The Effect of Optimal Self-Distillation in Noisy Gaussian Mixture Model

Kaito Takanami[1,2] Takashi Takahashi[1,3] Ayaka Sakata[2,3]

[1]The University of Tokyo [2]Ochanomizu University [3]RIKEN

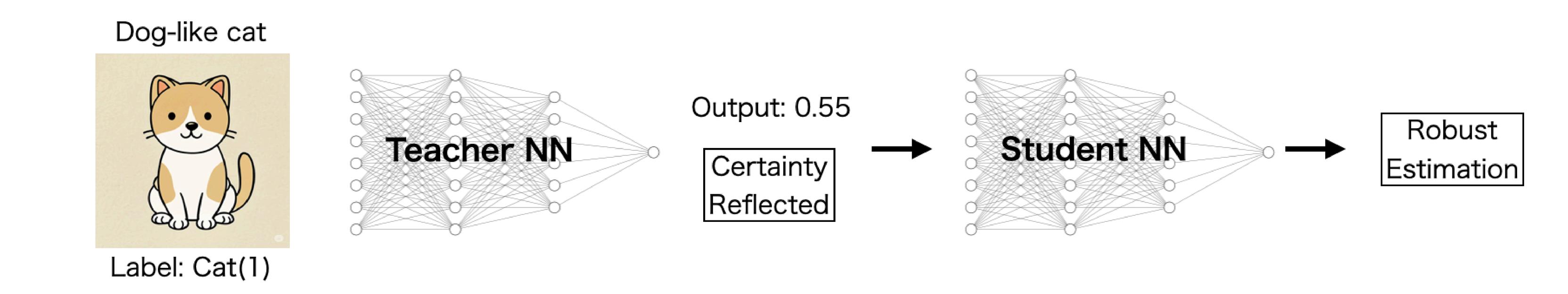
What is self-distillation (SD)?

- **Knowledge Distillation**: A large model teaches a small one for compression. (Teacher size > Student size)),
- **Self Distillation (SD)**: A model teaches itself for a performance boost. (Teacher size = Student size),

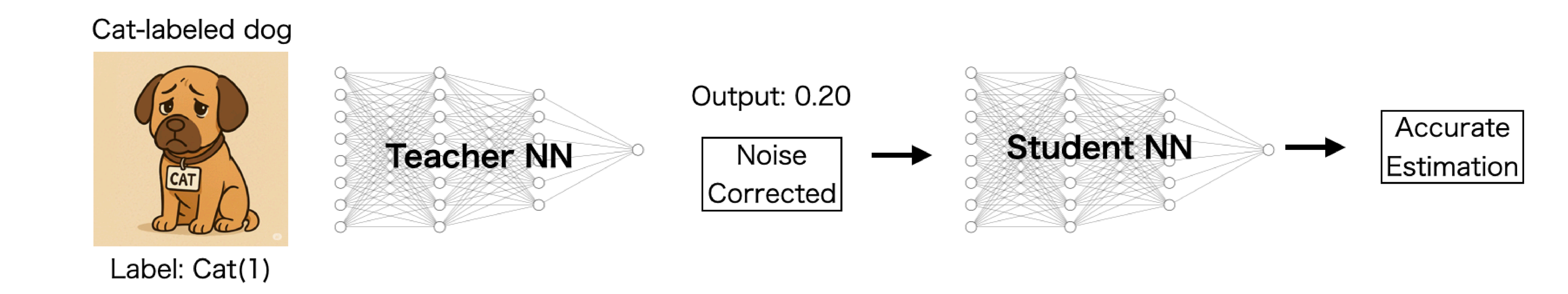
Why successful? (in classification)

- The model's output probabilities (**dark knowledge**) carry information beyond hard labels.
- SD can also remove noise from training labels (**denoising**).

Dark knowledge effect [Hinton et al., 2015]



Denoising effect [Das et al., 2023]



Question

- How do dark knowledge/denoising contribute to performance gains, and by how much?
- What improvements can multi-stage SD achieve when its hyperparameters are optimally tuned?
- What is the optimal update steps?
- What is the effect of bias learning in class-imbalanced dataset?

Approach

Statistical analysis based on inequality bounds offers only coarse insight into hyperparameter effects.
→ High-dimensional solvable models provide **precise** evaluation (exact expectaion behavior for hyperparameters).

Model

Binary classification of Gaussian mixture data with noisy labels using a single-layer neural network:

Data generation

$$\begin{cases} \alpha = (\#\text{ of Data}(=M))/(\text{Dimension}(=N)) \\ y_\mu^{\text{true}} \in \{0,1\}, \quad P(y_\mu^{\text{true}}=1)=\rho, \quad P(y_\mu \neq y_\mu^{\text{true}})=\theta \\ \boldsymbol{x}_\mu \sim (2y_\mu^{\text{true}}-1)\boldsymbol{v}/\sqrt{N} + \sqrt{\Delta}\mathcal{N}(0,1) \end{cases}$$

We learn from noisy labels $\mathcal{D} = \{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^M$.

