Denoising Score Matching with Random Features: Insights on Diffusion Models from Precise Learning Curves

Anand Jerry George
EPFL

Rodrigo Veiga
University of Nottingham

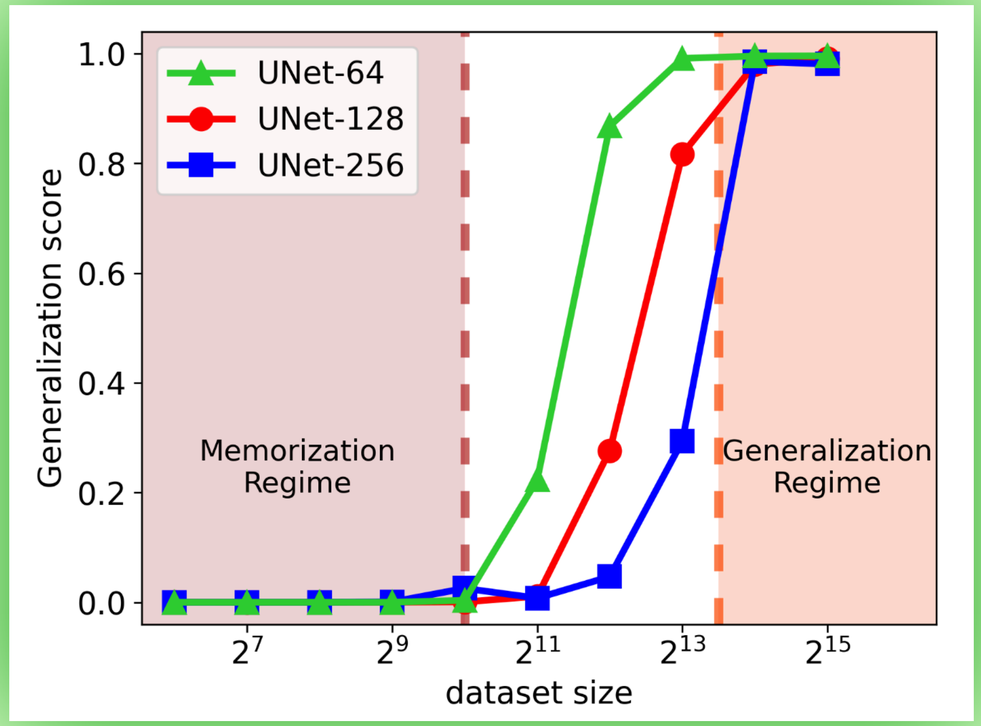
Nicolas Macris
EPFL

➢ Motivation

- Diffusion models are powerful generative models but often **memorize** training data.

➢ Motivation

- Diffusion models are powerful generative models but often **memorize** training data.
- Empirical findings suggest that there is a phase transition from memorization to generalization as we increase the training **dataset size**.

