The committee machine: Computational to statistical gaps in learning a two-layers neural network

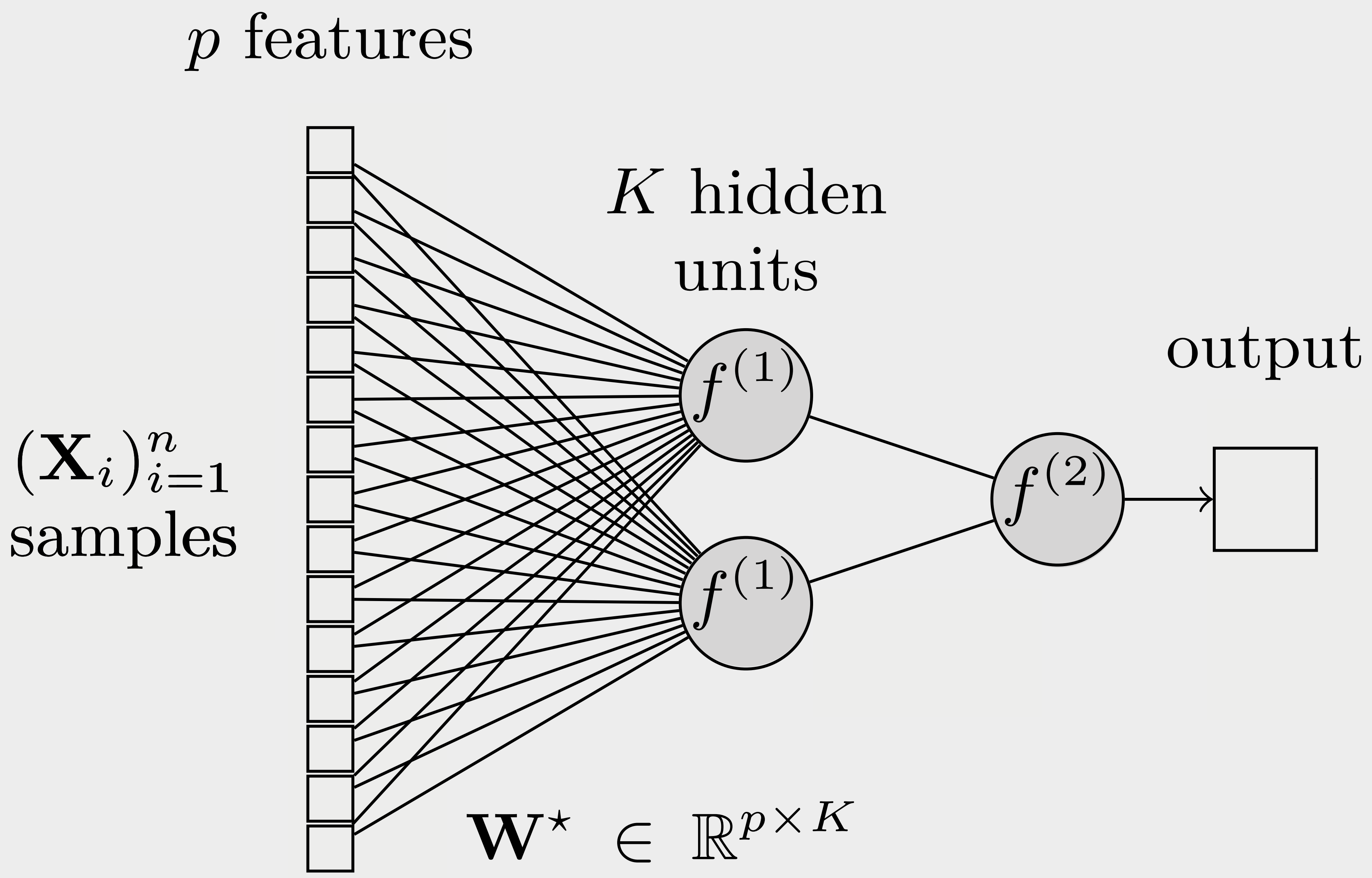
Benjamin Aubin, *Antoine Maillard, Jean Barbier*

Nicolas Macris, Florent Krzakala & Lenka Zdeborovà

« Can we efficiently learn a teacher network from a limited number of samples? »

« Can we efficiently learn a teacher network from a limited number of samples? »

- **Teacher:**

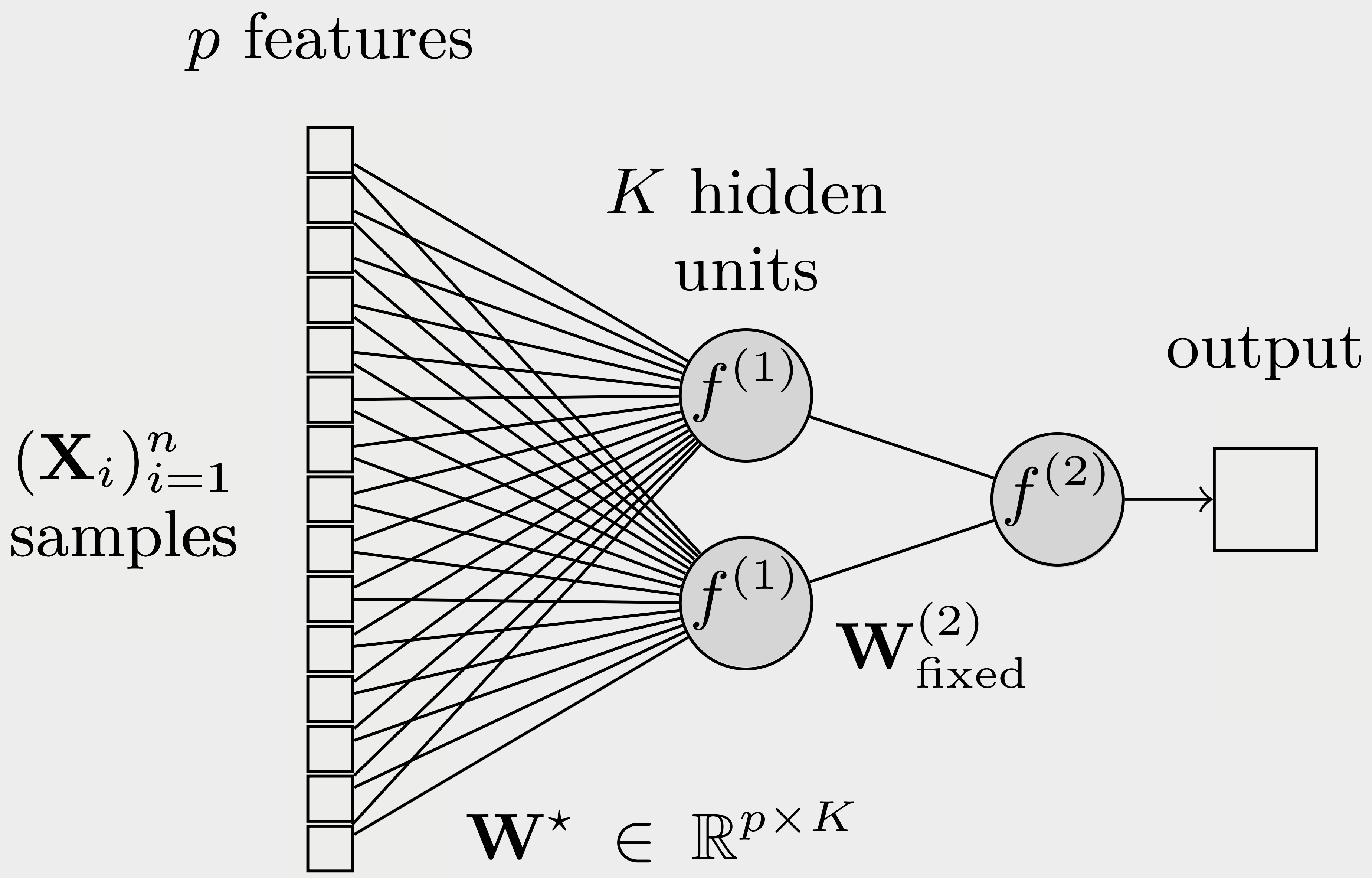


« Can we efficiently learn a teacher network from a limited number of samples? »

- **Teacher:**

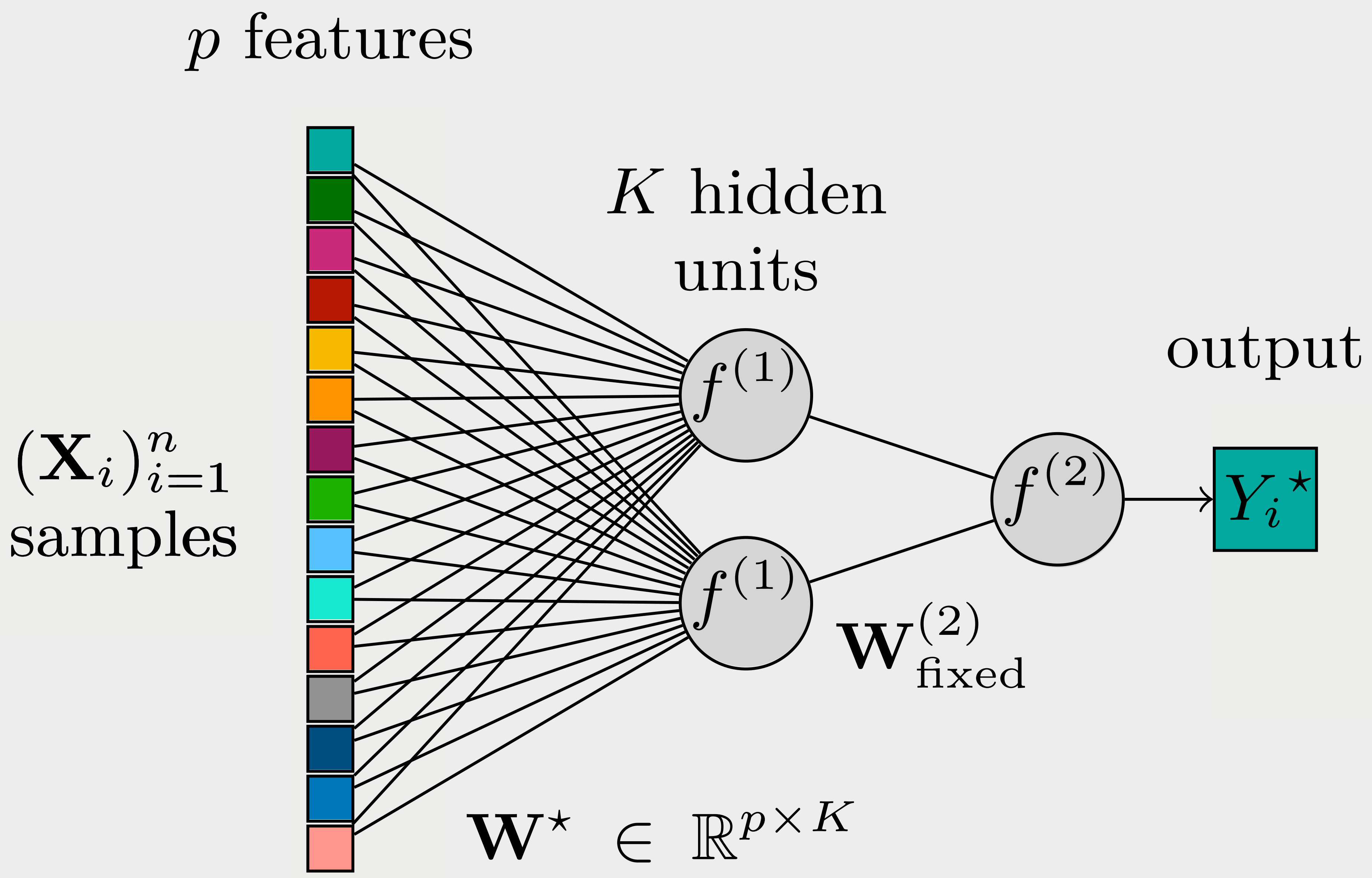
  ✓ Committee machine: second layer fixed

