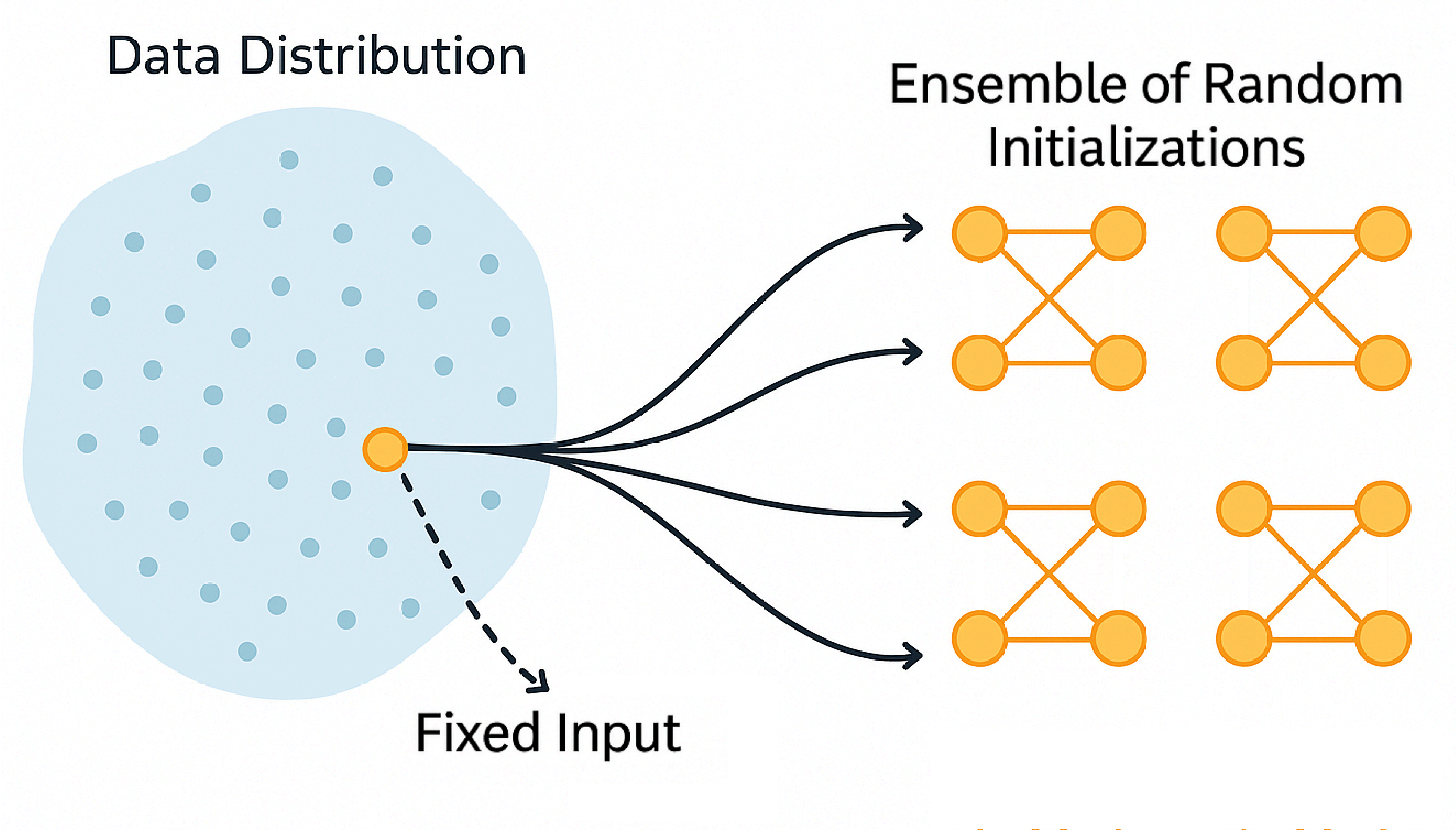
Mean Field (MF) Approach

- **IGB:**



- Fix **a random initialization**
- Forward the **entire dataset** through it
- Key quantity G_0 : averaged over inputs, not weights

- **MF:**



- Fix **a pair of inputs**
- Analyze how their correlation (MF key quantity) evolves across layers
- Correlation is computed over the **ensemble of random initializations**

Phase Diagrams

Propagation of sample “a“ through an MLP

$$Y_i^{(l)}(a) = \sum_{j=1}^{N_l} W_{ij}^{(l)} \phi\left(Y_i^{(l-1)}(a)\right) + B_i^{(l)}$$

Phase Diagrams

Propagation of sample "a" through an MLP

$$Y_i^{(l)}(a) = \sum_{j=1}^{N_l} W_{ij}^{(l)} \phi\left(Y_i^{(l-1)}(a)\right) + B_i^{(l)}$$

DNN parameters initialization:

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{N_l}\right)$$

$$B_i^{(l)} \sim \mathcal{N}\left(0, \sigma_b^2\right)$$

Phase Diagrams

Propagation of sample “a“ through an MLP

$$Y_i^{(l)}(a) = \sum_{j=1}^{N_l} W_{ij}^{(l)} \phi\left(Y_i^{(l-1)}(a)\right) + B_i^{(l)}$$

DNN parameters initialization:

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{N_l}\right)$$

$$B_i^{(l)} \sim \mathcal{N}\left(0, \sigma_b^2\right)$$

Correlation: $$c_{ab}^{(l)} = \frac{\mathbb{E}_{\mathcal{W}}\left(Y_i^{(l)}(a) Y_i^{(l)}(b)\right)}{\sqrt{\mathbb{E}_{\mathcal{W}}\left(\left(Y_i^{(l)}(a)\right)^2\right) \mathbb{E}_{\mathcal{W}}\left(\left(Y_i^{(l)}(b)\right)^2\right)}}$$

Phase Diagrams

Propagation of sample “a“ through an MLP

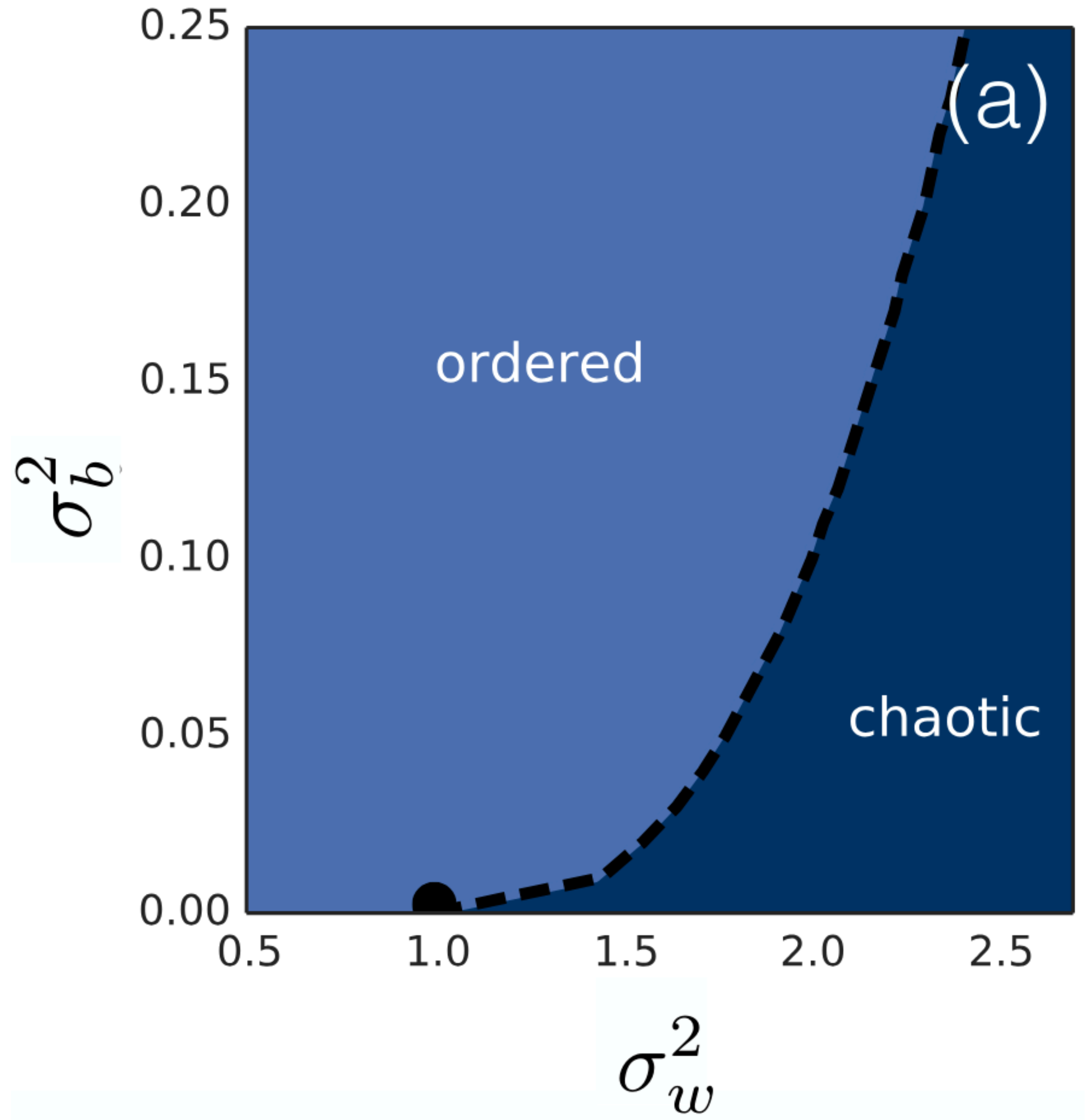
$$Y_i^{(l)}(a) = \sum_{j=1}^{N_l} W_{ij}^{(l)} \phi\left(Y_i^{(l-1)}(a)\right) + B_i^{(l)}$$