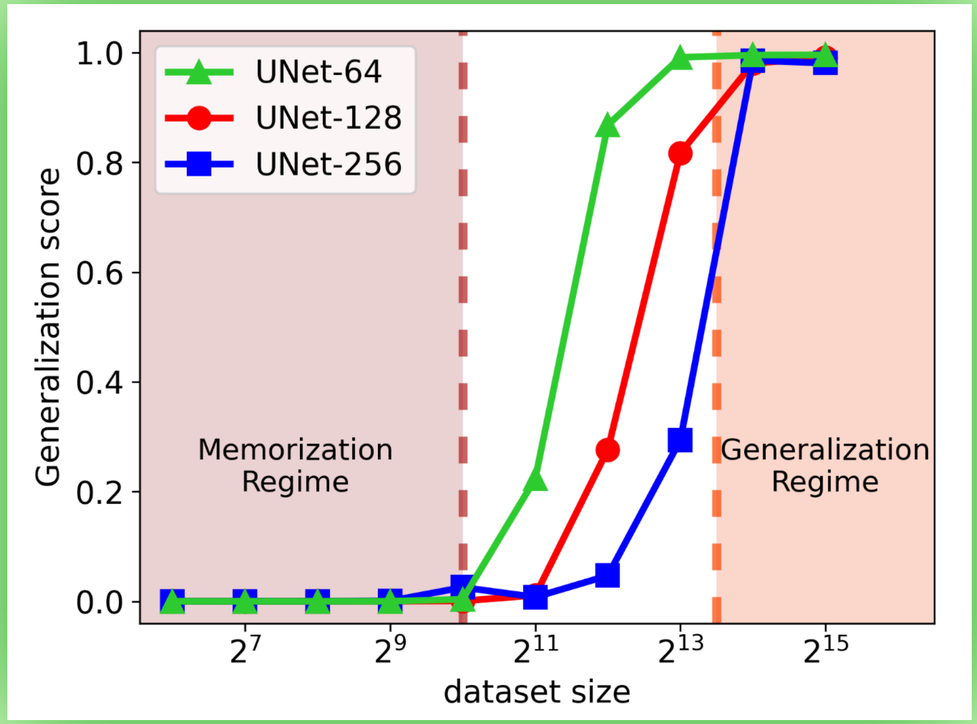


Zhang et. al. 2024

➢ Motivation

- Diffusion models are powerful generative models but often **memorize** training data.
- Empirical findings suggest that there is a phase transition from memorization to generalization as we increase the training **dataset size**.

Goal: To understand the mechanisms behind this phenomenon in an analytically tractable setting.



Zhang et. al. 2024

➢ Background

- Goal of generative models: Draw a **new** sample from P_0, given $\{x_i\}_{i=1}^n \overset{iid}{\sim} P_0$.

➢ Background

- Goal of generative models: Draw a **new** sample from P_0, given $\{x_i\}_{i=1}^n \overset{iid}{\sim} P_0$.
- Define a forward process: standard Ornstein-Uhlenbeck SDE:

$$dX_t = -X_t\,dt + \sqrt{2}\,dB_t\,, \quad X_0 \sim P_0(\mathbb{R}^d)\,. \tag{1}$$

➢ Background

- Goal of generative models: Draw a **new** sample from P_0, given $\{x_i\}_{i=1}^n \overset{iid}{\sim} P_0$.
- Define a forward process: standard Ornstein-Uhlenbeck SDE:

$$dX_t = -X_t\,dt + \sqrt{2}\,dB_t\,, \quad X_0 \sim P_0(\mathbb{R}^d)\,. \tag{1}$$

- Diffusion models operates by **time reversing** the above process

$$-dY_t = (Y_t + 2\nabla \log P_t(Y_t))\,dt + \sqrt{2}\,d\tilde{B}_t\,, \quad Y_T \sim P_T\,. \tag{2}$$

➢ Background

- Goal of generative models: Draw a **new** sample from P_0, given $\{x_i\}_{i=1}^n \overset{iid}{\sim} P_0$.
- Define a forward process: standard Ornstein-Uhlenbeck SDE:

$$dX_t = -X_t\,dt + \sqrt{2}\,dB_t\,, \quad X_0 \sim P_0(\mathbb{R}^d)\,. \tag{1}$$

- Diffusion models operates by **time reversing** the above process

$$-dY_t = (Y_t + 2\nabla \log P_t(Y_t))\,dt + \sqrt{2}\,d\tilde{B}_t\,, \quad Y_T \sim P_T\,. \tag{2}$$

- **Score function** $\nabla \log P_t$ is all you need for time reversing, learned through **Denoising Score Matching (DSM)**. Let $a_t = e^{-t}$, $h_t = 1 - e^{-2t}$, then

$$\mathscr{L}_{\mathsf{DSM}}(s) = \int_0^T dt \sum_i^n \mathbb{E}_{z\sim\mathcal{N}(0,I_d)} \left\| \sqrt{h_t} s(t, a_t x_i + \sqrt{h_t} z) + z \right\|^2 ; x_i \sim P_0\,.$$

➢ Problem setup

- We study the minimizer of the empirical DSM loss:

$$\mathscr{L}^{m}_{\mathsf{DSM}}(s) = \int_0^T dt \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} s(t, a_t x_i + \sqrt{h_t} z_{ij}^t) + z_{ij}^t \right\|^2 ; x_i \sim P_0,\ z_{ij}^t \sim \mathcal{N}(0, I_d)$$

➢ Problem setup

- We study the minimizer of the empirical DSM loss:

$$\mathscr{L}^{m}_{\mathsf{DSM}}(s) = \int_0^T dt \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} s(t, a_t x_i + \sqrt{h_t} z_{ij}^t) + z_{ij}^t \right\|^2 ; x_i \sim P_0,\ z_{ij}^t \sim \mathcal{N}(0, I_d)$$