  [Schwarze'93]



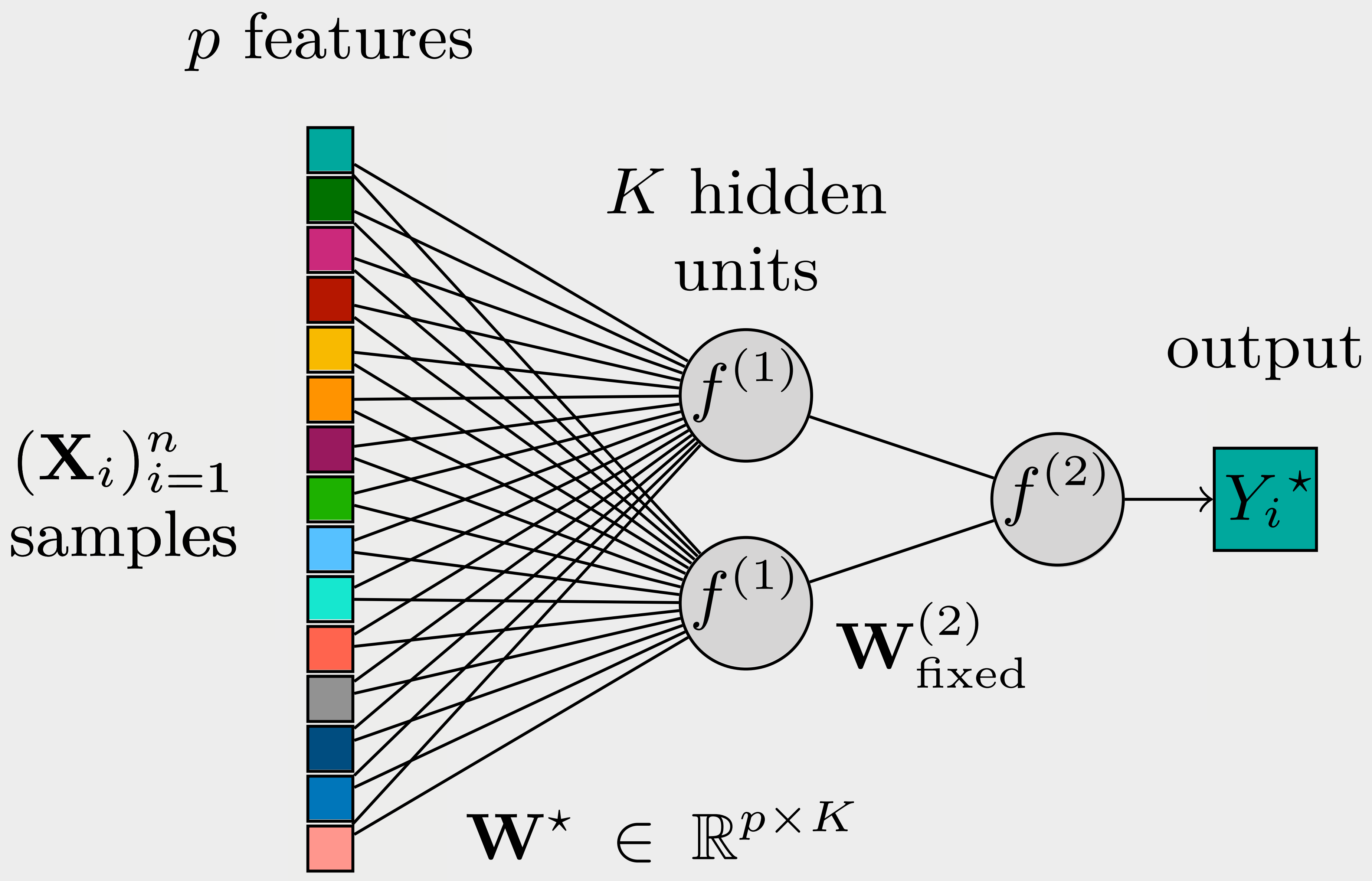
« Can we efficiently learn a teacher network from a limited number of samples? »

- **Teacher:**
  - ✓ Committee machine: second layer fixed
    [Schwarze'93]
  - ✓ *i.i.d* samples

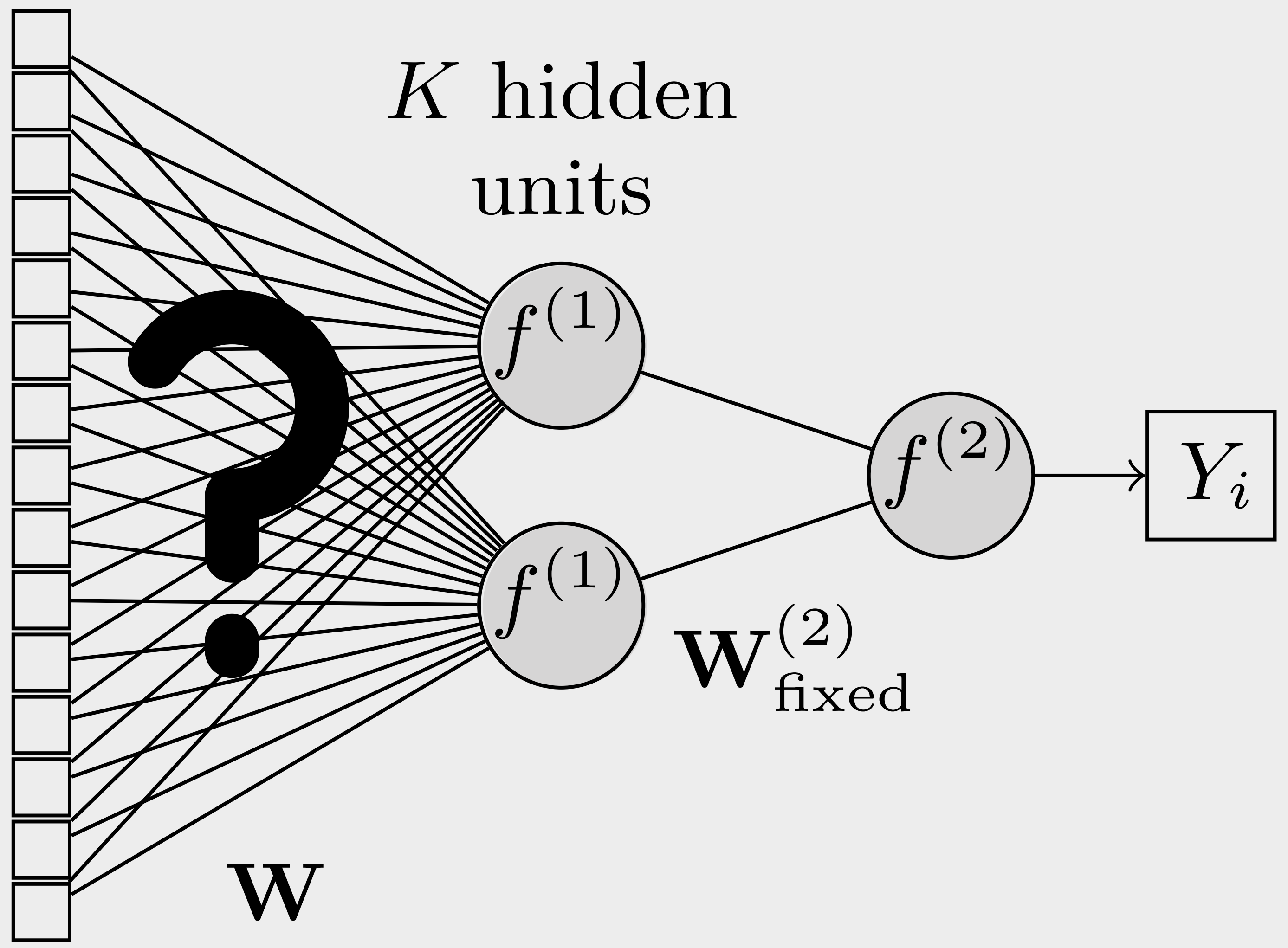


« Can we efficiently learn a teacher network from a limited number of samples? »

- **Teacher:**
  - ✓ Committee machine: second layer fixed
    [Schwarze'93]
  - ✓ *i.i.d* samples

