What Drives the Inlier-Memorization Effect? A Theory of Outlier Detection via Early Training Dynamics

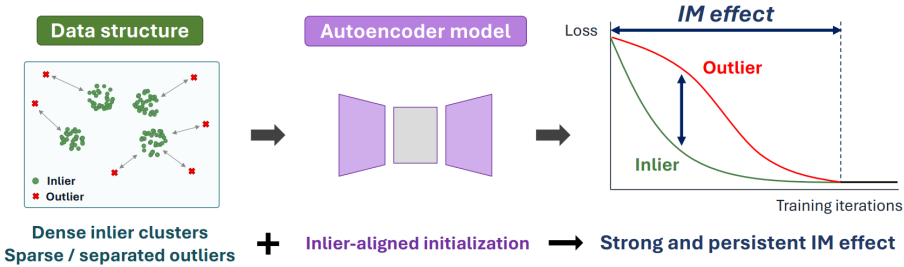
Anonymous Author(s)

Department of Statistics, Seoul National University
Presented by Sangmoon Han

2026-06-24

Table of Contents

Overview



- Recent advances have leveraged the inlier-memorization (**IM**) effect, a phenomenon in which deep models memorize inlier patterns earlier than those of outliers, as a powerful signal for distinguishing outliers.
- Despite its empirical success, the theoretical understanding of the IM effect remains limited.
- In this paper, the authors present a theoretical study of the **IM** effect.
- The conditions under which the **IM** effect exhibits **persistence** and **strength** are characterized from two key perspectives: **the data distribution** and **weight initialization**.
  - **Persistence** : The range of training steps over which the effect holds.
  - **Strength** : Reflecting how many outliers can be separated from inliers.

Notations & Assumptions

- The authors assume the observed dataset $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$ has a clustered structure and is decomposed into three distinct factors.
- Given a cluster-assignment map $g : [n] \to [K]$ with K cluster centers $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$, each sample is decomposed as:

$$x_i = \widetilde{x}_i + \xi_i + o_i, \quad i \in [n] \tag{1}$$

  where:

  - $\widetilde{x}_i := \mu_{g(i)}$ is the associated cluster center.
  - $\xi_i \in \mathbb{R}^p$ represents within-cluster variation.
  - $o_i \in \mathbb{R}^p$ denotes an additional perturbation corresponding to outliers.
- For each cluster $k \in [K]$, they define the following:
  - $\Lambda_k := \{i \in [n] : g(i) = k\}$.
  - $n_{\min} := \min_{k \in [K]} n_k$ and $n_{\max} := \max_{k \in [K]} n_k$.
  - $\delta := \min_{a \neq b \in [K]} \|\mu_a - \mu_b\|_2$.
- Under this decomposition, they define x_i as an inlier when $o_i = 0$, and as an outlier otherwise.
  - $\mathcal{O} := \{i \in [n] : o_i \neq 0\}$.
  - The outlier rate of cluster k is defined as $\frac{|\mathcal{O} \cap \Lambda_k|}{n_k}$.

Model Setup

- They consider a simple autoencoder $F_W : \mathbb{R}^p \to \mathbb{R}^p$ defined as

$$F_W(x) := A\phi(Wx), \tag{2}$$

where $W \in \mathbb{R}^{H\times p}$ is the trainable encoder weight matrix, $A \in \mathbb{R}^{p\times H}$ is the fixed decoder weight matrix, and ϕ is an activation function applied elementwise.

- For theoretical simplicity, as mentioned above, they assume that **only the encoder weight matrix W is trained, while the decoder weight matrix A is fixed during optimization**.
- To learn the encoder parameters W, The author utilize the following mean squared error

$$\mathcal{L}(W) := \frac{1}{2}\sum_{i=1}^{n} \|x_i - F_W(x_i)\|_2^2 = \frac{1}{2}\|X - \mathcal{F}_W(X)\|_2^2 \tag{3}$$

where $X := \left[x_1^\top, \ldots, x_n^\top\right]^\top \in \mathbb{R}^{np}$ and $\mathcal{F}_W(X) := \left[F_W(x_1)^\top, \ldots, F_W(x_n)^\top\right]^\top \in \mathbb{R}^{np}$.