DNN parameters initialization:

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{N_l}\right)$$

$$B_i^{(l)} \sim \mathcal{N}\left(0, \sigma_b^2\right)$$

Correlation: $$c_{ab}^{(l)} = \frac{\mathbb{E}_{\mathcal{W}}\left(Y_i^{(l)}(a) Y_i^{(l)}(b)\right)}{\sqrt{\mathbb{E}_{\mathcal{W}}\left(\left(Y_i^{(l)}(a)\right)^2\right) \mathbb{E}_{\mathcal{W}}\left(\left(Y_i^{(l)}(b)\right)^2\right)}}$$



Control parameters: (σ_w^2, σ_b^2)

Order parameter: $\lim_{l \to \infty} c_{ab}^{(l)} = c$

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. arXiv preprint arXiv:1611.01232, 2016.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In International conference on machine learning, pages 2672–2680. PMLR, 2019.

Phase Diagrams

Propagation of sample "a" through an MLP

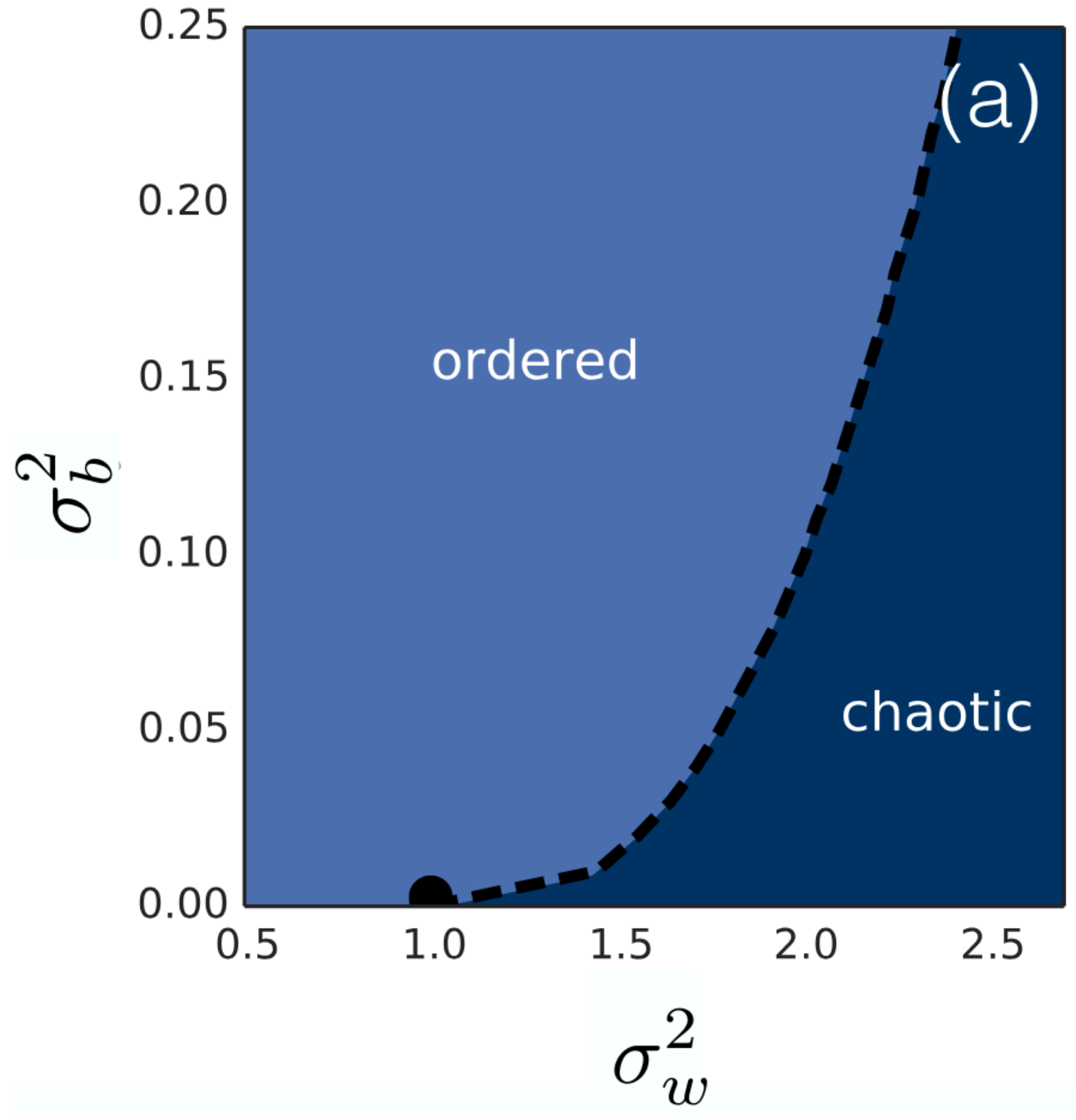
$$Y_i^{(l)}(a) = \sum_{j=1}^{N_l} W_{ij}^{(l)} \phi\left(Y_i^{(l-1)}(a)\right) + B_i^{(l)}$$

DNN parameters initialization:

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{N_l}\right)$$

$$B_i^{(l)} \sim \mathcal{N}\left(0, \sigma_b^2\right)$$

Correlation: $c_{ab}^{(l)} = \dfrac{\mathbb{E}_{\mathcal{W}}\left(Y_i^{(l)}(a) Y_i^{(l)}(b)\right)}{\sqrt{\mathbb{E}_{\mathcal{W}}\left(\left(Y_i^{(l)}(a)\right)^2\right) \mathbb{E}_{\mathcal{W}}\left(\left(Y_i^{(l)}(b)\right)^2\right)}}$



Control parameters: (σ_w^2, σ_b^2)

Order parameter: $\lim_{l \to \infty} c_{ab}^{(l)} = c$

Ordered phase: $c = 1$ is stable

Chaotic phase: $c = 1$ is unstable; converges to $c < 1$

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. arXiv preprint arXiv:1611.01232, 2016.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In International conference on machine learning, pages 2672–2680. PMLR, 2019.

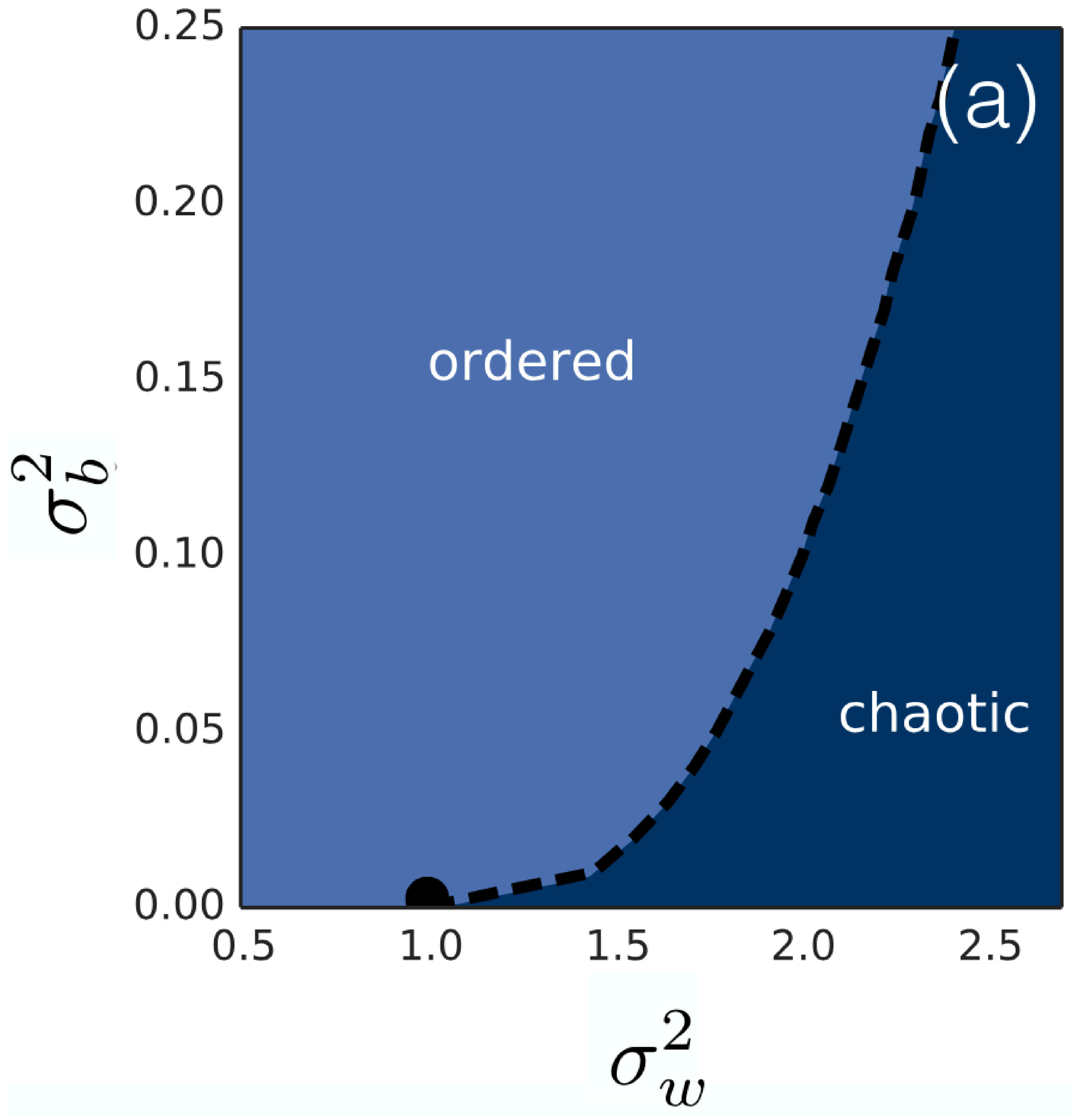
Gradient Behavior Across Phases

The two phases correspond to distinct gradient behaviors:

Ordered phase: Vanishing gradients backward decay leads to **persistence** of the initial state.