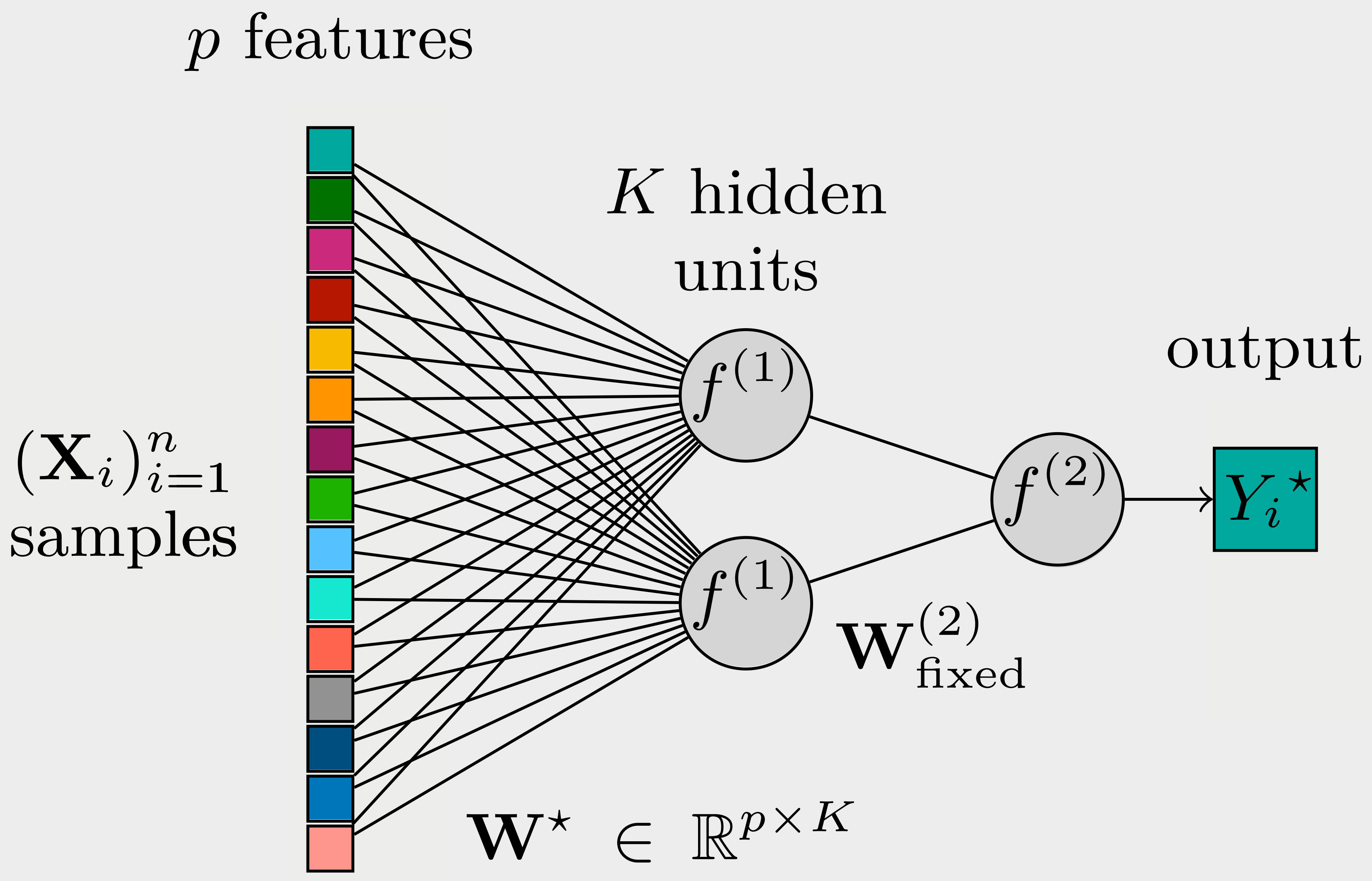


- **Student:**

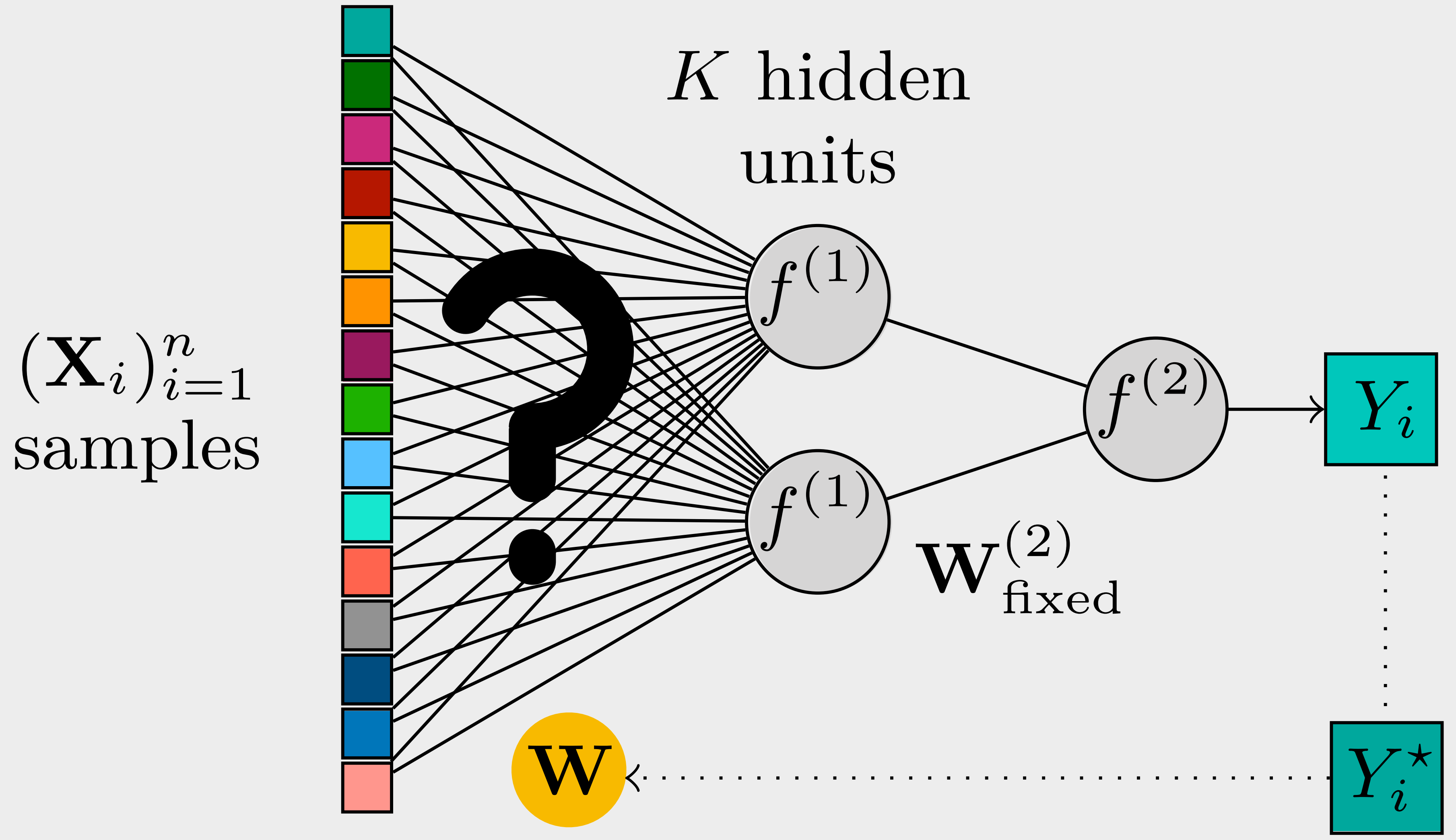


« Can we efficiently learn a teacher network from a limited number of samples? »

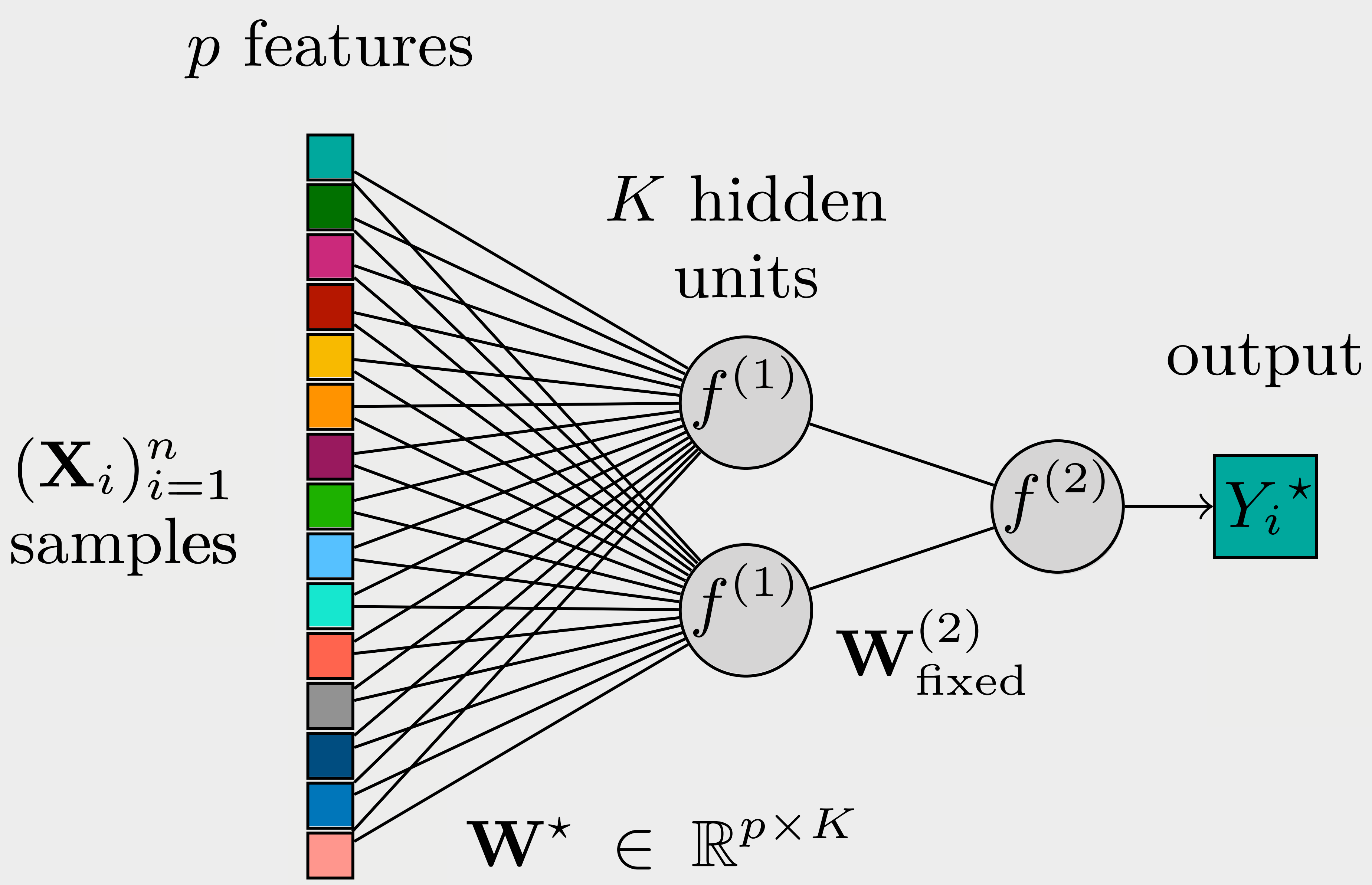
- **Teacher:**
  - ✓ Committee machine: second layer fixed [Schwarze'93]
  - ✓ *i.i.d* samples



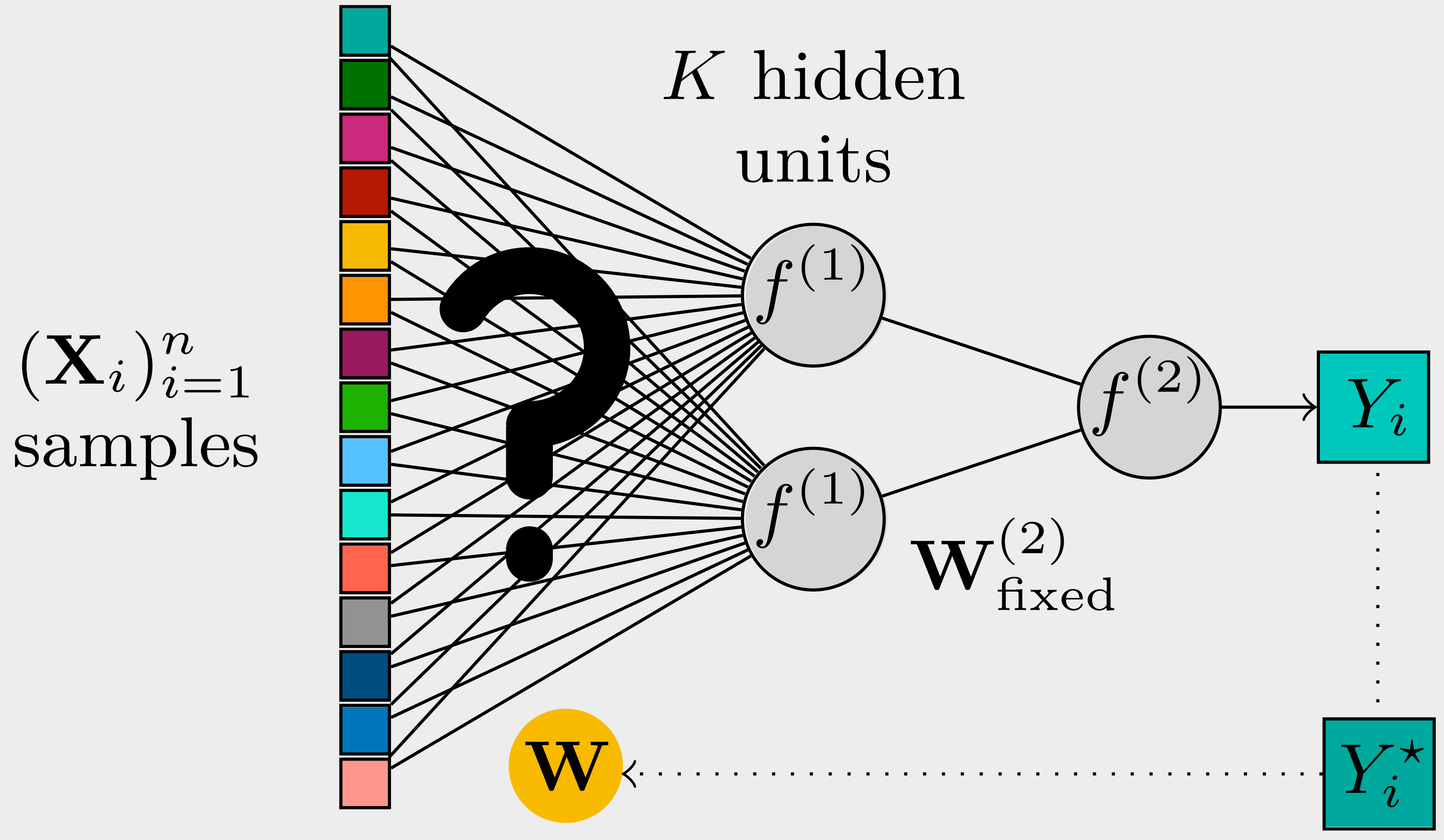
- **Student:**



« Can we efficiently learn a teacher network from a limited number of samples? »