Training via Gradient Descent

- Let $w := \mathrm{vec}(W) \in \mathbb{R}^{Hp}$ denote the vectorized parameter.
- They optimize this objective (3) using Gradient Descent(**GD**) with learning rate $\eta > 0$.
- The GD algorithm at each iteration step $\tau \in \mathbb{Z}_{\geq 0}$ updates the parameter as $w_{\tau+1} = w_\tau - \eta \nabla_w \mathcal{L}(W_\tau)$, given an initial weight w_0.
- The update can be equivalently written as

$$w_{\tau+1} = w_\tau - \eta J(W_\tau)^\top r_\tau, \tag{4}$$

  where $J(W_\tau) := \frac{\partial}{\partial w} \mathcal{F}_{W_\tau}(X) \in \mathbb{R}^{np \times Hp}$ and $r_\tau := \mathcal{F}_{W_\tau}(X) - X \in \mathbb{R}^{np}$.
- Specially, they define the initialization error on cluster centers $\widetilde{X}$ as $\widetilde{R}_0 := \|\mathcal{F}_{W_0}(\widetilde{X}) - \widetilde{X}\|_2$. A smaller $\widetilde{R}_0$ indicates that the initial weights W_0 better capture the underlying inlier structure.

Assumptions on Data and Outliers

To establish our main results, they introduce the following assumptions regarding the data distribution and outliers:

- **Assumption A. 1 (Bounded Sample Norms):** $\|x_i\|_2 \leq 1, {}^{\forall} i \in [n]$.
- **Assumption 3.1 (Small within-cluster variation):**
  $\|\xi_i\|_2 \leq \varepsilon_c, {}^{\forall} i \in [n]$, for some constant $\varepsilon_c \geq 0$.
- **Assumption 3.2 (Bounded outlier rate):**
  The outlier rate is at most ρ for some constant $\rho > 0$, i.e., $\max_{k \in [K]} \frac{|\mathcal{O} \cap \Lambda_k|}{n_k} \leq \rho$.
- **Assumption 3.3 (Well-separated outlier):**
  Let x_i be an outlier with perturbation o_i. We say that x_i is well-separated if $\|o_i\|_2 > \delta + 2\varepsilon_c$.

Here, the quantity $\delta + 2\varepsilon_c$ corresponds to an upper bound on the distance between inliers from the two closest clusters. Thus, a **well-separated** outlier lies sufficiently far from all inliers.

Assumptions on Data and Outliers

- Based on these assumptions, main theorem characterizes the **IM effect** using two critical iterations, T_1 and T_2.
- These iterations depend on key factors: $n_{\min}$, $n_{\max}$, ε_c, $\widetilde{R}_0$, ρ, and δ.
- For simplicity, They introduce a specific notation to denote these dependencies:

$$c = c(a_1 \uparrow, \ldots, a_m \downarrow)$$

This indicates that the constant c depends on the parameters $a_1, \ldots, a_m$. The arrows $\uparrow$ and $\downarrow$ explicitly denote that c is monotonically **increasing** or **decreasing** with respect to the corresponding argument, respectively.

Main Result: IM Effect

Theorem (IM Effect)

Assume that the training data $\{x_i\}_{i=1}^n$*, containing both inliers and outliers, satisfy* **Assumptions 3.1 and 3.2** *as well as regularity conditions in* **Appendix A**.
Suppose that the autoencoder in **(2)** *is trained by minimizing the reconstruction error in* **(3)** *via GD, updating the encoder parameter* W *according to* **(4)**. *Then, there exist a learning rate* $\eta > 0$ *and two integers* $0 < T_1 < T_2$ *with*

$$T_1 = c(n_{\min} \downarrow) \quad \textit{and} \quad T_2 = c(n_{\min} \uparrow, n_{\max} \downarrow, \varepsilon_c \downarrow, \widetilde{R}_0 \downarrow, \rho \downarrow, \delta \uparrow) \tag{5}$$

such that for any iteration $\tau \in [T_1, T_2]$*, inliers have smaller reconstruction loss than well-separated outliers. In particular, for any outlier* x_i *satisfying* **Assumption 3.3**, *we have*

$$\max_{j \in \mathcal{O}^c} \|x_j - F_{W_\tau}(x_j)\|_2^2 < \|x_i - F_{W_\tau}(x_i)\|_2^2. \tag{6}$$

Corollary: Perfect Separation

Corollary (Perfect separation via the IM effect)

Under the assumptions of **Theorem 3.1** *and with the interval* $[T_1, T_2]$ *defined therein, suppose that* **Assumption 3.3** *holds for all* $i \in \mathcal{O}$.