Multi-stage SD

The learning process at each stage $t = 0, 1, \ldots, T$ is given by

$$\hat{\boldsymbol{w}}^t, \hat{B}^t = \underset{\boldsymbol{w}^t, B^t}{\operatorname{argmin}} \left[\sum_{\mu=1}^M \ell\left(y_\mu^t, \sigma\left(\frac{\boldsymbol{w}^t \cdot \boldsymbol{x}_\mu}{\sqrt{N}} + B^t \right) \right) + \frac{\lambda^t}{2} \|\boldsymbol{w}^t\|^2 \right],$$

where y_μ^t is the label we use at stage t:

$$\begin{cases} y_\mu^0 = y_\mu \ (\text{observed label}) \\ y_\mu^t = \sigma\left(\beta^t\left((\hat{\boldsymbol{w}}^{t-1}\cdot\boldsymbol{x}_\mu)/\sqrt{N} + \hat{B}^{t-1}\right)\right)(t \geq 1) \ (\text{pseudo-label}) \end{cases}$$

Evaluation of the optimal multi-stage SD

- **Error metric**: $\mathcal{E}^t =$ (data-averaged 0-1 generalization error)
- **Hyperparameters**: $\lambda^t (t = 0, 1, \ldots)$ and $\beta^t (t = 1, 2, \ldots)$
- **Optimal SD**: $\mathcal{E}^{t*} = (\mathcal{E}^t$ at the optimal hyperparameters)
- **Optimal SD effect**: $\mathcal{E}^{0*} - \mathcal{E}^{t*}$

Technical Details: Replica method for dynamics

The dynamics of $\varphi^t = (\hat{\boldsymbol{w}}^t, \hat{B}^t)(t = 0, 1, \ldots)$ is given by

$$p(\varphi^T, \varphi^{T-1}, \ldots, \varphi^1 \mid \varphi^0, \mathcal{D}) = \lim_{\beta\to\infty} \prod_{t=1}^T \frac{\exp(-\beta \mathcal{L}(\varphi^t \mid \varphi^{t-1}, \mathcal{D}))}{Z^t(\varphi^{t-1}, \mathcal{D})},$$

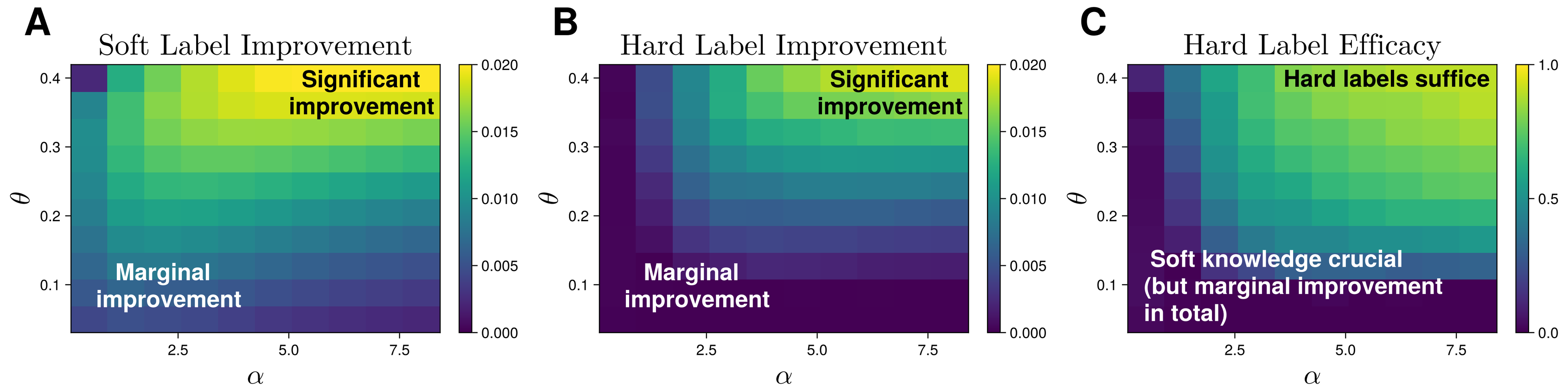
where the denominator makes data averaging difficult.

Multi-Stage Replica Method

Replicating partition function $1/(Z^t(\varphi^{t-1}, \mathcal{D})) = \lim_{n^t\to 0} (Z^t(\varphi^{t-1}, \mathcal{D}))^{n^t-1}$ for all t eliminates the averaging issue.