- Score parameterized by **Random features** neural network $s_{A_t}(t,x) = A_t \varrho\left(W_t x\right).$

➢ Problem setup

- We study the minimizer of the empirical DSM loss:

$$\mathscr{L}^{m}_{\mathsf{DSM}}(s) = \int_0^T dt \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} s(t, a_t x_i + \sqrt{h_t} z_{ij}^t) + z_{ij}^t \right\|^2 ; x_i \sim P_0,\ z_{ij}^t \sim \mathcal{N}(0, I_d)$$

- Score parameterized by **Random features** neural network $s_{A_t}(t,x) = A_t \varrho\left(W_t x\right).$
- Data: $P_0 \sim \mathcal{N}(0,C)$; C can be isotropic, or have low dimensional structure.

➢ Problem setup

- We study the minimizer of the empirical DSM loss:

$$\mathscr{L}^{m}_{\mathsf{DSM}}(s) = \int_0^T dt \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} s(t, a_t x_i + \sqrt{h_t} z_{ij}^t) + z_{ij}^t \right\|^2 ; x_i \sim P_0,\ z_{ij}^t \sim \mathcal{N}(0, I_d)$$

- Score parameterized by **Random features** neural network $s_{A_t}(t,x) = A_t \varrho\left(W_t x\right).$
- Data: $P_0 \sim \mathcal{N}(0,C)$; C can be isotropic, or have low dimensional structure.
- Scaling regime: $d, D, n, p \to \infty,\ \psi_D = D/d,\ \psi_n = n/d,\ \psi_p = p/d$ fixed, where

  d : input dimension, D: dimension of latent space, n : # of training samples,

  p : # of random features, m : # of noise samples per training sample.

➢ Main Ideas

Metrics used: For the minimizer of the DSM loss $\hat{A}_t$, we define

$$\mathscr{E}^m_{\text{test}}(\hat{A}_t) = \frac{1}{d}\mathbb{E}_{x\sim P_t}\left\| \hat{A}_t\varrho\left(W_t x\right) - \nabla \log P_t(x)\right\|^2$$

$$\mathscr{E}^m_{\text{train}}(\hat{A}_t) = \frac{1}{dnm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\| \sqrt{h_t}\hat{A}_t\varrho\left(W_t(a_t x_i + \sqrt{h_t}z_{ij}^t)\right) + z_{ij}^t\right\|^2$$

➢ Main Ideas

Metrics used: For the minimizer of the DSM loss $\hat{A}_t$, we define

$$\mathscr{E}^m_{\text{test}}(\hat{A}_t) = \frac{1}{d}\mathbb{E}_{x\sim P_t}\left\| \hat{A}_t\varrho\left(W_t x\right) - \nabla \log P_t(x)\right\|^2$$

$$\mathscr{E}^m_{\text{train}}(\hat{A}_t) = \frac{1}{dnm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\| \sqrt{h_t}\hat{A}_t\varrho\left(W_t(a_t x_i + \sqrt{h_t}z^t_{ij})\right) + z^t_{ij}\right\|^2$$

- We derive asymptotically precise expressions for $\mathscr{E}^m_{\text{test}}(\hat{A}_t)$ and $\mathscr{E}^m_{\text{train}}(\hat{A}_t)$ as a function of ψ_D, ψ_n, ψ_p, and t, for $m = 1,\infty$. **Linear pencils** technique from Random Matrix Theory is used to derive the learning curves.