Then for any iteration $\tau \in [T_1, T_2]$, *the reconstruction error of every outlier is strictly larger than that of every inlier, i.e.,*

$$\max_{j \in \mathcal{O}^c} \|x_j - F_{W_\tau}(x_j)\|_2^2 < \min_{i \in \mathcal{O}} \|x_i - F_{W_\tau}(x_i)\|_2^2.$$

Simulation Study: Experimental Setup

- **Experiment Setting**
  - $n = 2,000$, $p = 2$, $H = 32$, $K = 3$.
  - $\mu_k \in [-10, 10]^2$, separation $\delta = 16$.
  - $n_{\min} = 666$, $n_{\max} = 667$.
  - $\rho = 0.05$ (100 samples), Uniform in $[-18, 18]^2$, minimum distance 10 from all μ_k.
  - All features min-max scaled to $[0, 1/\sqrt{2}]$ ($\|x_i\|_2 \leq 1$).
  - $\phi(x) = \tanh(x)$ activation, trained via Adam ($\eta = 2 \times 10^{-4}$). Kaiming uniform initialization.
  - A batch size of 500 for $1,000$ epoch.

- **Evaluation Metrics**
  - **AUROC(τ)**: At epoch τ, compute the outlier score $\|x_i - F_{W_\tau}(x_i)\|_2^2$ and evaluate **AUROC**.
  - **IM Window:** The maximal contiguous interval of epochs where **AUROC(τ)** ≥ 0.9, serving as an empirical approximation of $[T_1, T_2]$.

- They conduct a simulation study focusing on four key factors affecting the **IM** effect:
  - Within-cluster variation: ε_c
  - Cluster-size balance: $n_{\min}$ and $n_{\max}$
  - Outlier rate: ρ
  - Initialization quality: $\widetilde{R}_0$

Experiment

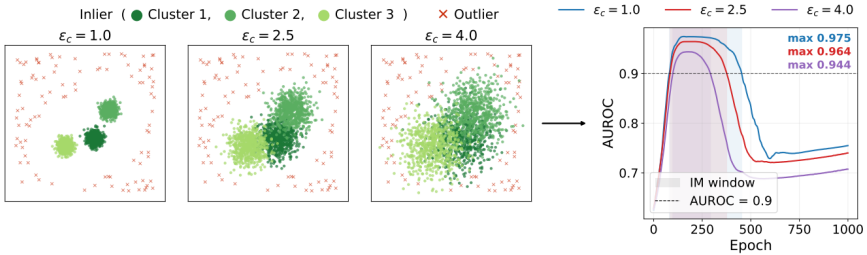


Figure 2: Effect of within-cluster variation ε_c. (Left) Data distributions for $\varepsilon_c \in \{1.0, 2.5, 4.0\}$. (Right) AUROC trajectories; shaded regions indicate the IM window.

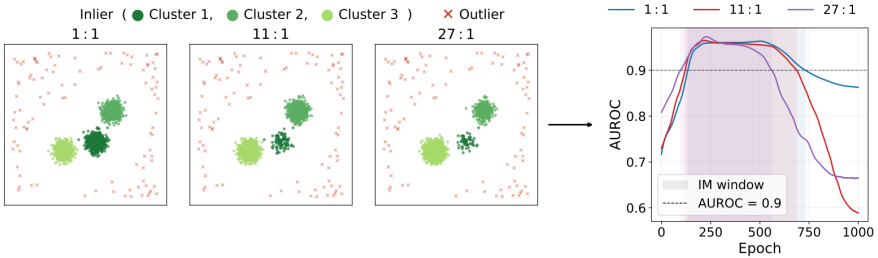


Figure 3: Effect of cluster-size balance $n_{\max} : n_{\min}$. (Left) Data distributions for $(n_{\max} : n_{\min}) \in \{(1:1), (11:1), (27:1)\}$. (Right) AUROC trajectories; shaded regions indicate the IM window.