Chaotic phase: Exploding gradients backward amplification causes **instability**.

Edge of Chaos: Stable gradients enables **effective training**.



Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. arXiv preprint arXiv:1611.01232, 2016.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In International conference on machine learning, pages 2672–2680. PMLR, 2019.

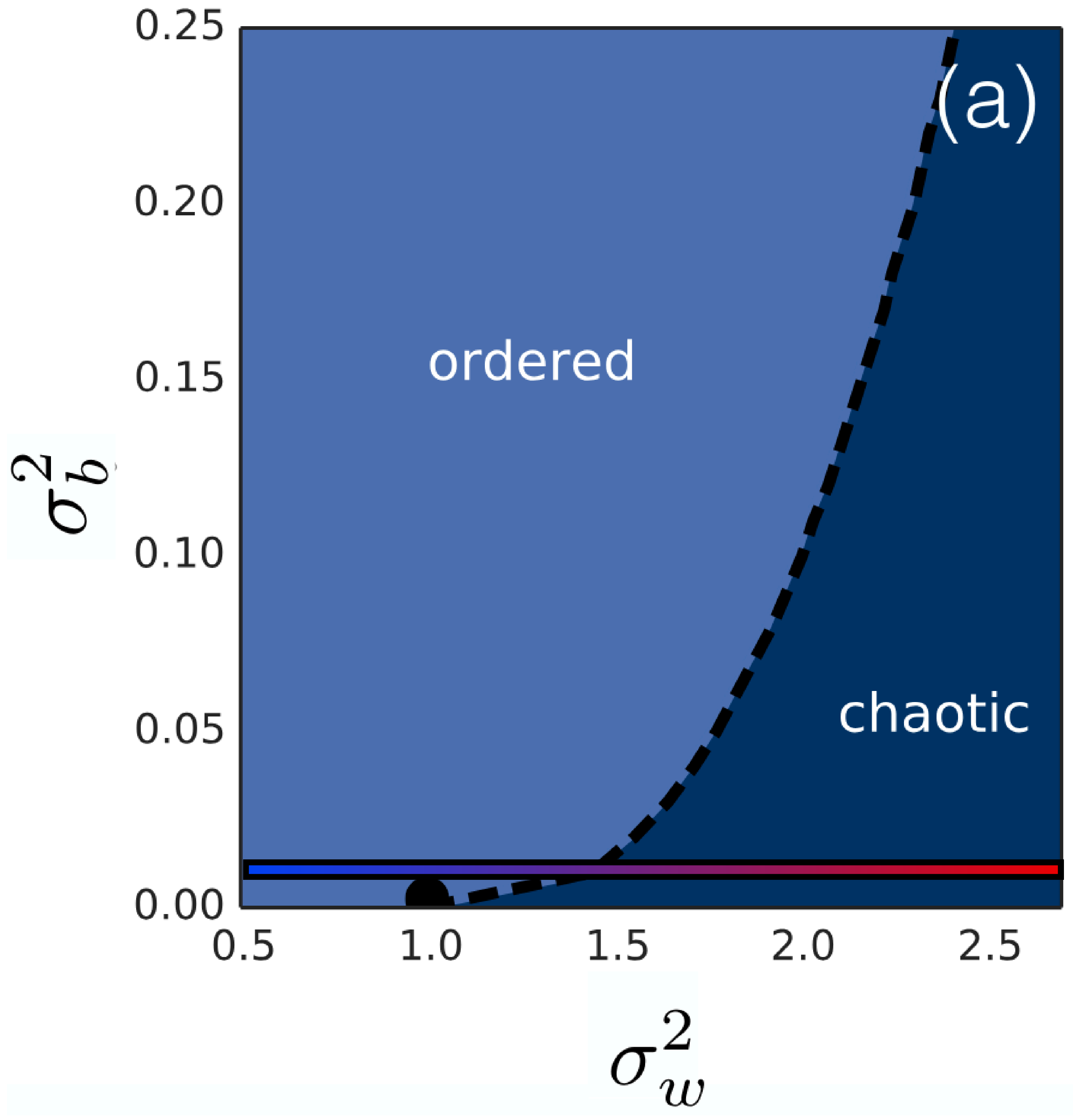
Gradient Behavior Across Phases

The two phases correspond to distinct gradient behaviors:

Ordered phase: Vanishing gradients backward decay leads to **persistence** of the initial state.

Chaotic phase: Exploding gradients backward amplification causes **instability**.

Edge of Chaos: Stable gradients enables **effective training**.



Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. arXiv preprint arXiv:1611.01232, 2016.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In International conference on machine learning, pages 2672–2680. PMLR, 2019.