➢ Main Ideas

Metrics used: For the minimizer of the DSM loss $\hat{A}_t$, we define

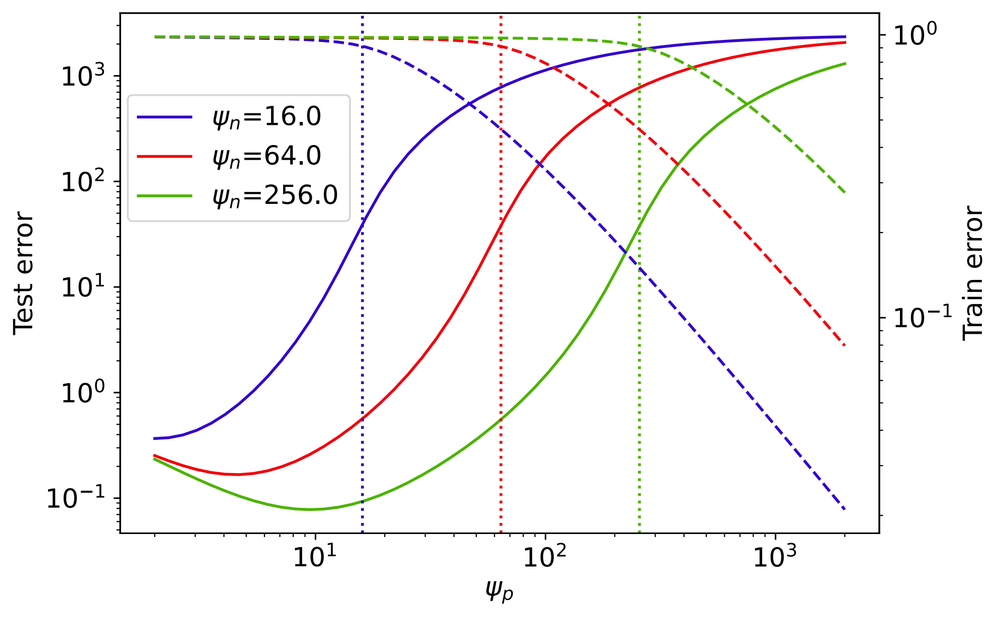
$$\mathscr{E}^m_{\text{test}}(\hat{A}_t) = \frac{1}{d}\mathbb{E}_{x\sim P_t}\left\| \hat{A}_t\varrho\left(W_t x\right) - \nabla \log P_t(x)\right\|^2$$

$$\mathscr{E}^m_{\text{train}}(\hat{A}_t) = \frac{1}{dnm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\| \sqrt{h_t}\hat{A}_t\varrho\left(W_t(a_t x_i + \sqrt{h_t}z_{ij}^t)\right) + z_{ij}^t\right\|^2$$

- We derive asymptotically precise expressions for $\mathscr{E}^m_{\text{test}}(\hat{A}_t)$ and $\mathscr{E}^m_{\text{train}}(\hat{A}_t)$ as a function of ψ_D, ψ_n, ψ_p, and t, for $m = 1,\infty$. **Linear pencils** technique from Random Matrix Theory is used to derive the learning curves.
- We interpret the obtained learning curves to identify regimes of memorization and generalization.

➢ Results

Learning curves for small t:

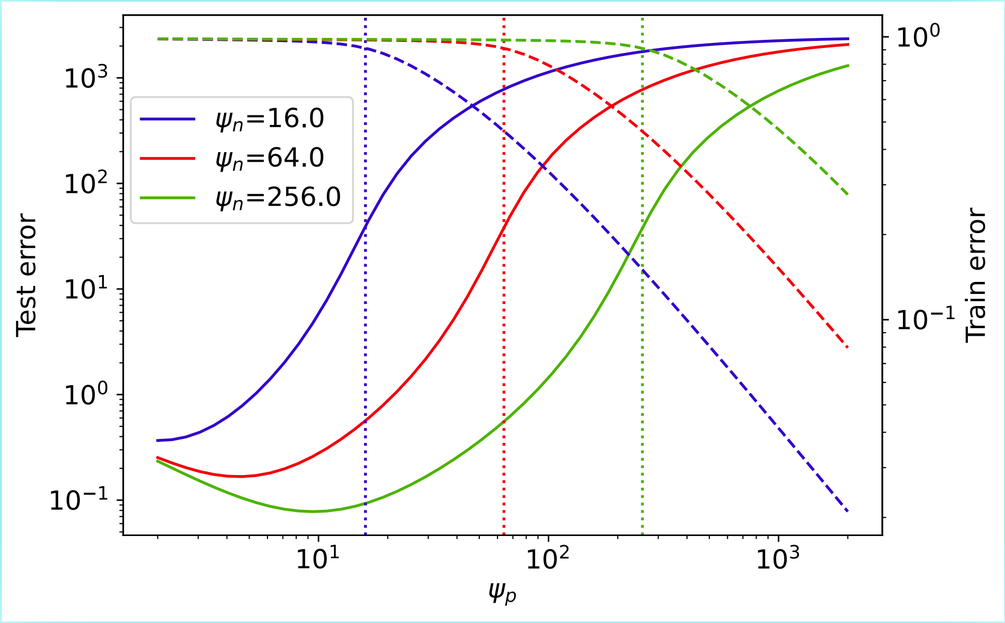


$m = \infty$

Classical U-shaped learning curves

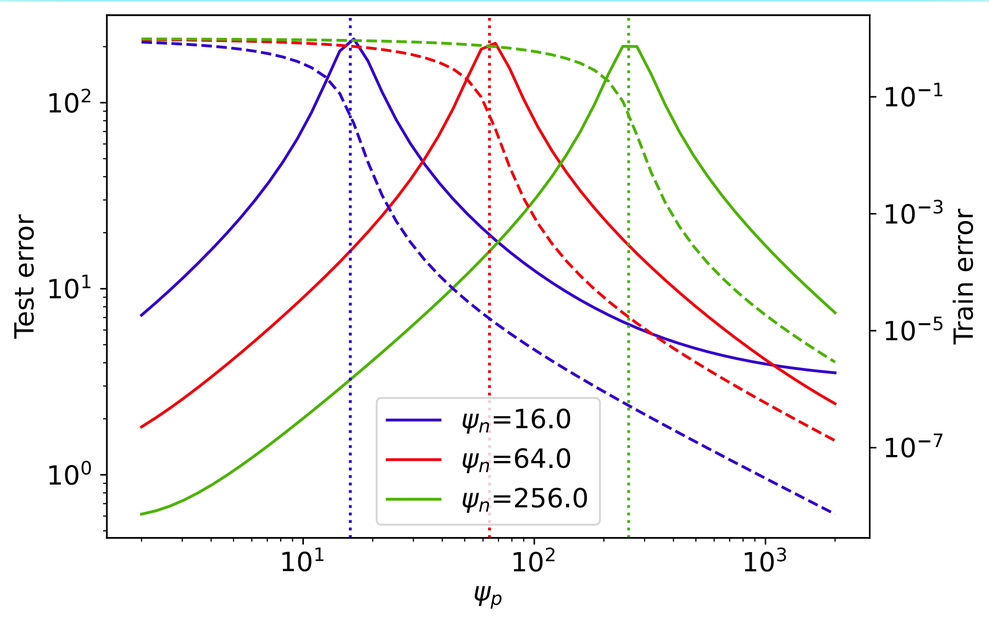
➢ Results

Learning curves for small t:



$m = \infty$

Classical U-shaped learning curves

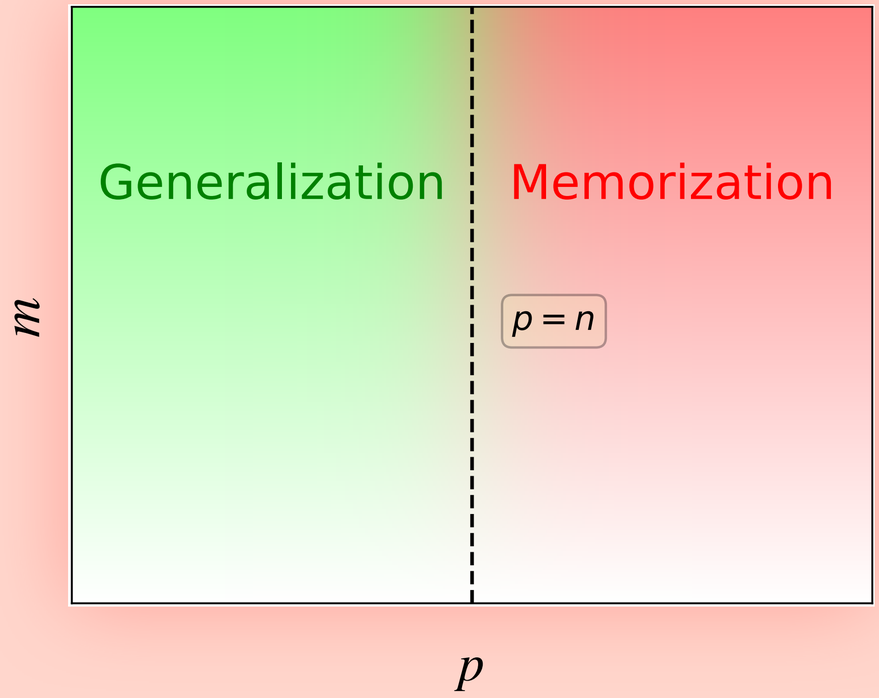


$m = 1$

Double-descent type learning curves

➢ Key take aways

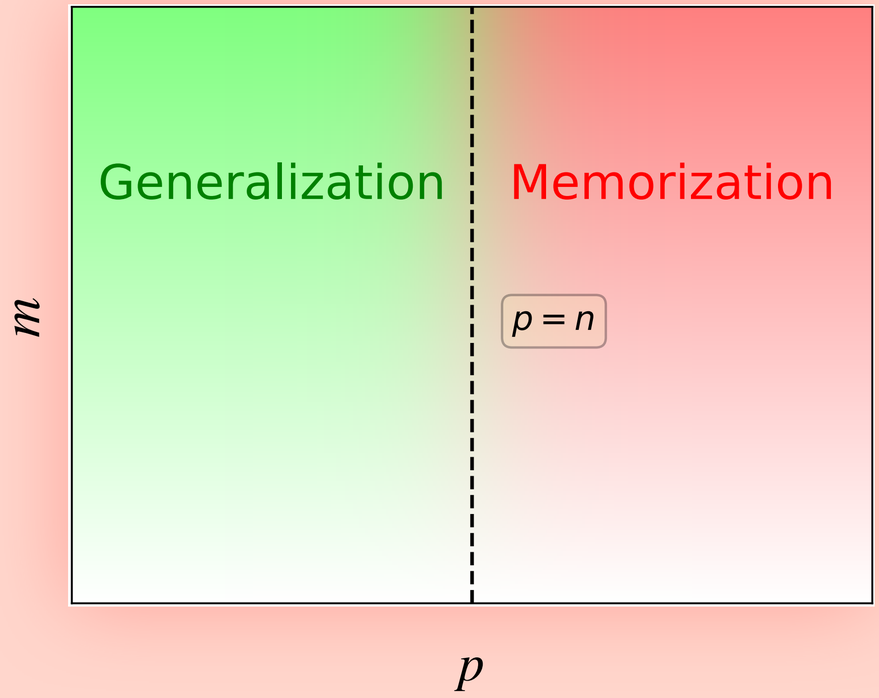
- Memorization increases as p increases, with a phase transition at $p = n$.



➢ Key take aways

- Memorization increases as p increases, with a phase transition at $p = n$.
- For $p \gg n$:

  Memorization increases as m increases.



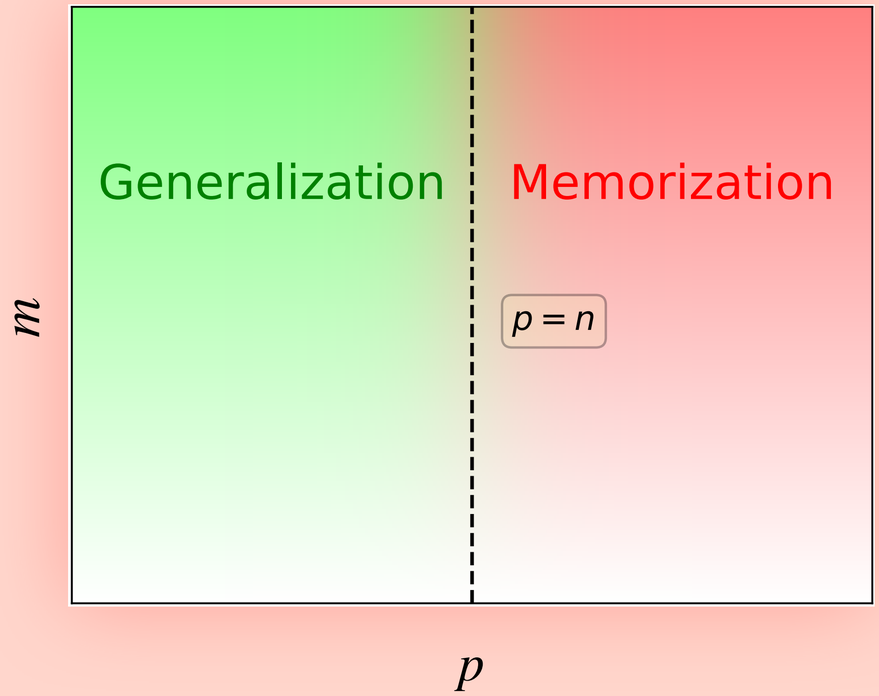
➢ Key take aways

- Memorization increases as p increases,

  with a phase transition at $p = n$.
- For $p \gg n$:

  Memorization increases as m increases.
- For $n \gg p$:

  Generalization increases as m increases.



Thank you!

Stop by our poster at location 186