- **Teacher:**
  - ✓ Committee machine: second layer fixed [Schwarze'93]
  - ✓ *i.i.d* samples

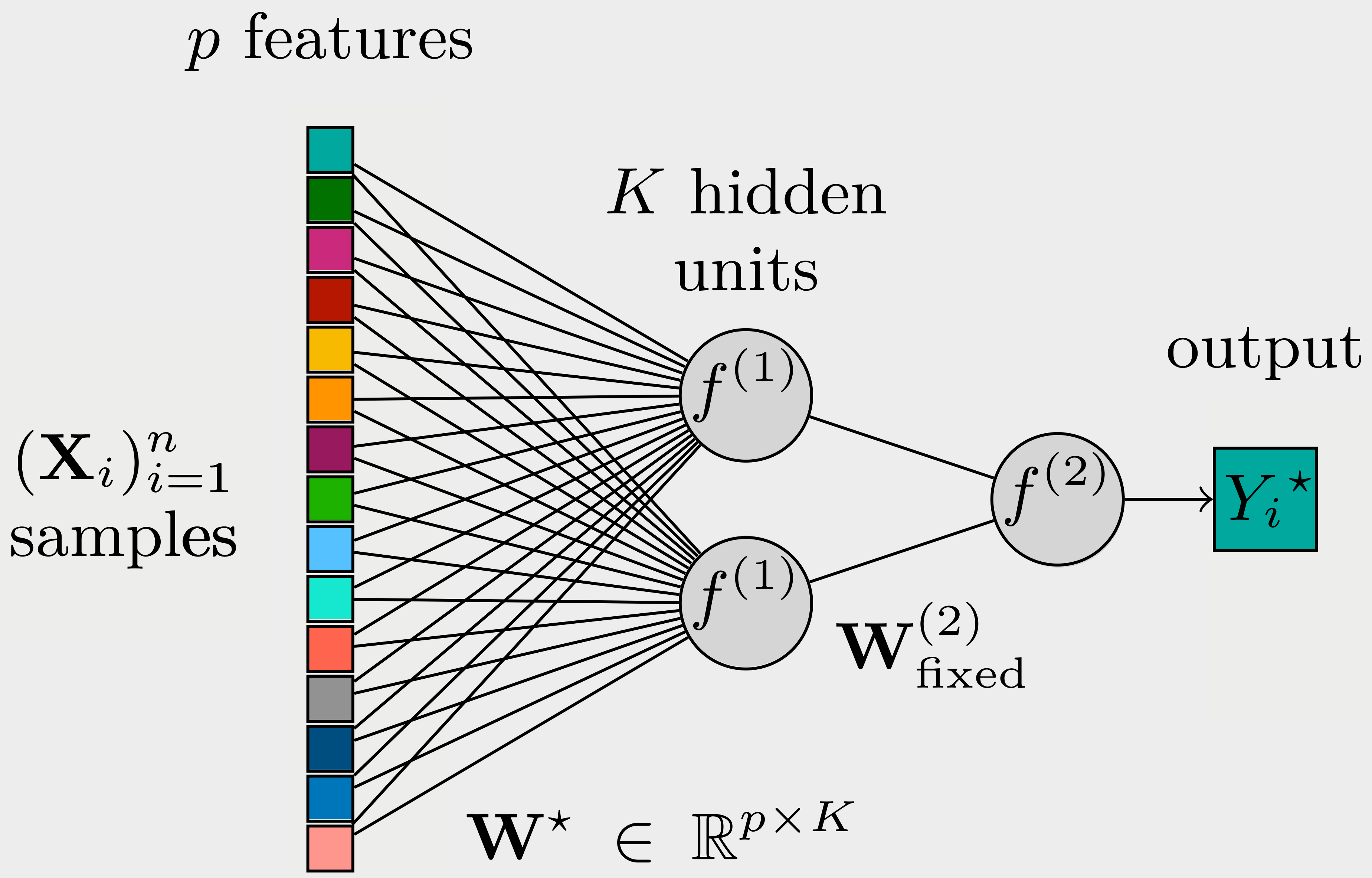


- **Student:**
  - ✓ Learning task possible ?

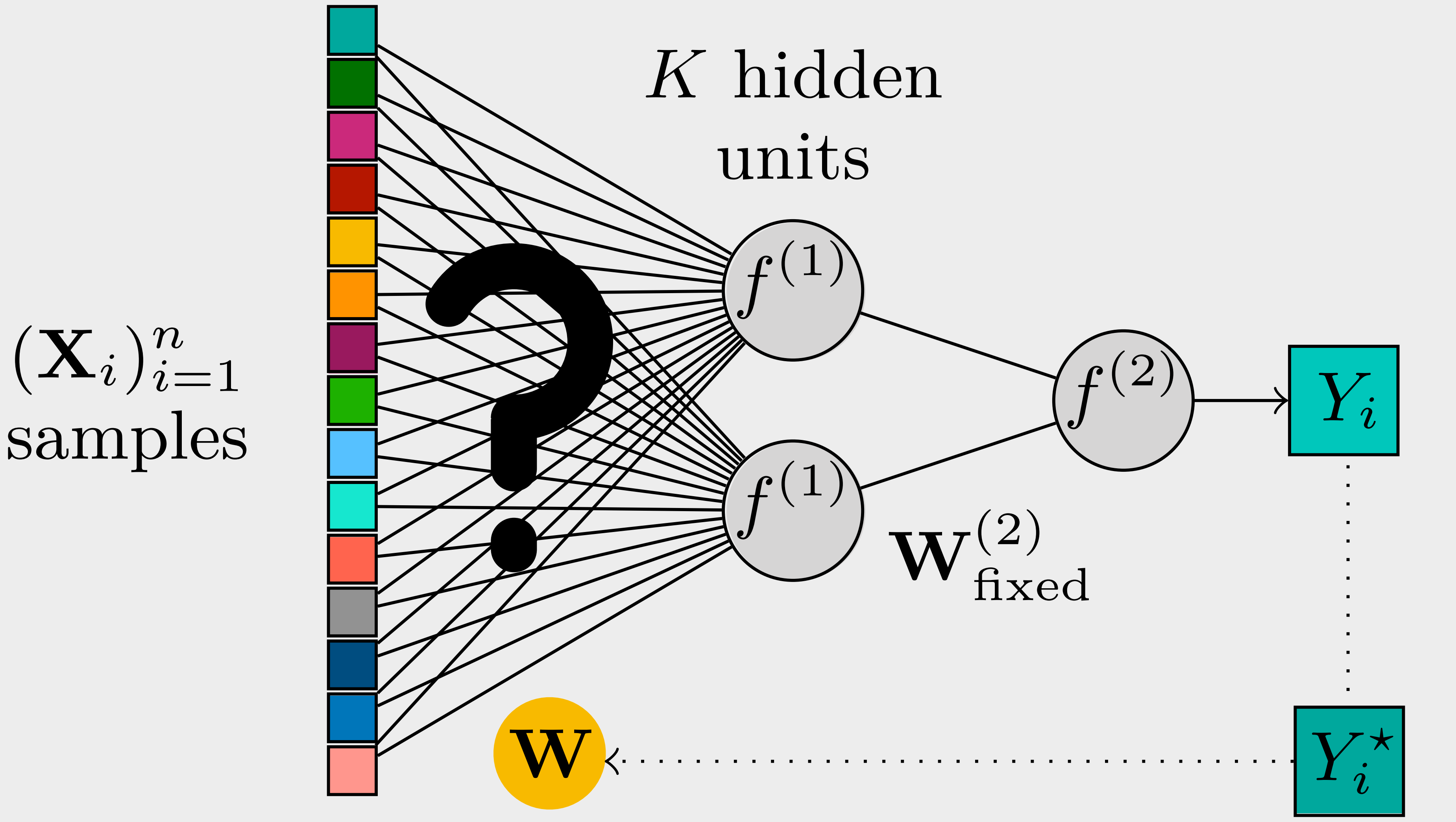
« Can we efficiently learn a teacher network from a limited number of samples? »

- **Teacher:**
  - ✓ Committee machine: second layer fixed

    [Schwarze'93]
  - ✓ *i.i.d* samples



- **Student:**
  - ✓ Learning task possible ?
  - ✓ Computational complexity?



Motivation

➡ **Traditional approach**

- Worst case scenario/PAC bounds: VC-dim & Rademacher complexity
- Numerical experiments

Motivation

➡ **Traditional approach**

- Worst case scenario/PAC bounds: VC-dim & Rademacher complexity
- Numerical experiments

➡ **Complementary approach**

✓ Revisit the statistical physics typical case scenario [Sompolinsky'92, Mezard'87] :
i.i.d data coming from a probabilistic model

Motivation

➡ Traditional approach

- Worst case scenario/PAC bounds: VC-dim & Rademacher complexity
- Numerical experiments

➡ Complementary approach

- ✓ Revisit the statistical physics typical case scenario [Sompolinsky'92, Mezard'87] : *i.i.d* data coming from a probabilistic model
- ✓ Theoretical understanding of the generalization performance
- ✓ Regime: $p \to \infty, \ \frac{n}{p} = \Theta(1)$

Main result (1) - Generalization error

- **Information theoretically optimal generalization error** (Bayes optimal case)

$$\epsilon_g^{(p)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{W}|\mathbf{X}}\left[Y(\mathbf{XW})\right] - Y^\star(\mathbf{XW}^\star)\right)^2\right] \xrightarrow[p\to\infty]{} \epsilon_g(q^*)$$

Main result (1) - Generalization error

- **Information theoretically optimal generalization error**
  (Bayes optimal case)

$$\epsilon_g^{(p)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{W}|\mathbf{X}}\left[Y(\mathbf{XW})\right] - Y^\star(\mathbf{XW}^\star)\right)^2\right] \xrightarrow[p\to\infty]{} \epsilon_g(q^*)$$

EXPLICIT

Main result (1) - Generalization error

- **Information theoretically optimal generalization error**
  (Bayes optimal case)

$$\epsilon_g^{(p)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{W}|\mathbf{X}}\left[Y(\mathbf{XW})\right] - Y^\star(\mathbf{XW}^\star)\right)^2\right] \xrightarrow[p\to\infty]{} \epsilon_g(q^*)$$