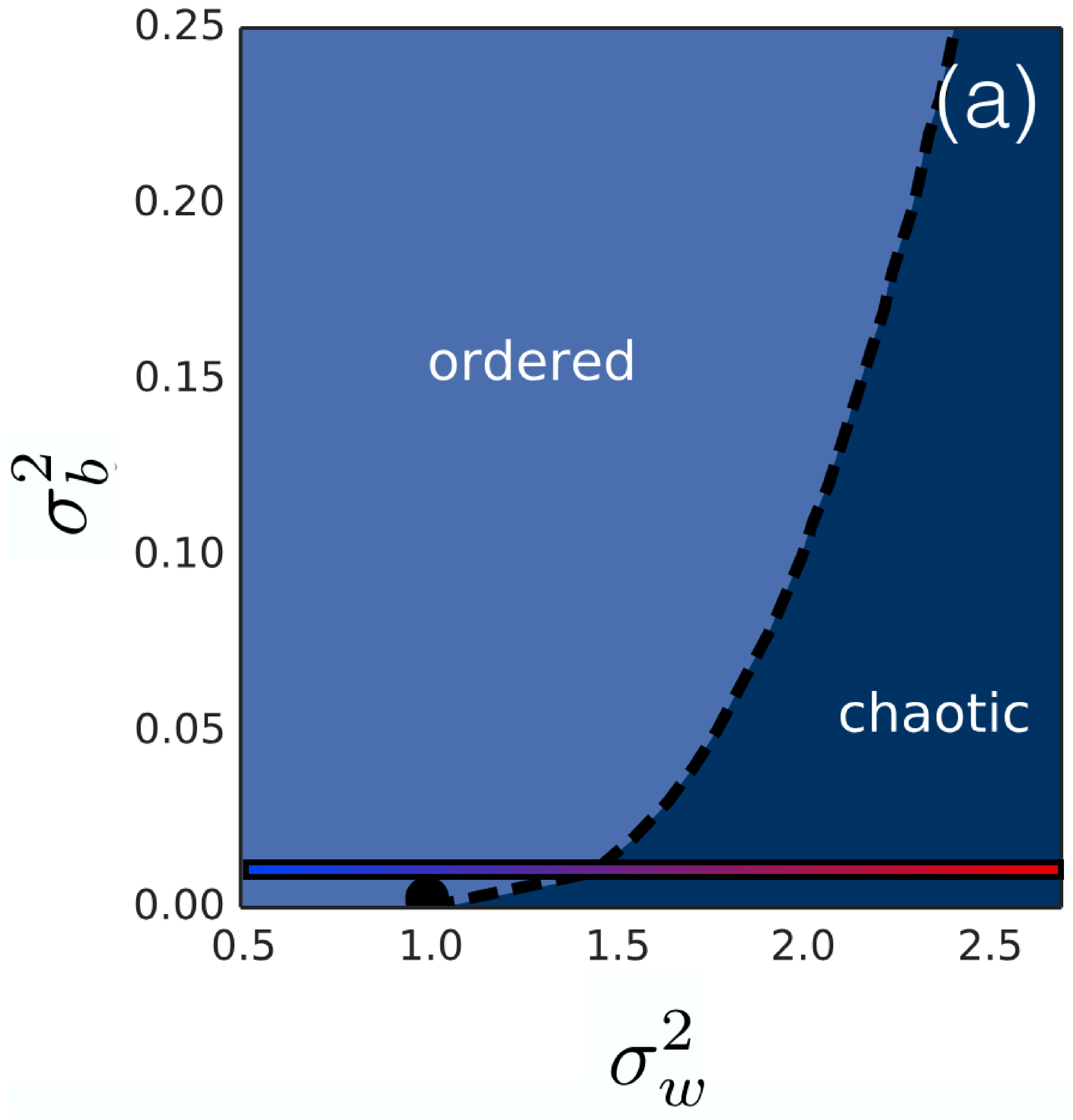
Gradient Behavior Across Phases

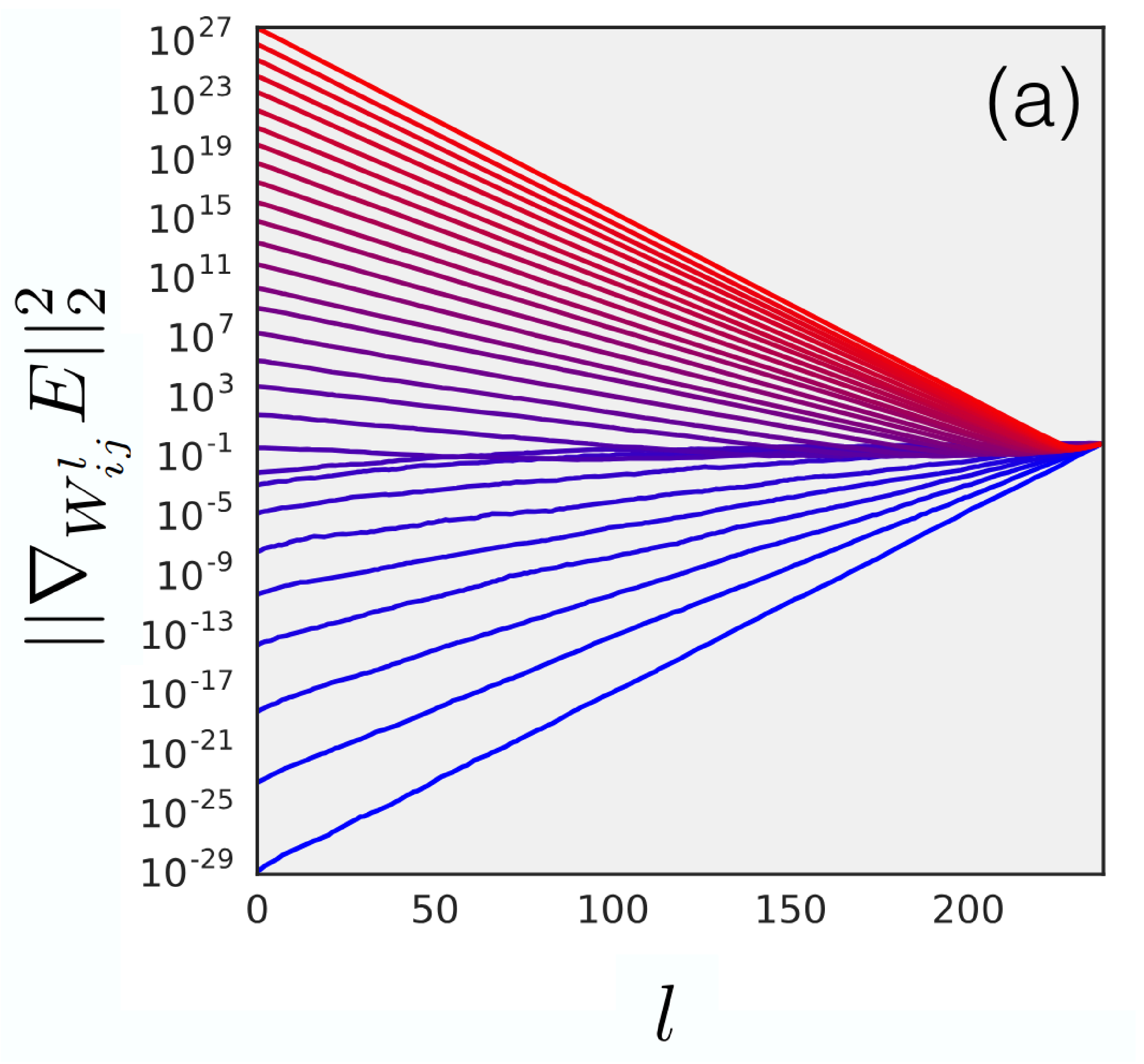
The two phases correspond to distinct gradient behaviors:

Ordered phase: Vanishing gradients backward decay leads to **persistence** of the initial state.

Chaotic phase: Exploding gradients backward amplification causes **instability**.

Edge of Chaos: Stable gradients enables **effective training**.





Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. arXiv preprint arXiv:1611.01232, 2016.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In International conference on machine learning, pages 2672–2680. PMLR, 2019.

Summary Of Key Concepts: IGB Vs. MF

IGB (Initial Guessing Bias)

Captures how **architecture** design shapes **initial prediction**:
→ Neutral vs. Prejudiced behavior

Measured by γ:

- $\gamma \gg 1$: deep prejudice
- $\gamma \ll 1$: neutrality

MF (Mean Field Theory)

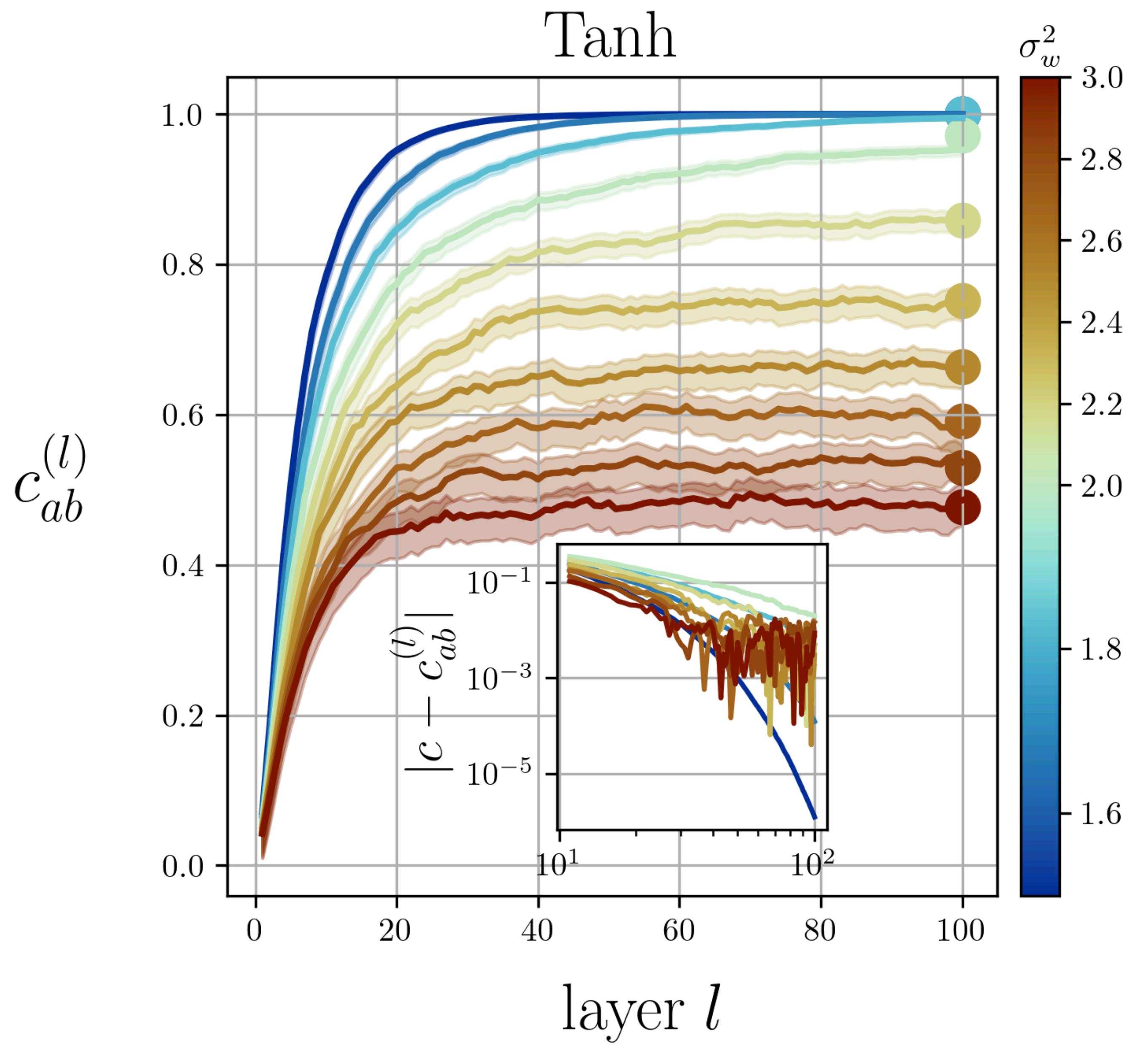
Captures how **hyperparameter** choices shape **trainability**:
→ Ordered vs. Chaotic phases

Described by correlation fixed point c :