Experiment

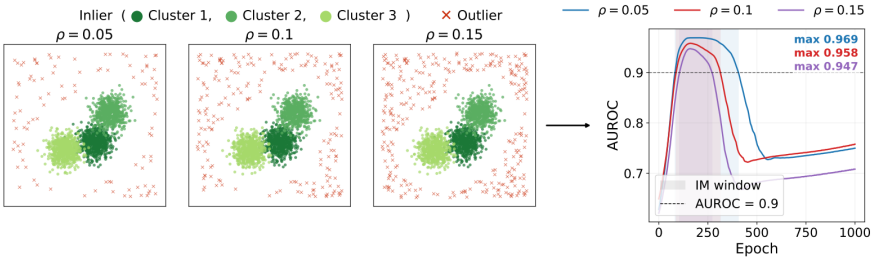


Figure 4: Effect of outlier rate ρ. (Left) Data distributions for $\rho \in \{0.03, 0.10, 0.25\}$. (Right) AUROC trajectories; shaded regions indicate the IM window.

Table 1: Effect of initialization error R_0 on the IM window and detection performance (best AUROC).

$\widetilde{R}_0$	T_1	T_2	IM window length $(T_2 - T_1)$	Best AUROC
0.40	37	509	472	0.961
0.55	56	433	377	0.962
0.60	94	456	362	0.964

Experiment

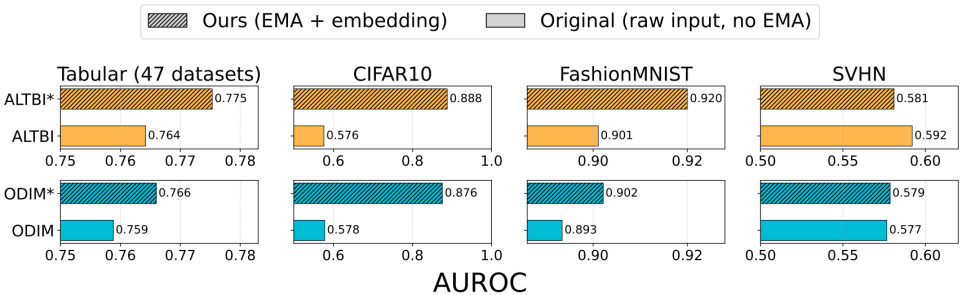


Figure 5: Average over 47 tabular datasets (leftmost) and results on `CIFAR10`, `FashionMNIST`, and `SVHN`. Hatched bars (ALTBI* and ODIM*) are the proposed variants combining TabPFN/ViT embeddings with EMA warm-up, while the plain bars (ALTBI and ODIM) correspond to training on raw inputs without EMA.

Appendix A: Regularity Conditions

- **Assumption A. 4 (Cluster-Center Jacobian):**
  Fix an initialization $W_0 \in \mathbb{R}^{H \times p}$ and a radius $R_{\text{loc}} > 0$. Define

$$\mathcal{D}(W_0, R_{\text{loc}}) := \left\{ W \in \mathbb{R}^{H \times p} : \|W - W_0\|_F \leq R_{\text{loc}} \right\}.$$

  Assume that there exist constants $\alpha_c, \beta_c, L_c > 0$ such that, for every $W \in \mathcal{D}(W_0, R_{\text{loc}})$,
  - For every unit vector $z \in \mathbb{R}^{Kp}$, $\alpha_c \leq \left\| J_c(W)^\top z \right\|_2 \leq \beta_c$
  - For all $W_1, W_2 \in \mathcal{D}(W_0, R_{\text{loc}})$, $\|J_c(W_1) - J_c(W_2)\| \leq L_c \|W_1 - W_2\|_F$.
- **Assumption A. 6 (Activation bounds):**
  Assume that $|\phi'(t)| \leq B_1$ and $|\phi''(t)| \leq B_2$ for all $t \in \mathbb{R}$.

Return to Main Theorem

Appendix A: Other Assumptions.

- **Assumption A. 1 (Bounded Sample Norms):** $\|x_i\|_2 \leq 1, {}^{\forall} i \in [n]$.
- **Assumption A. 2 (Bounded Outlier Magnitude):** $\|o_i\|_2 \leq \varepsilon_o$ for all $i \in \mathcal{O}$.
- **Assumption A. 3 (No Within-Cluster Variation):** $\xi_i = 0$ for all $i = 1, \ldots, n$.