EXPLICIT

Main result (1) - Generalization error

- **Information theoretically optimal generalization error** (Bayes optimal case)

$$\epsilon_g^{(p)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{W}|\mathbf{X}}\left[Y(\mathbf{XW})\right] - Y^\star(\mathbf{XW}^\star)\right)^2\right] \xrightarrow[p\to\infty]{} \epsilon_g(q^*)$$

EXPLICIT

- q^*: extremizing the variational formulation of this ***mutual information***:

$$\lim_{p\to\infty}\frac{1}{p}I(\mathbf{W};\mathbf{Y}|\mathbf{X}) = -\sup_{r\in\mathcal{S}_K^+}\inf_{q\in\mathcal{S}_K^+}\left\{\psi_{P_0}(r) + \alpha\Psi_{\text{out}}(q) - \frac{1}{2}\text{Tr}(rq)\right\} + \text{cst}$$

Main result (1) - Generalization error

- **Information theoretically optimal generalization error**
  (Bayes optimal case)

$$\epsilon_g^{(p)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{W}|\mathbf{X}}\left[Y(\mathbf{XW})\right] - Y^\star(\mathbf{XW}^\star)\right)^2\right] \xrightarrow[p\to\infty]{} \epsilon_g(q^*)$$

EXPLICIT

- q^*: extremizing the variational formulation of this ***mutual information***:

$$\lim_{p\to\infty}\frac{1}{p}I(\mathbf{W};\mathbf{Y}|\mathbf{X}) = -\sup_{r\in\mathcal{S}_K^+}\inf_{q\in\mathcal{S}_K^+}\left\{\psi_{P_0}(r) + \alpha\Psi_{\text{out}}(q) - \frac{1}{2}\text{Tr}(rq)\right\} + \text{cst}$$

Heuristic replica mutual information well known in statistical physics since 80's

Main result (1) - Generalization error

- **Information theoretically optimal generalization error** (Bayes optimal case)

$$\epsilon_g^{(p)} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{W}|\mathbf{X}}\left[Y(\mathbf{X}\mathbf{W})\right] - Y^\star(\mathbf{X}\mathbf{W}^\star)\right)^2\right] \xrightarrow[p\to\infty]{} \epsilon_g(q^*)$$

EXPLICIT

- q^*: extremizing the variational formulation of this ***mutual information***:

$$\lim_{p\to\infty}\frac{1}{p}I(\mathbf{W};\mathbf{Y}|\mathbf{X}) = -\sup_{r\in\mathcal{S}_K^+}\inf_{q\in\mathcal{S}_K^+}\left\{\psi_{P_0}(r) + \alpha\Psi_{\text{out}}(q) - \frac{1}{2}\text{Tr}(rq)\right\} + \text{cst}$$

Heuristic replica mutual information well known in statistical physics since 80's

✓ **Main contribution:** rigorous proof by adaptive (Guerra) interpolation

Main result (2) - Message Passing Algorithm

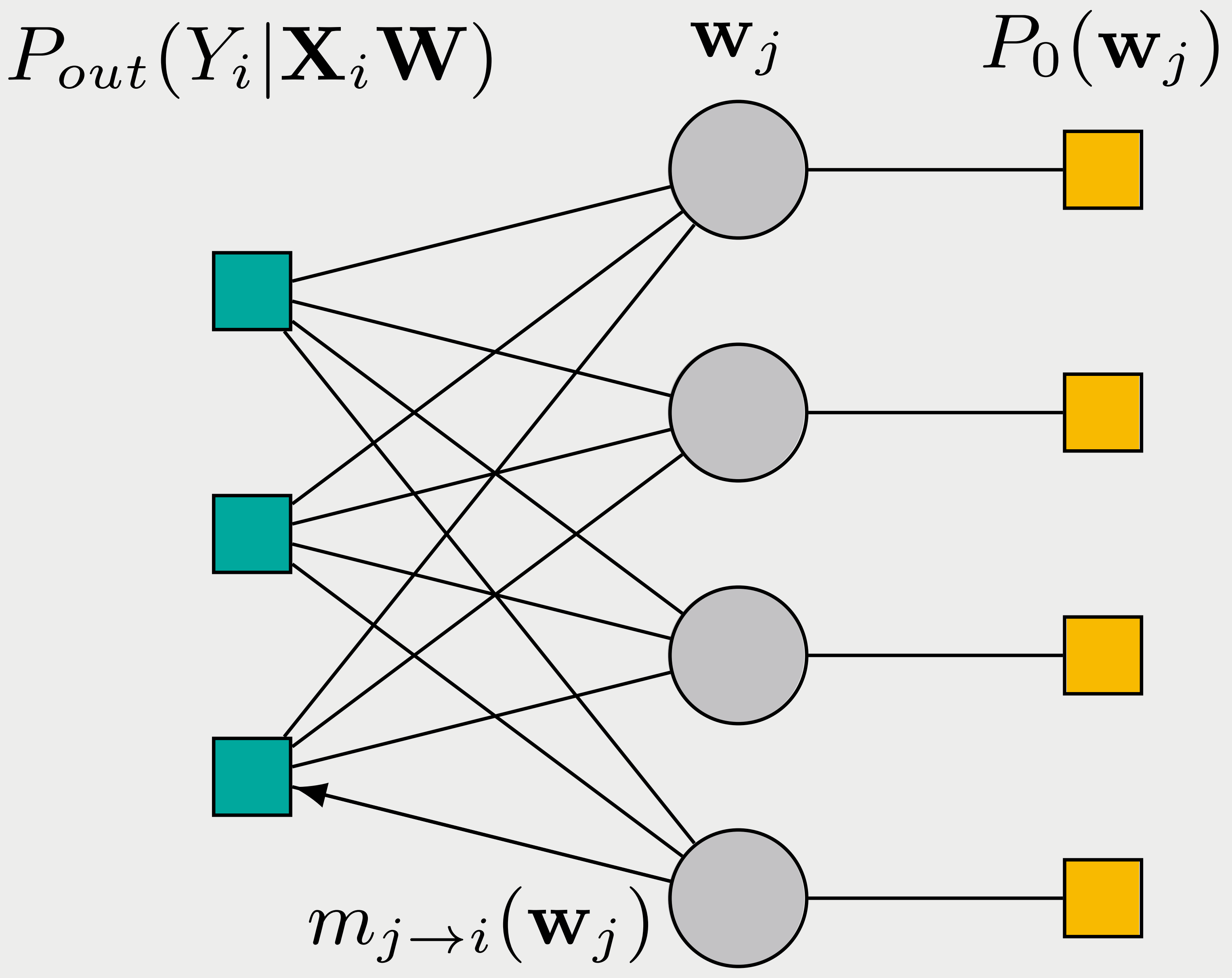
Main result (2) - Message Passing Algorithm

- **Traditional approach:**
  - Minimize a loss function. Not optimal for limited number of samples.

Main result (2) - Message Passing Algorithm

- **Traditional approach:**
  - Minimize a loss function. Not optimal for limited number of samples.

- **Approximate Message Passing (AMP) algorithm:**
  - Expansion of BP equations on a factor graph.
    Closed set of iterative equations.
    Estimates marginal probabilities $m_j(\mathbf{w}_j)$



Factor graph representation of the committee machine

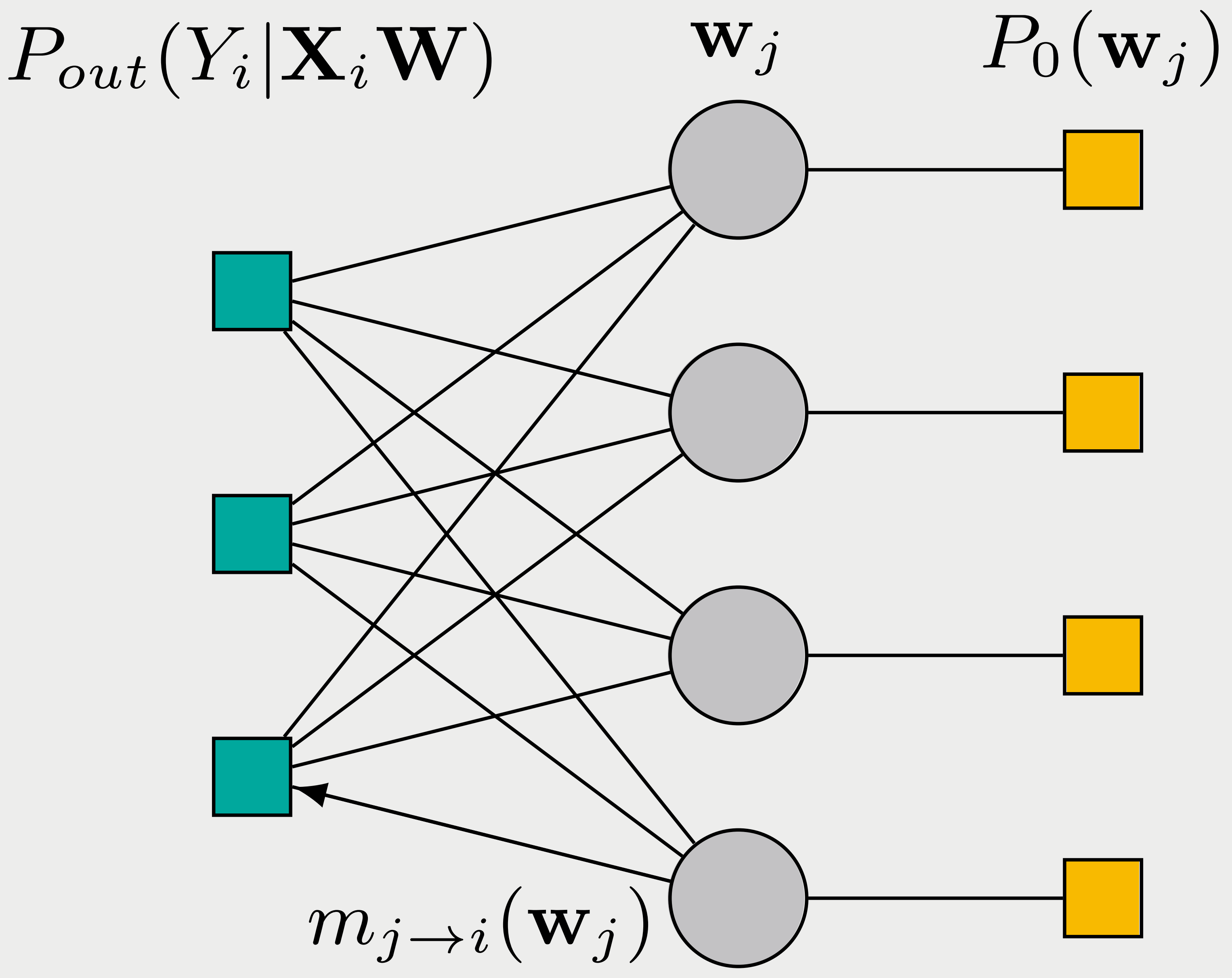
Main result (2) - Message Passing Algorithm

- **Traditional approach:**
  - Minimize a loss function. Not optimal for limited number of samples.