- $c = 1$: ordered phase / edge of chaos
- $c < 1$: chaotic phase

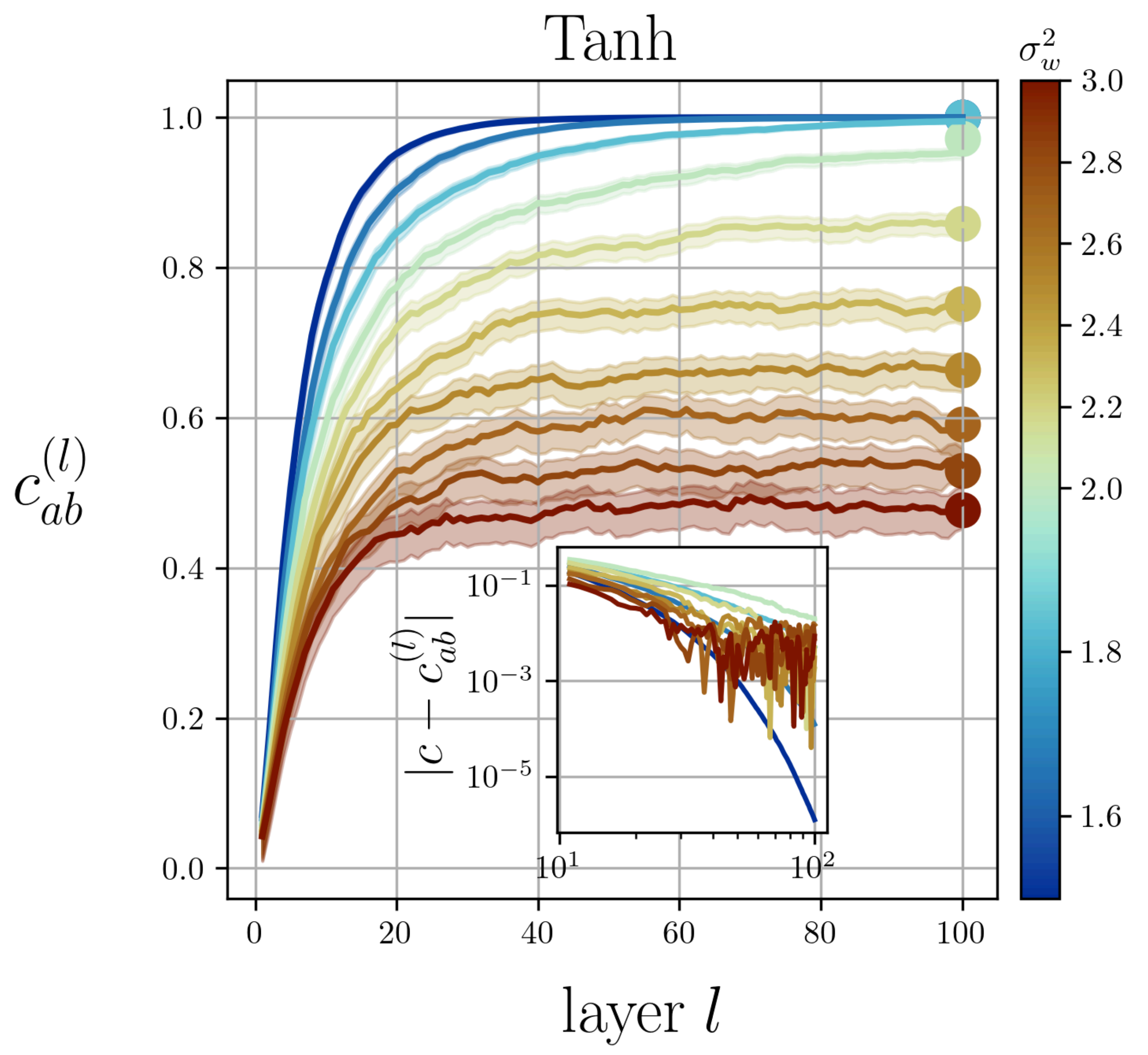
Connecting IGB And MF Frameworks

Link between key quantities: $$c = \frac{\gamma}{1+\gamma}$$



Connecting IGB And MF Frameworks

Link between key quantities: $$c = \frac{\gamma}{1+\gamma}$$



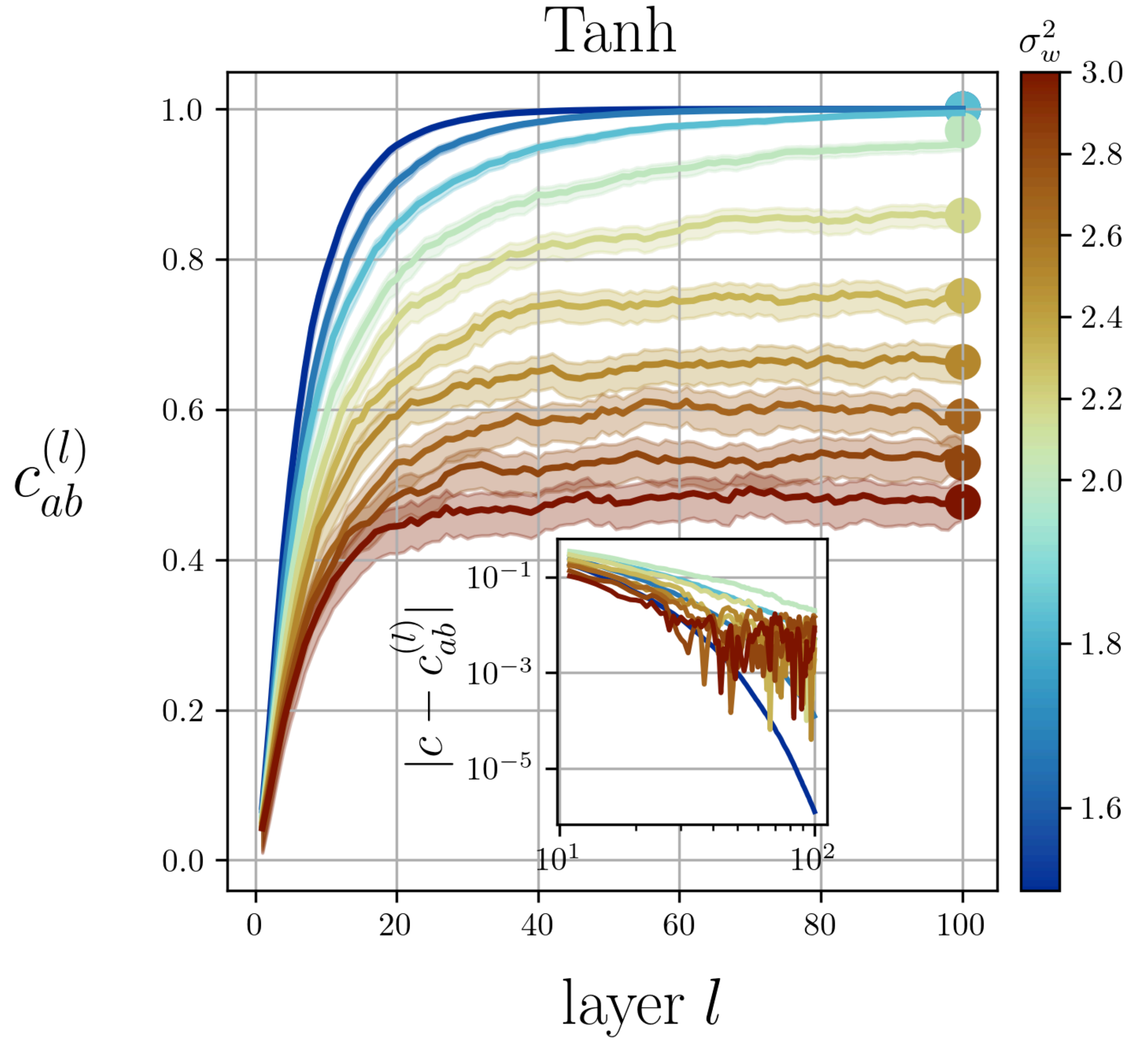
Reveals interplay between design and hyperparameters