Results

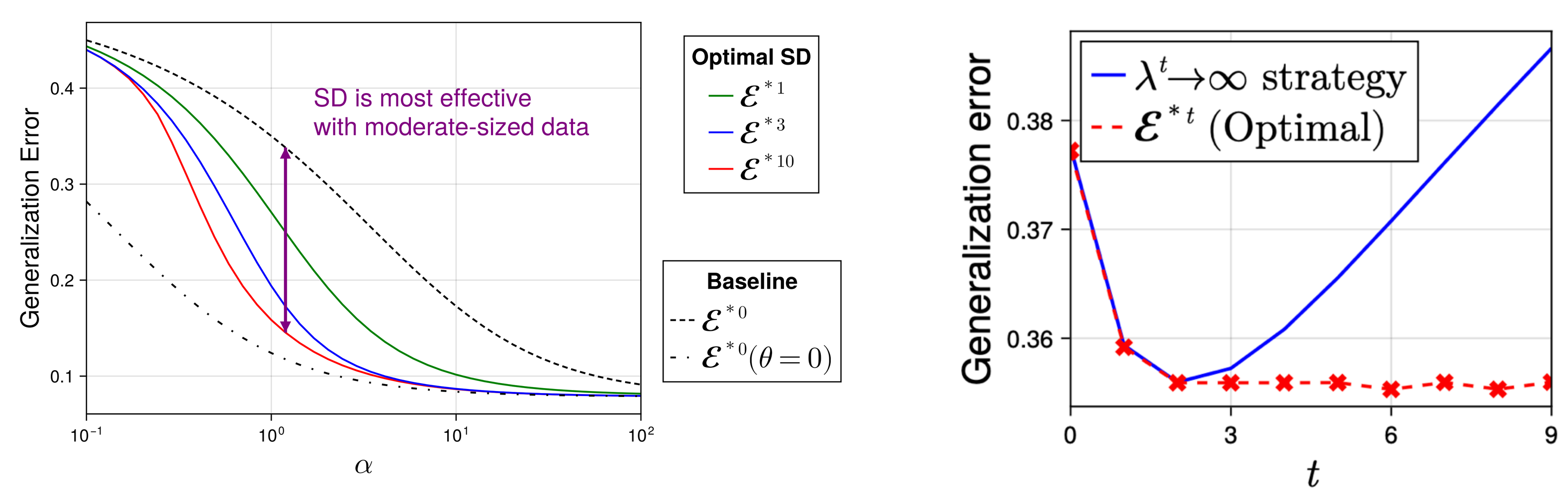
The Role of Soft Labels in SD (T=1)



Takeaways

Hard label improvement ≈ Soft error improvement
→ **Dark knowledge effect is limited**, denoising is the key factor for SD.

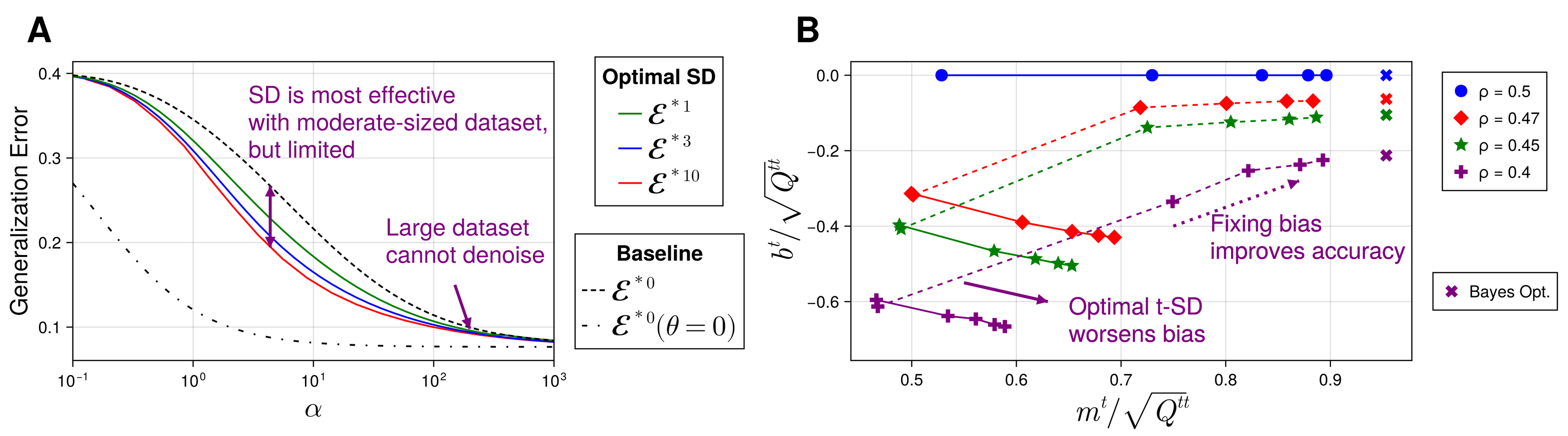
The Effect of Multi-Stage SD in Balanced Data



Takeaways

- The benefit of SD peaks at intermediate dataset size.
- Naive SD can perform as poorly as random guessing.
  → **Early stopping** heuristic matches optimal SD.

Hardness of learning bias



Takeaways

aligning both the bias and the decision-boundary orientation at once is hard under imbalanced dataset
→ **Bias fixing** heuristic approaches Bayes opt.

The study was conducted as part of the exploratory project "Mathematical Exploration of Universal Structures in Multicomponent and Polydisperse Systems", supported by the Toyota Konpon Research Institute, Inc. This work was also supported by JSPS KAKENHI Grant Numbers 22H05117 and 23K16960, JST ACT-X Grant Number JPMJAX24CG and JST BOOST NAIS Grant Number JPMJBS2418.