- **Approximate Message Passing (AMP) algorithm:**
  - Expansion of BP equations on a factor graph. Closed set of iterative equations. Estimates marginal probabilities $m_j(\mathbf{w}_j)$

  ✓ Conjectured to be **optimal** among polynomial algorithms

  ✓ Can be **tracked rigorously** (state evolution given by critical points of the replica mutual information) [Montanari-Bayati '10]



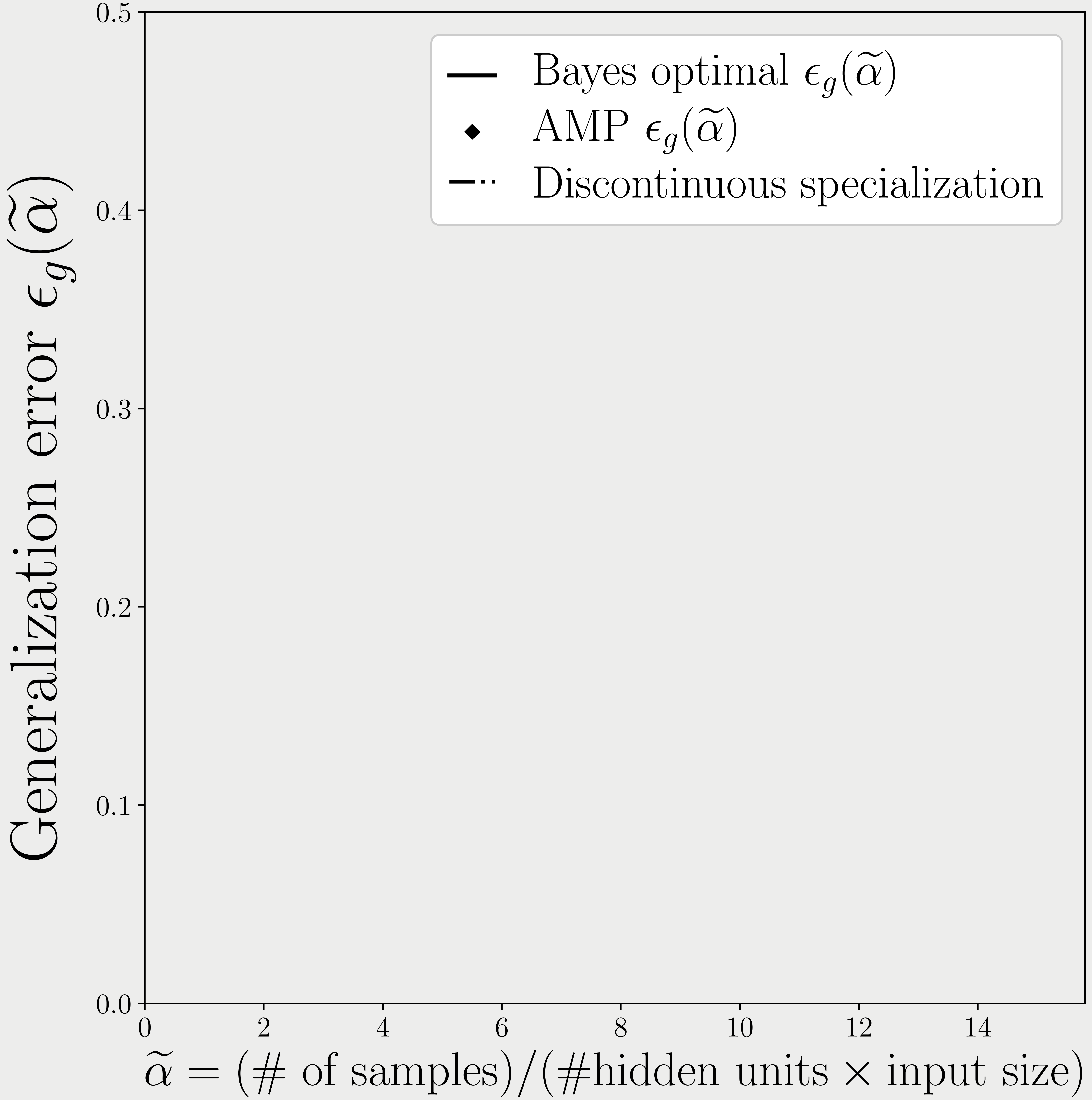
Factor graph representation of the committee machine

Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$

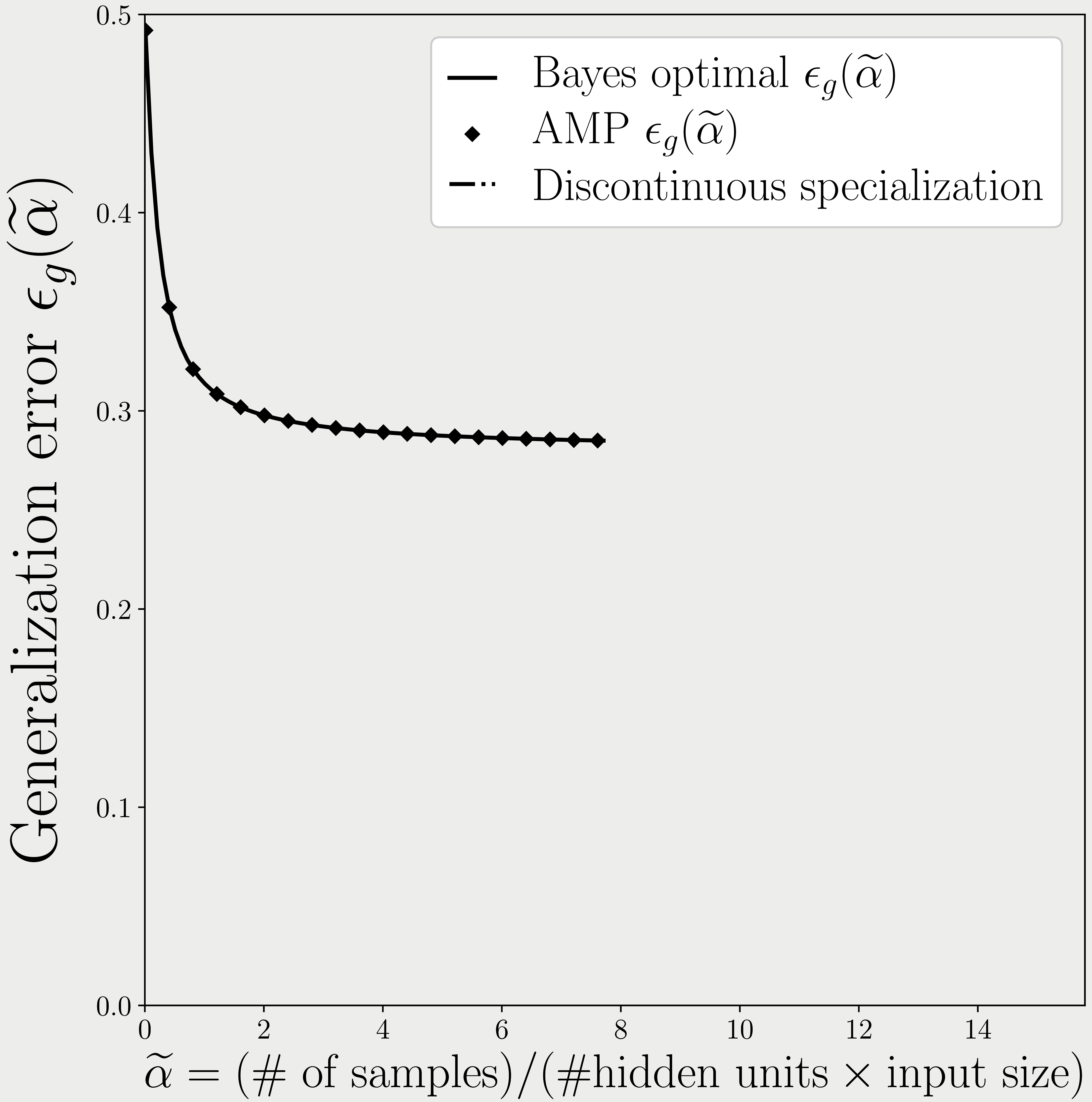
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



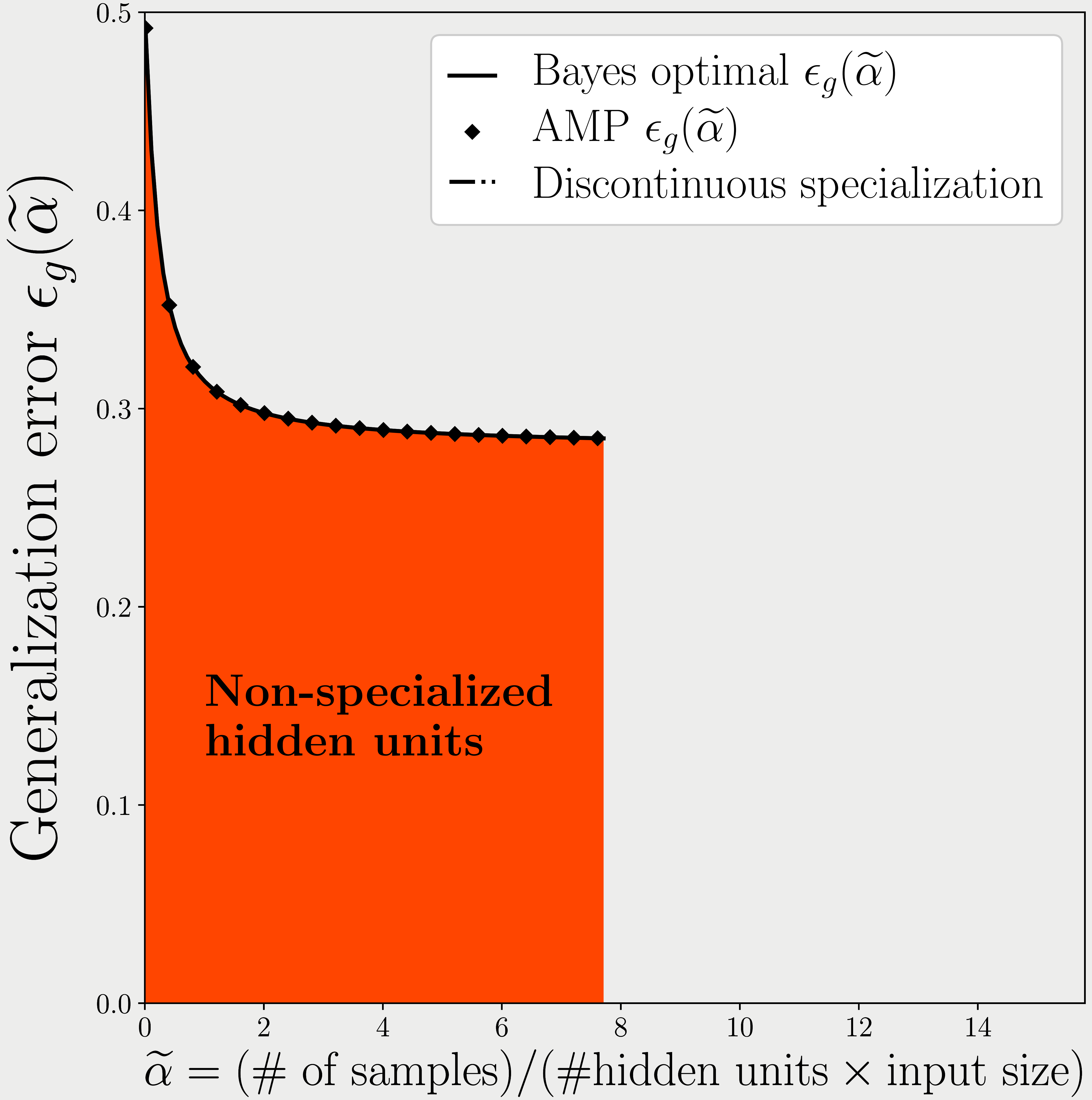
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