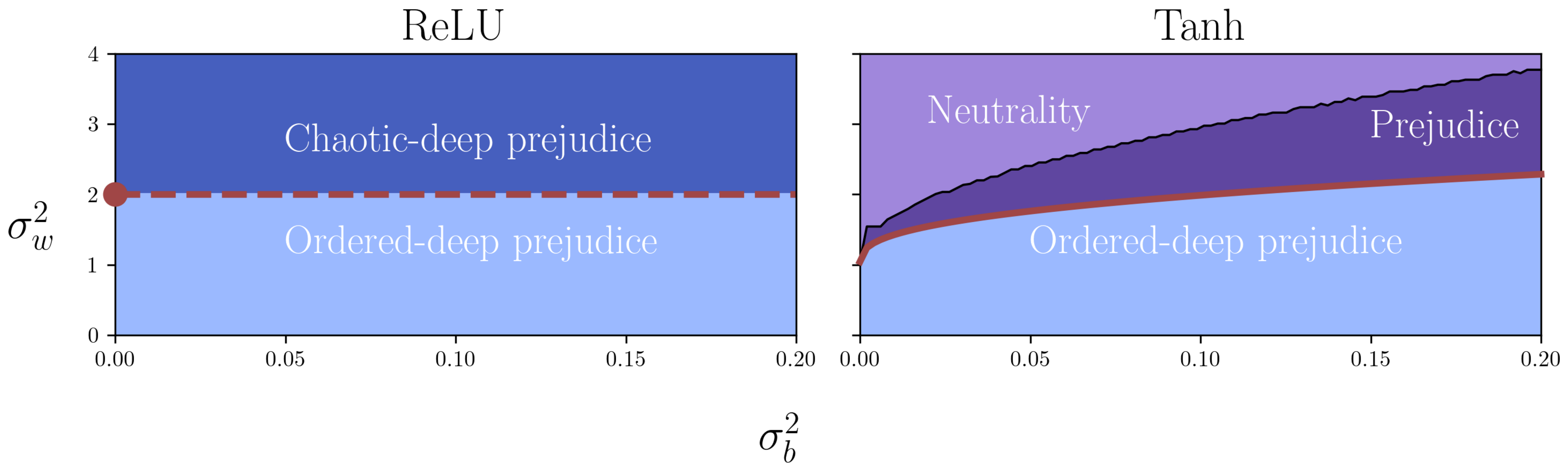
Connecting IGB And MF Frameworks

Link between key quantities: $$c = \frac{\gamma}{1+\gamma}$$

Reveals interplay between design and hyperparameters

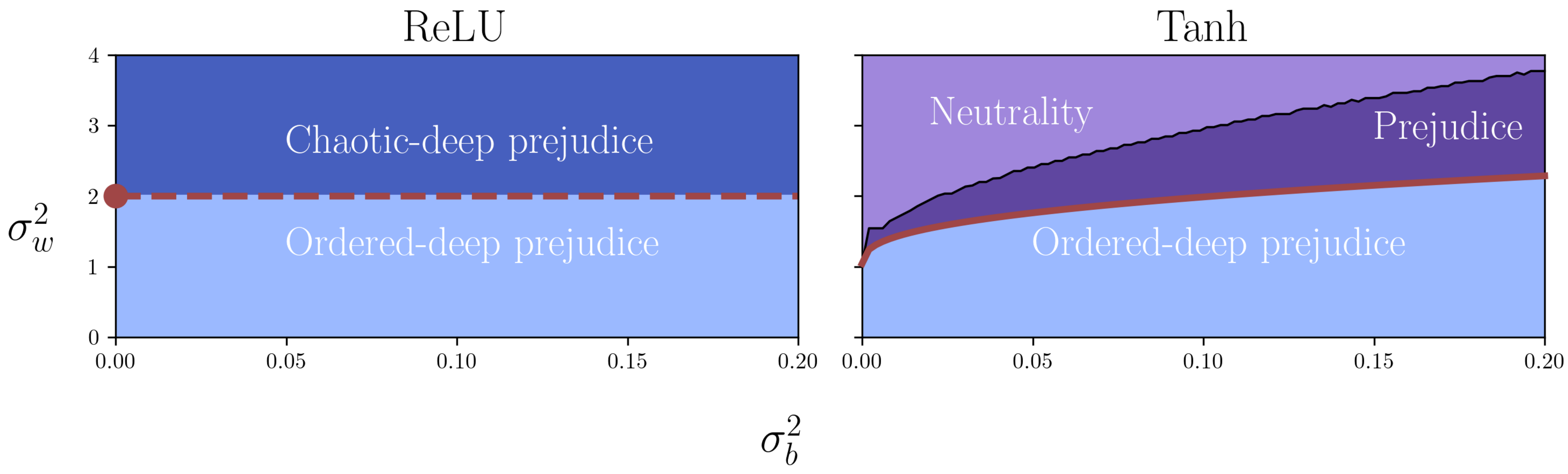
Connects initial bias (IGB) with trainability regimes (MF)





Initial Prejudice And Trainability

$$c = \frac{\gamma}{1+\gamma}$$

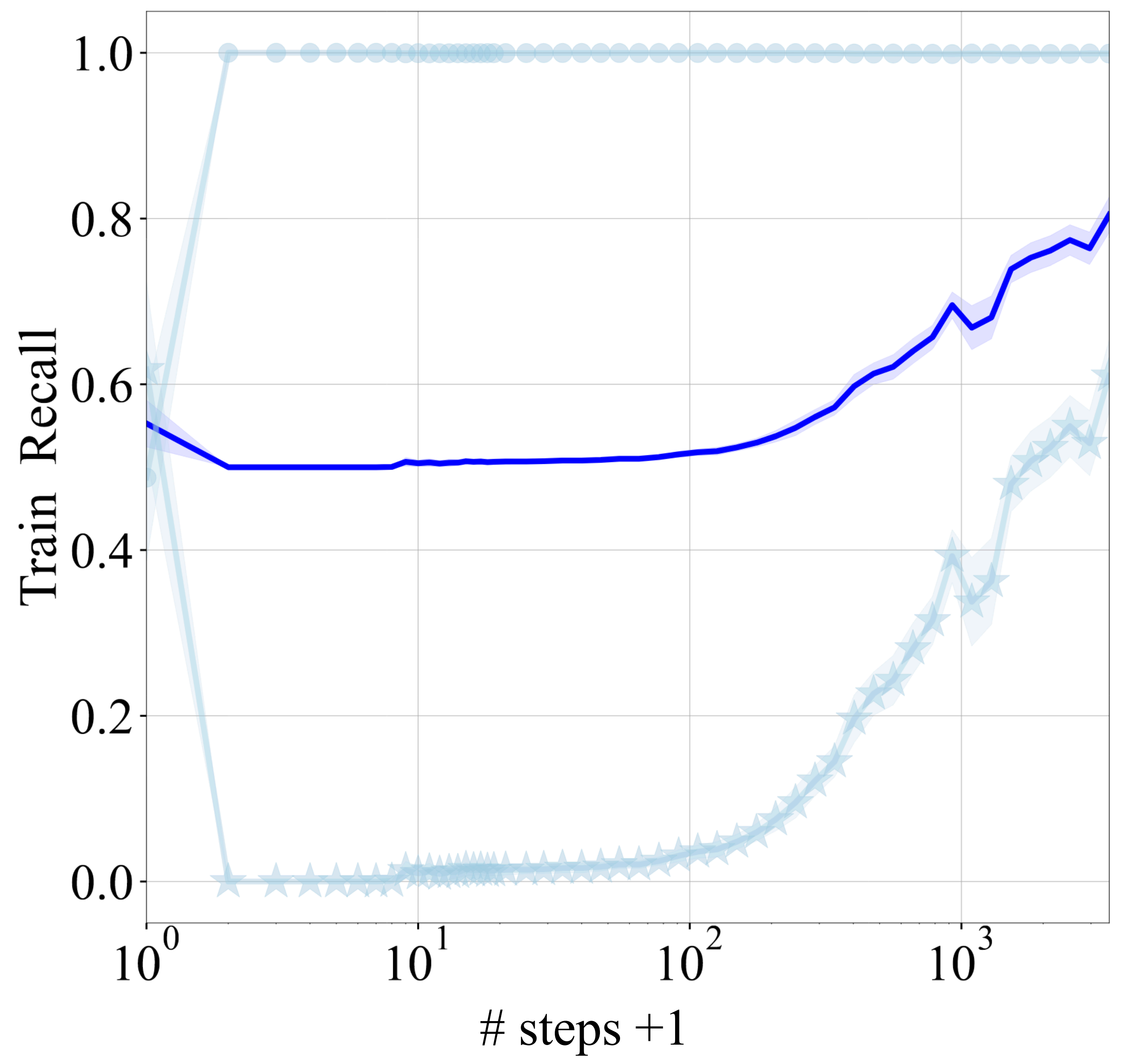


Edge of chaos $(c = 1) \Rightarrow \gamma = \infty$

Trainability peaks not at neutrality, but at deep prejudice.

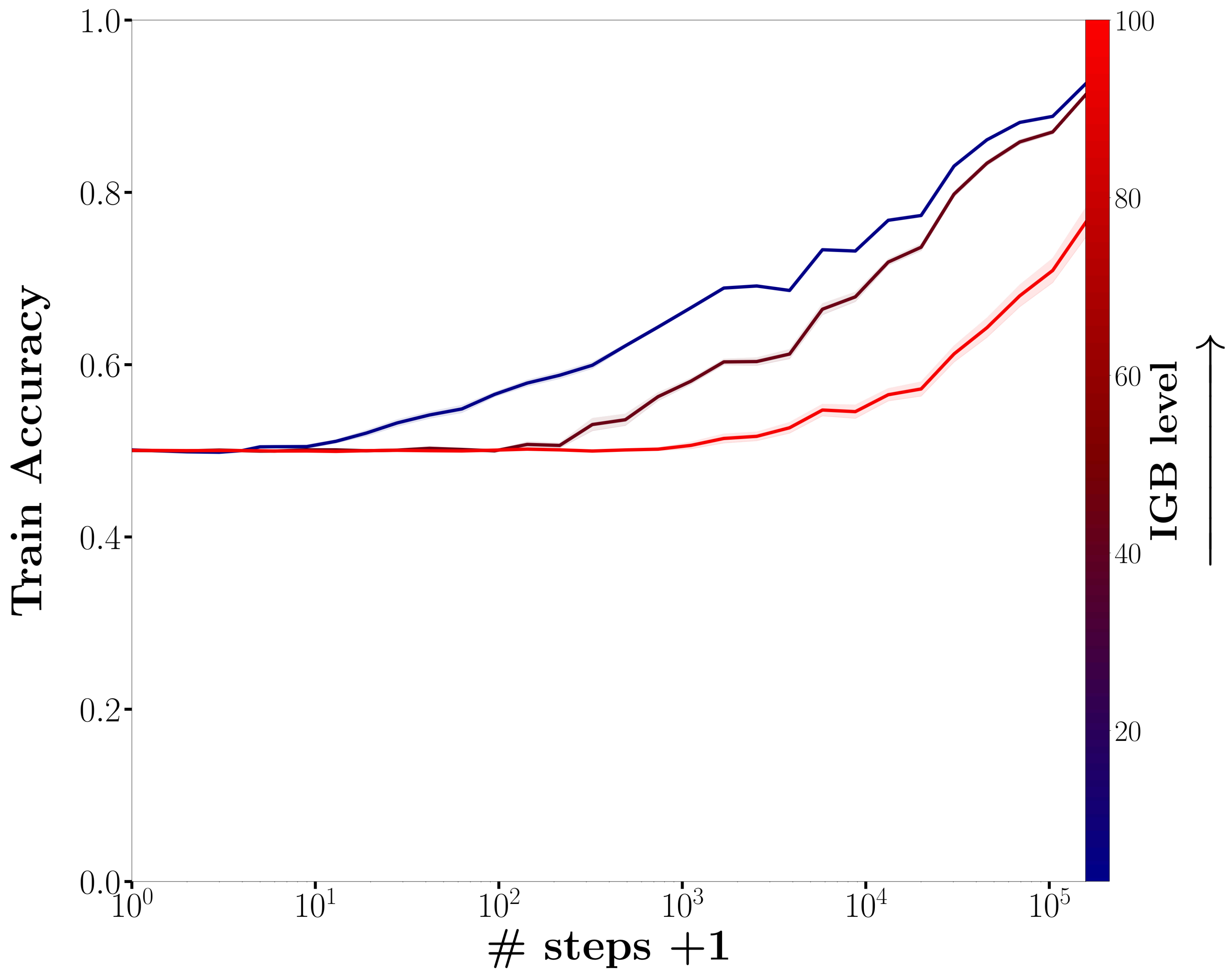
Class Imbalance

arXiv:2207.00391 - **ICML 2023**

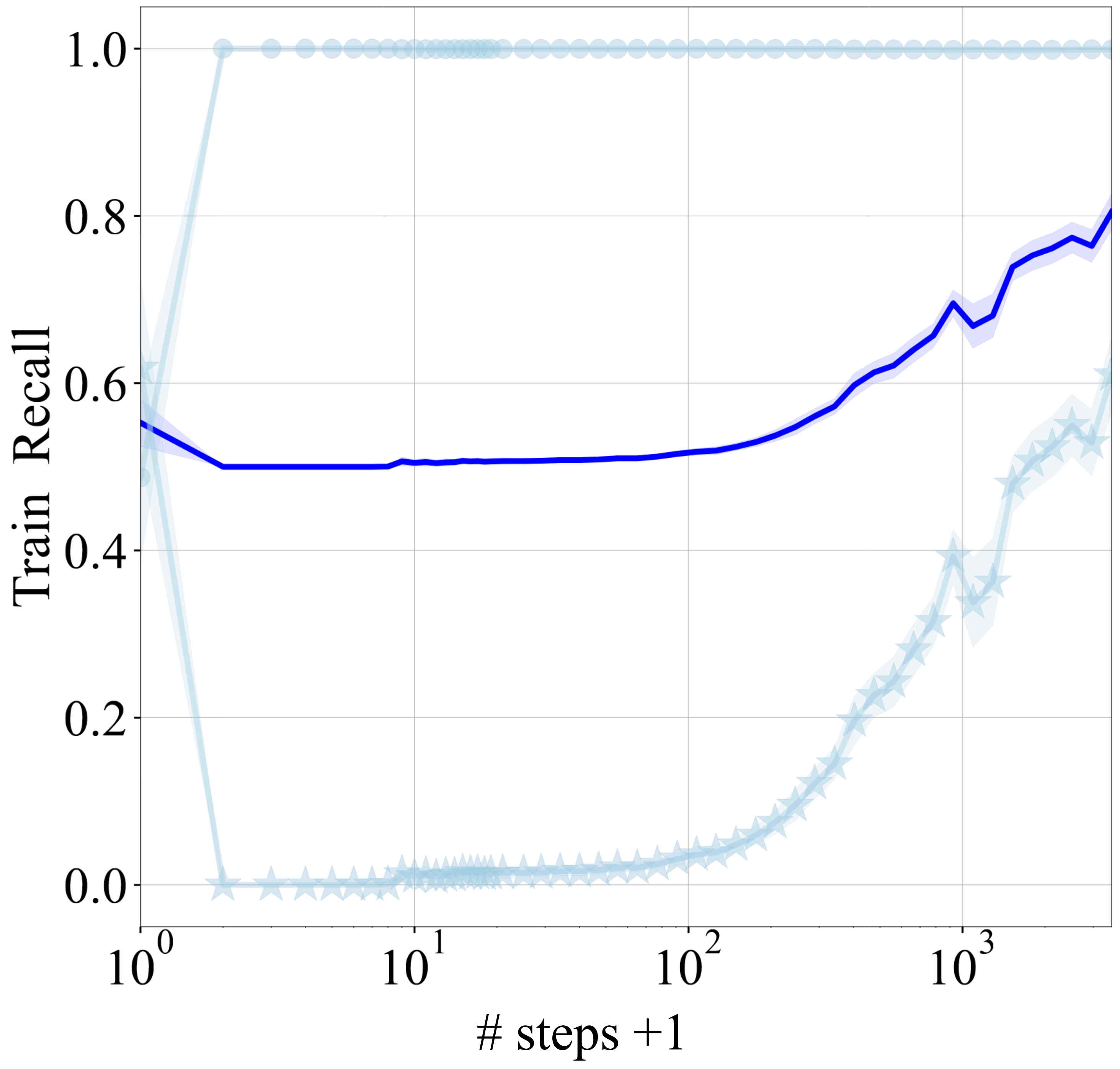


Initial Guessing Bias

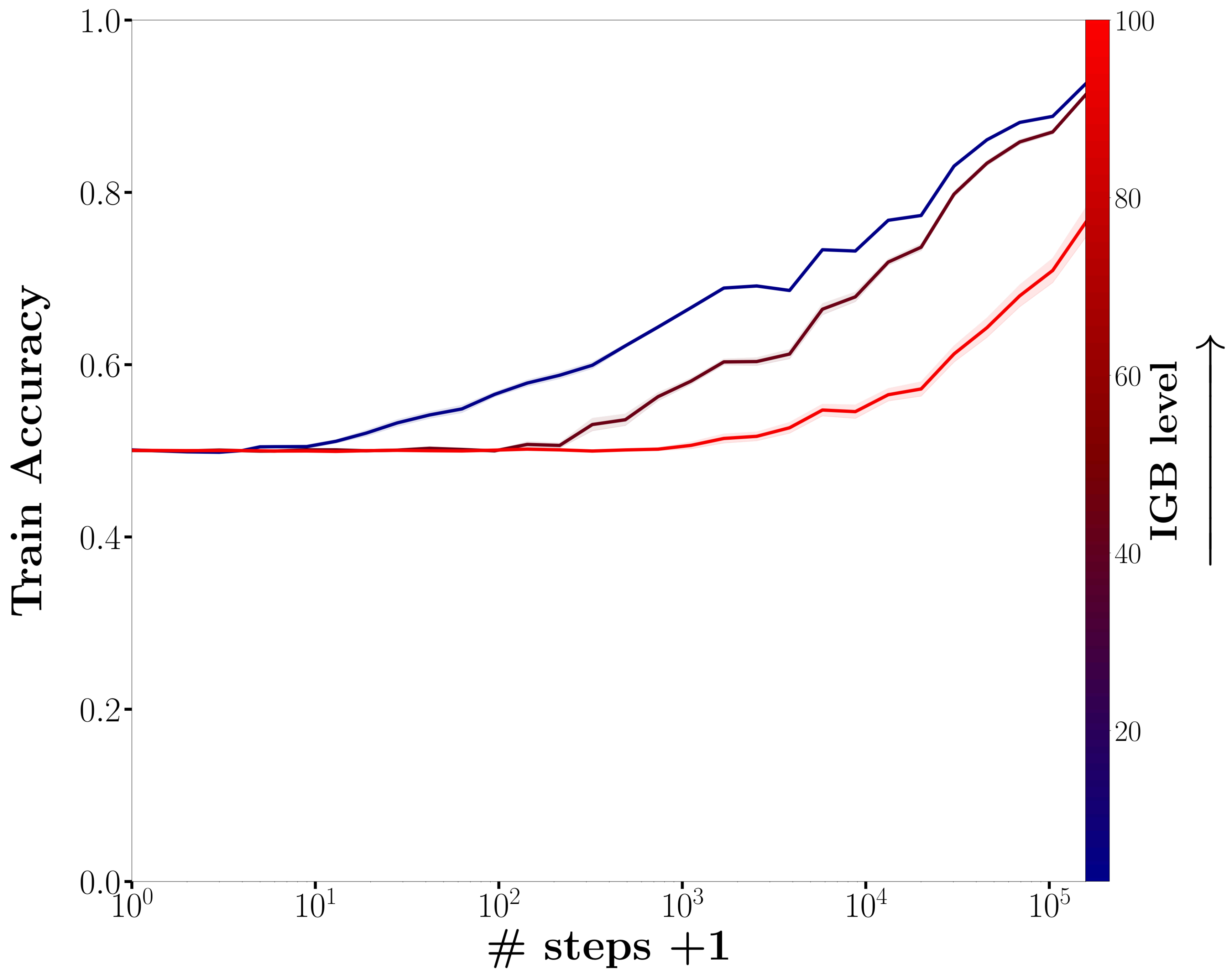
arXiv:2306.00809 - **ICML 2024**



Class Imbalance
arXiv:2207.00391 - **ICML 2023**

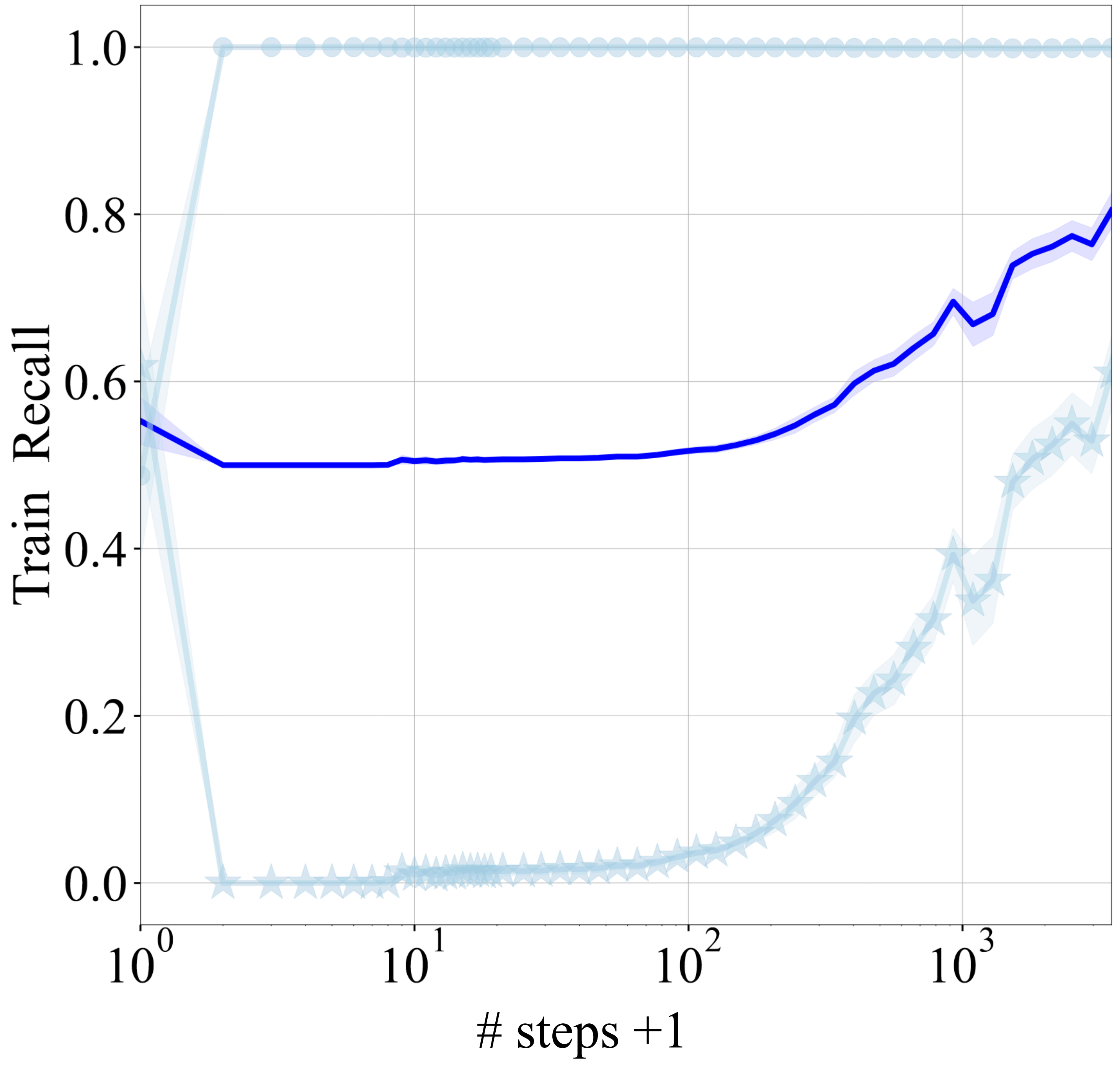


Initial Guessing Bias
arXiv:2306.00809 - **ICML 2024**

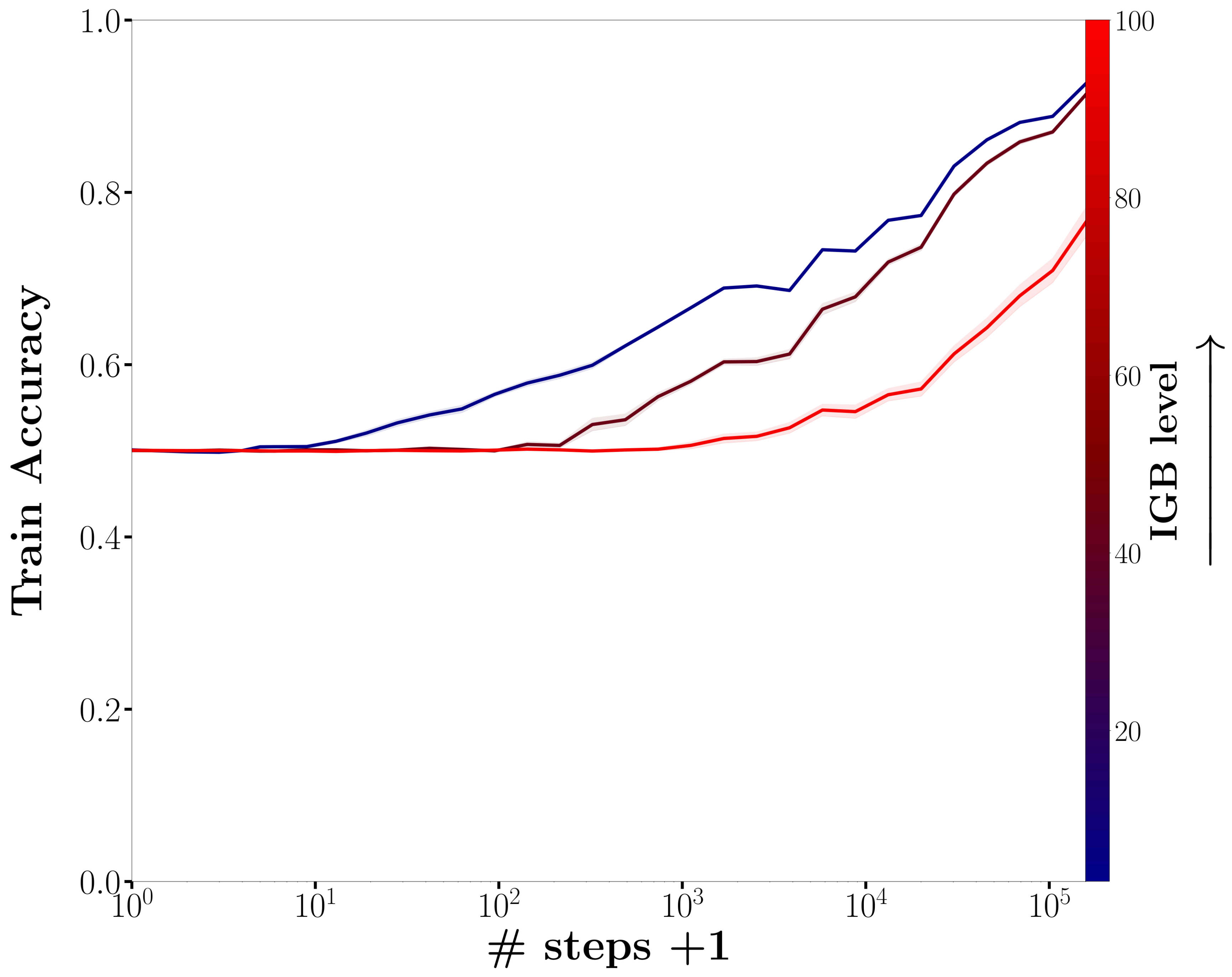


Interplay between Class Imbalance and IGB?

Class Imbalance
arXiv:2207.00391 - **ICML 2023**



Initial Guessing Bias
arXiv:2306.00809 - **ICML 2024**



Interplay between Class Imbalance and IGB?

Summary & Open Questions

How a network is built determines how it starts to guess and how it learns.

Future Directions:

- Dynamics Theory
- Interplay between dataset and model effects
- Interplay between IGB and Class Imbalance

Main References:

- Initial Guessing Bias— arXiv:2306.00809 (ICML 2024)
- Bias and Normalization — arXiv:2505.11312 (under review)
- Bias and Trainability — arXiv:2505.12096 (under review)

Summary & Open Questions

How a network is built determines how it starts to guess and how it learns.

Future Directions:

- Dynamics Theory
- Interplay between dataset and model effects
- Interplay between IGB and Class Imbalance

Main References:

- Initial Guessing Bias— arXiv:2306.00809 (ICML 2024)
- Bias and Normalization — arXiv:2505.11312 (under review)
- Bias and Trainability — arXiv:2505.12096 (under review)

Thanks!