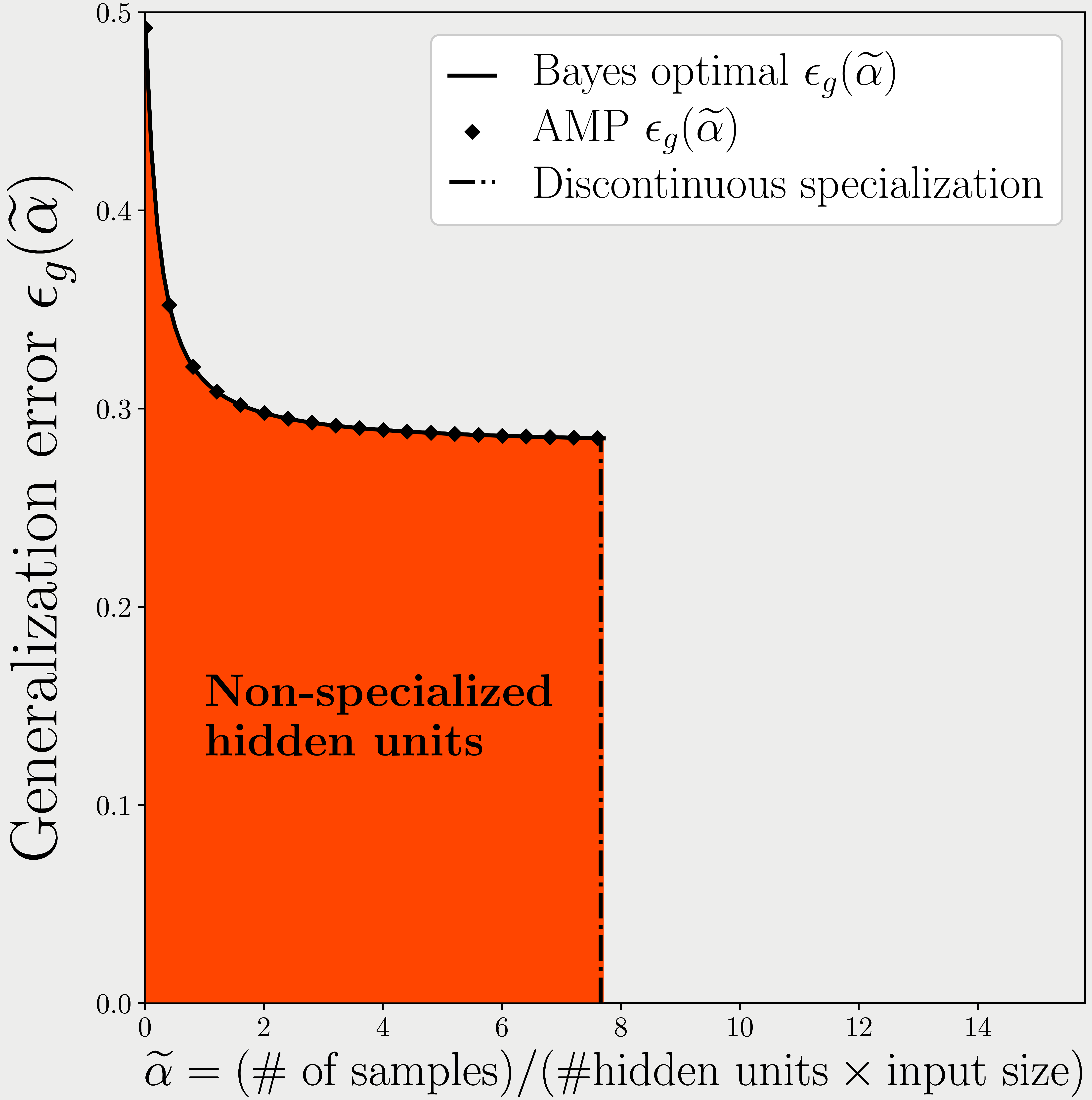
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



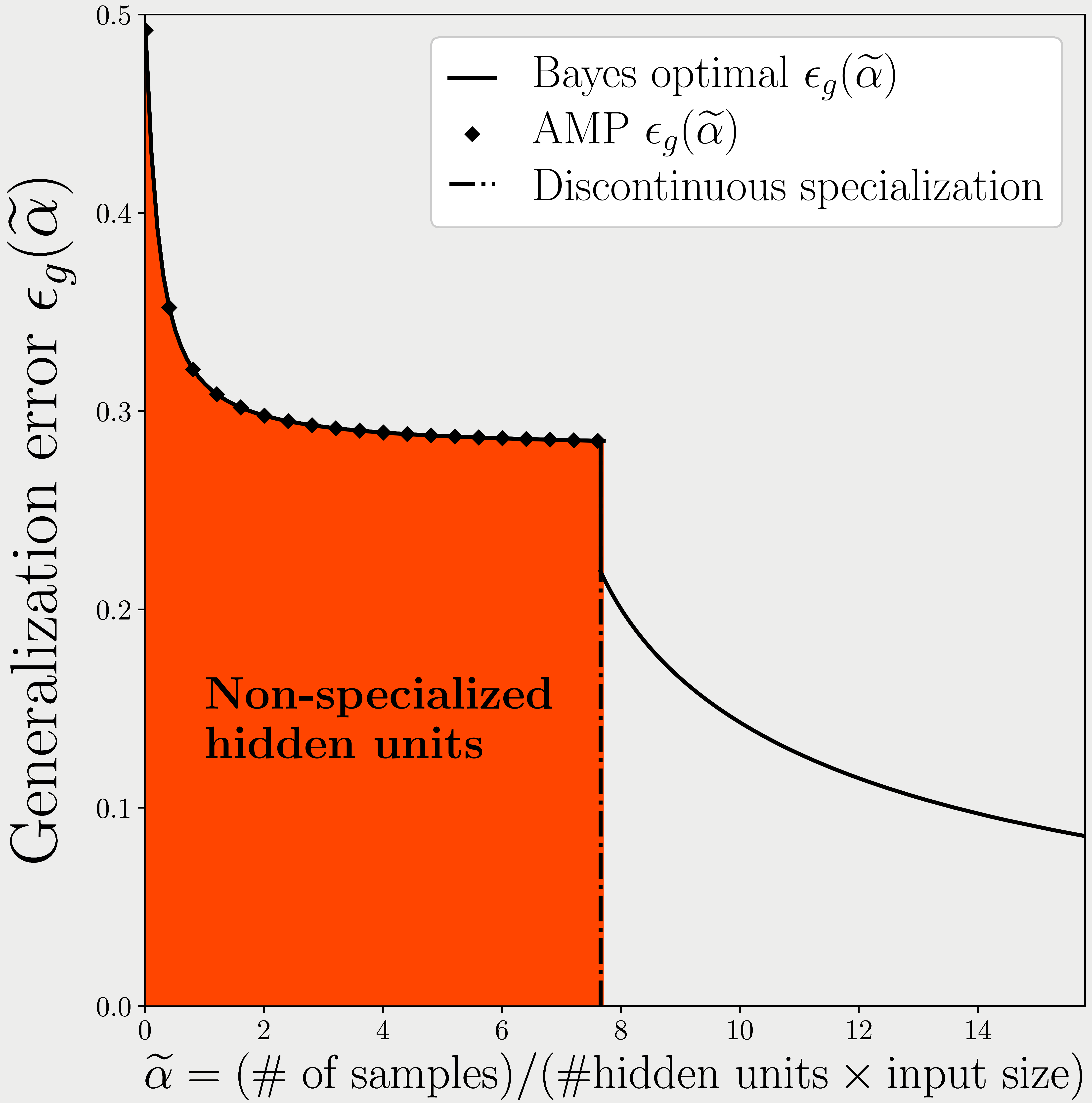
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



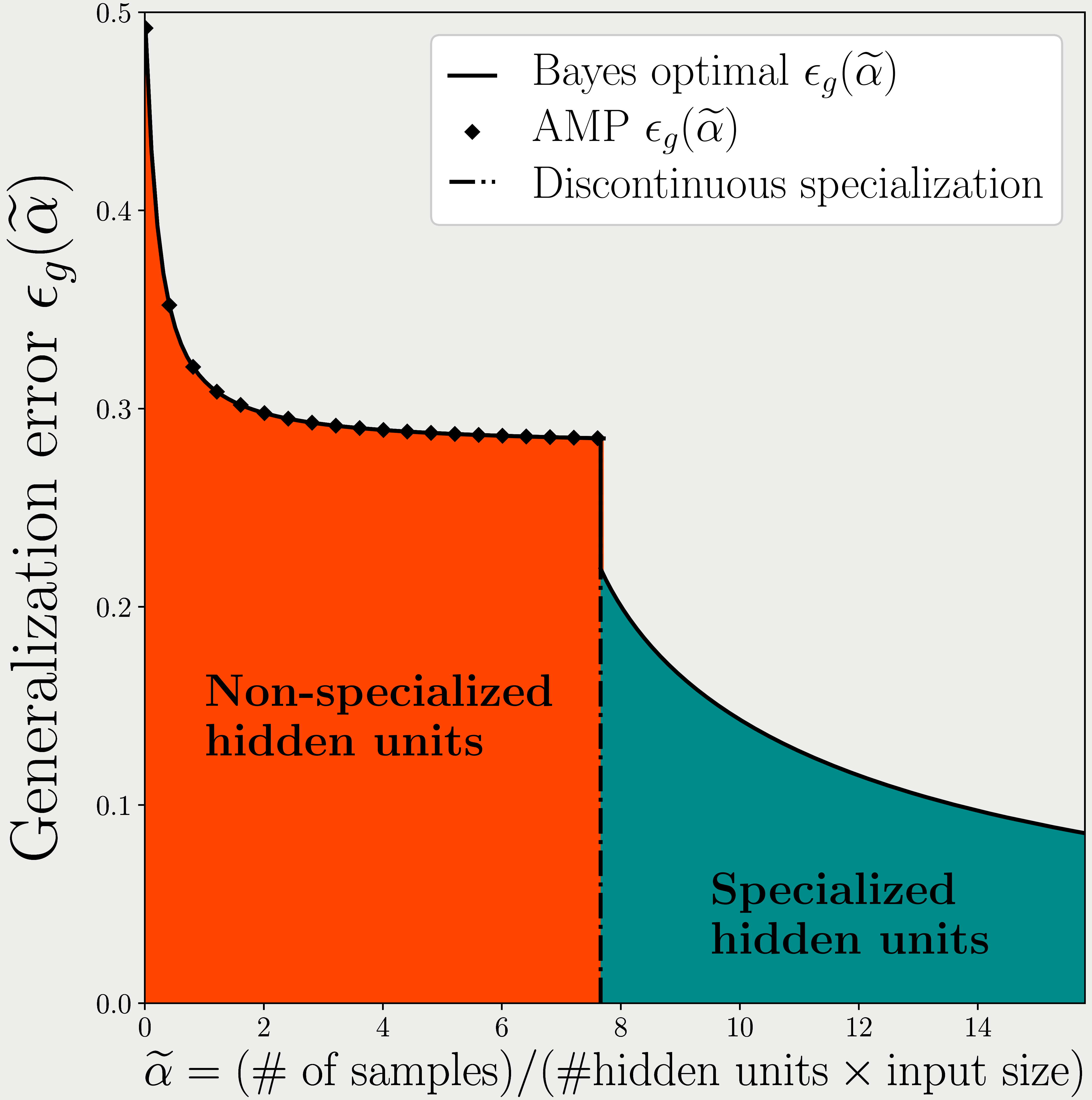
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



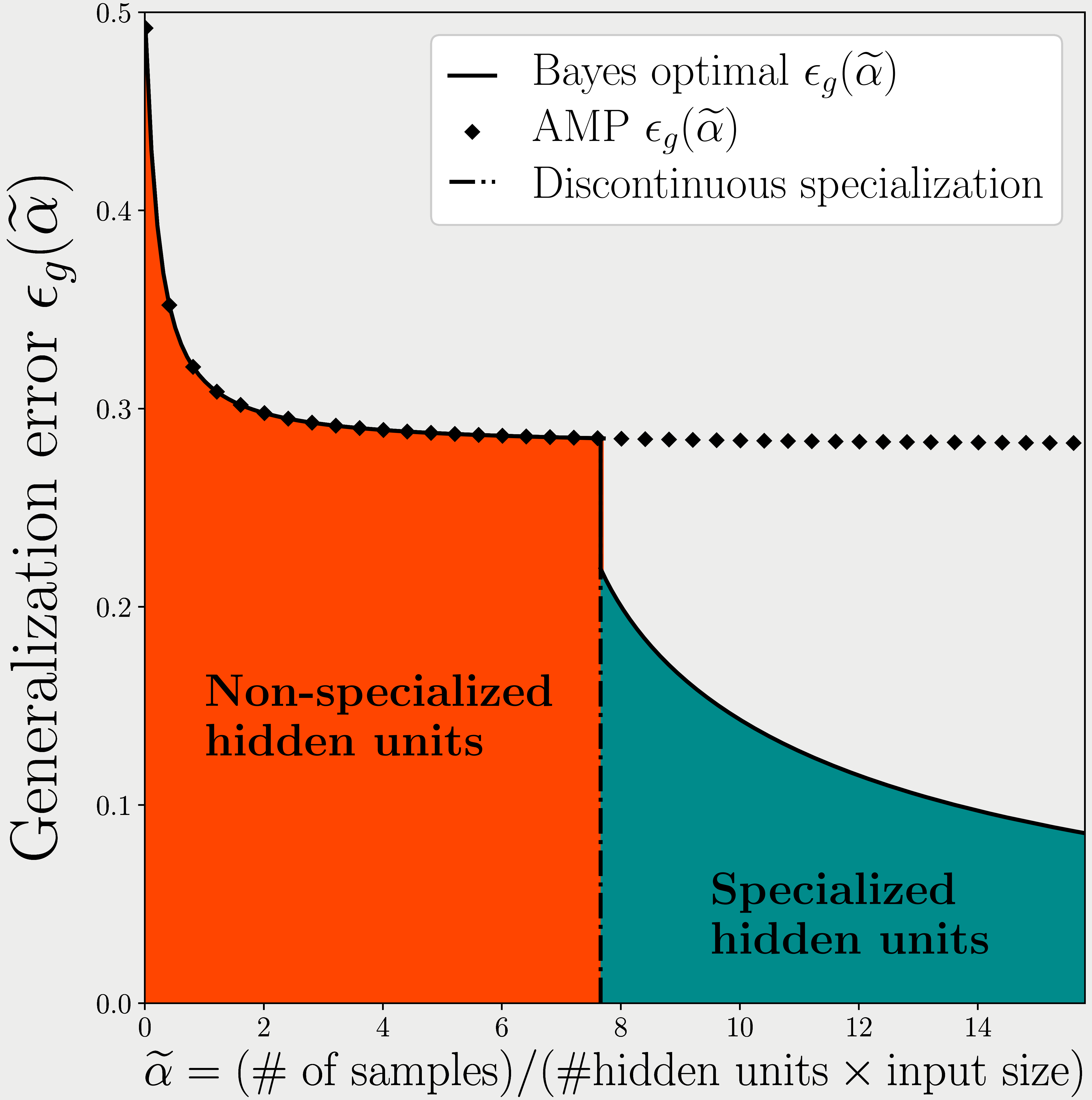
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



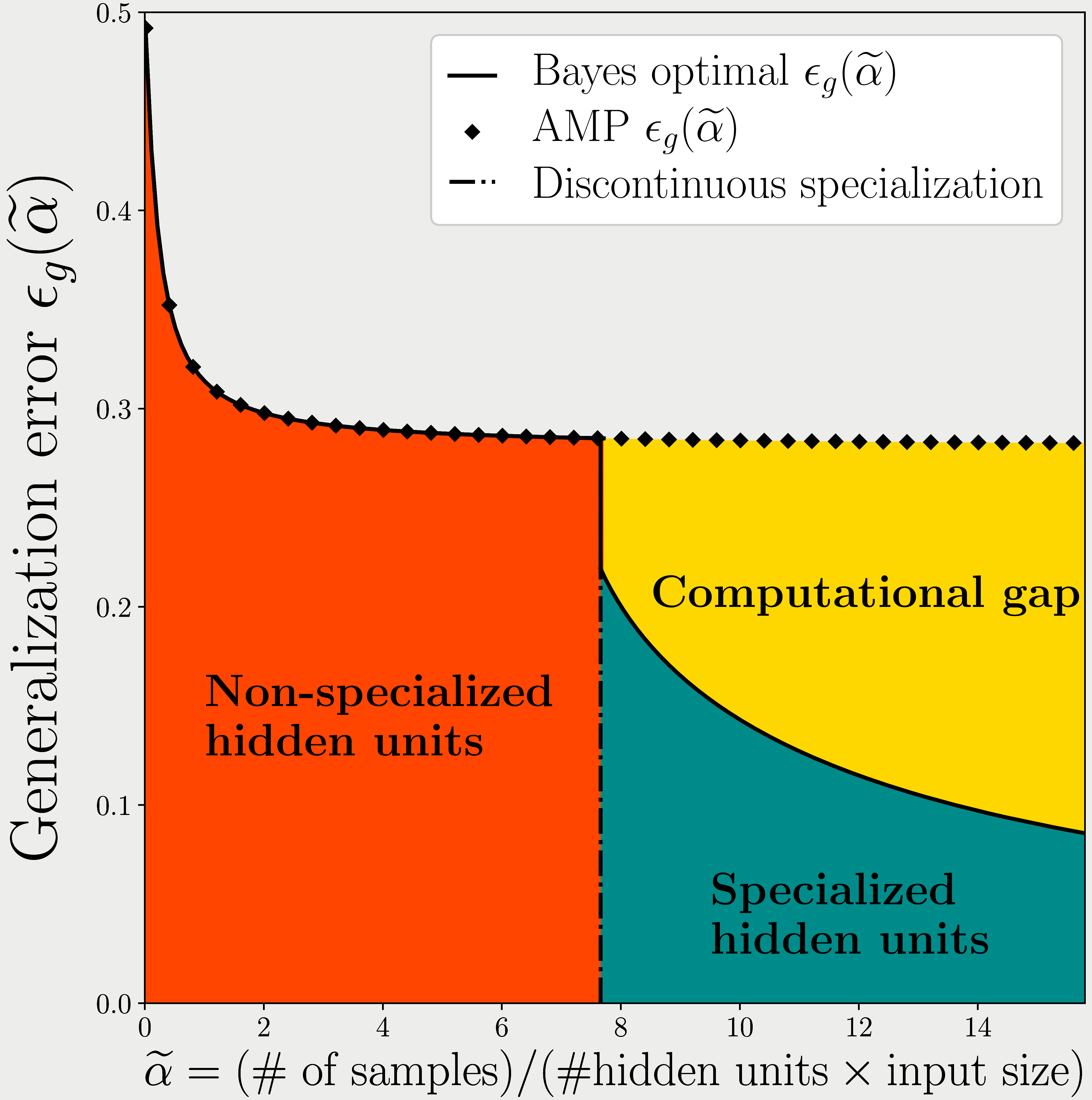
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



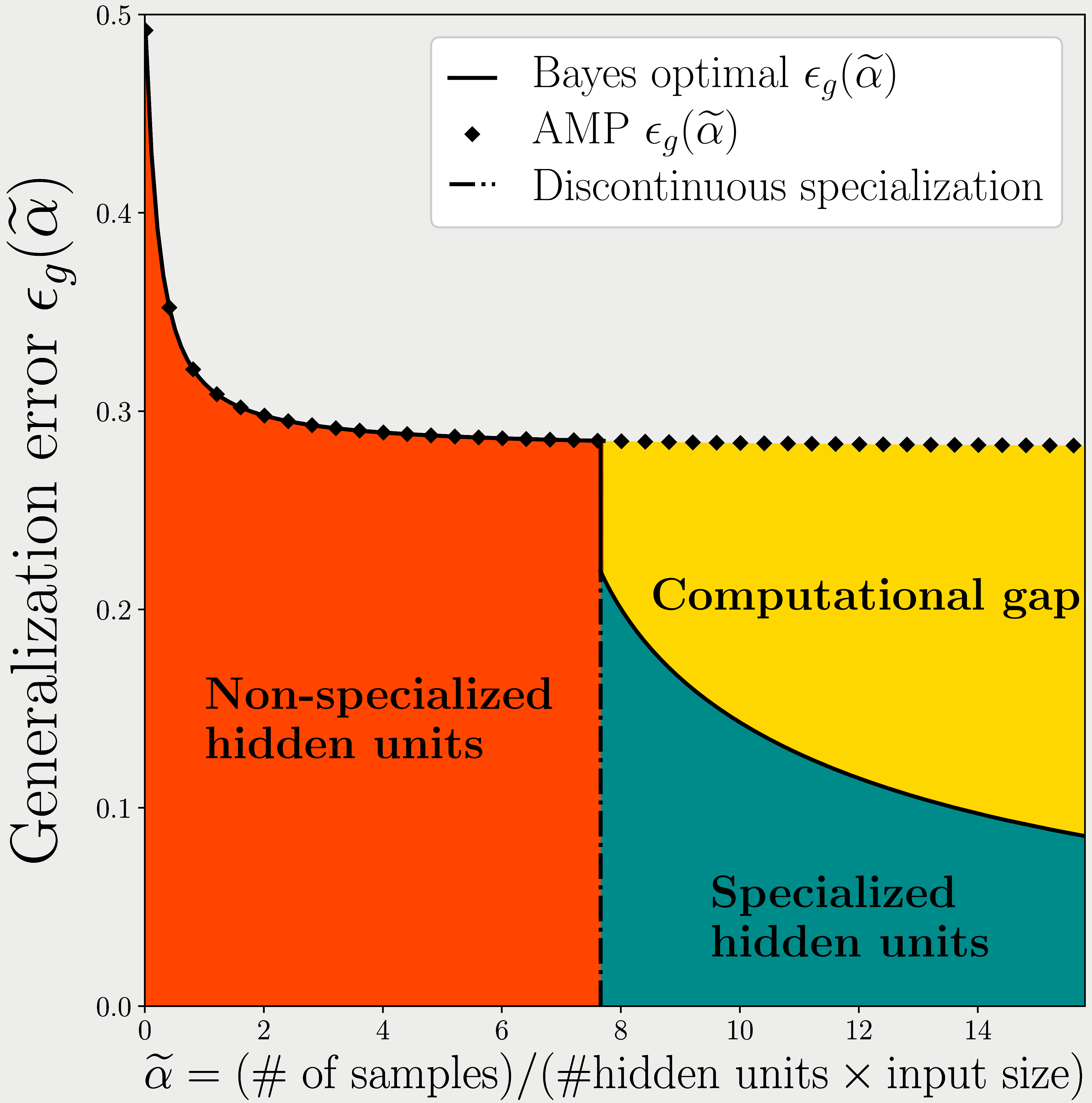
Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



Gaussian weights - sign activation

Large number of hidden units K $= \Theta_p(1)$



TO KNOW MORE:

Poster #111

https://github.com/benjaminaubin/TheCommitteeMachine