Inconsistency-Aware Minimization

Improving Generalization with Unlabeled Data

Hee-Sung Kim Hyeonseong Kim Sungyoon Lee

ICML 2026

Department of Computer Science
Hanyang University, Seoul, Korea

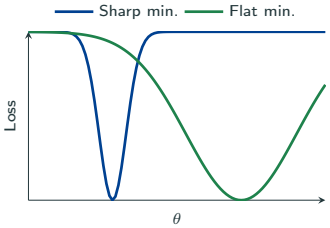
Motivation: Improving Generalization without Labels

The core question

- Can we **optimize** for better generalization?
- Can we do this **without labels?**

Existing approaches & limits

- **Sharpness** (SAM): needs label
- **Disagreement**: needs **two models** and is non-differentiable
- **Inconsistency**: needs **multiple models**



Optimize toward flat minima without labels.

Our goal: a measure that is

✓ **single-model**
✓ **label-free**
✓ **differentiable**

Local Inconsistency $S_\rho(\theta)$

Definition — worst-case output shift under parameter perturbation:

$$S_\rho(\theta) \;=\; \max_{\|\delta\| \le \rho} \; \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \Big[\mathrm{KL}\big(f(\boldsymbol{x}; \theta) \,\big\|\, f(\boldsymbol{x}; \theta + \delta) \big) \Big]$$

Reading the formula

- Sensitivity of **output distribution** to a small parameter perturbation
- High $S_\rho(\theta)$ $\Rightarrow$ outputs change a lot in the ball $\Rightarrow$ likely poor generalization
- Computed from **unlabeled data only** — uses $\boldsymbol{x}$, not labels

Theory: Local Inconsistency → FIM → Hessian

Quadratic approximation of the KL divergence:

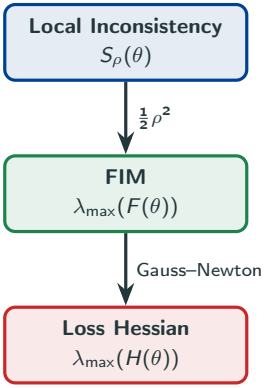
$$\mathbb{E}_{\boldsymbol{x}}\big[\mathrm{KL}(f(\boldsymbol{x};\theta)\|f(\boldsymbol{x};\theta+\delta))\big] \approx \tfrac{1}{2}\delta^{\top} F(\theta)\,\delta$$

$\Rightarrow$ Maximizing over $\|\delta\| \le \rho$:

$$S_\rho(\theta) \;\approx\; \tfrac{1}{2}\rho^2\,\lambda_{\max}\big(F(\theta)\big)$$

Bridge to the loss Hessian

- For CE loss: $H(\theta) \approx F(\theta)$ (Gauss–Newton)
- Captures **curvature** of loss landscape at θ



Why Local Inconsistency? Practical Comparison

- Sharpness : $\max_{\|\delta\| \le \rho} CE(f_{\theta+\delta}(x), y)$
- Disagreement : $\mathbb{E}_x[1(\hat{y}_\theta(x) \neq \hat{y}_{\theta'}(x))]$
- Inconsistency : $\mathbb{E}_{Z_n}\mathbb{E}_{\theta,\theta' \sim \Theta_{P|Z_n}}\mathbb{E}_{x \sim Z_n}[\mathrm{KL}(f(x;\theta)\|f(x;\theta'))]$.

Measure	Sharpness Keskar'17, SAM	Disagreement Jiang'22	Inconsistency Johnson'23	Local Inconsistency (Ours)
Single model	✓	×	×	✓
Label-free	×	✓	✓	✓
Differentiable	✓	×	✓	✓

Local Inconsistency uniquely combines all three practical advantages,
enabling its use as a **regularizer** during training — even **without labels.**

Inconsistency-Aware Minimization (IAM)

Two strategies to use S_ρ in training:

IAM-D — Direct regularization

$$L_{\text{IAM-D}}(\theta) = L(\theta) + \beta\, S_\rho(\theta)$$

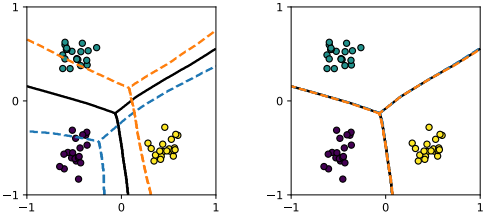
IAM-S — SAM-like perturbation

$$L_{\text{IAM-S}}(\theta) = L(\theta + \delta^*)$$

where $\delta^* = \text{argmax}_{\|\delta\| \le \rho} \mathbb{E}_{\boldsymbol{x}}[\text{KL}]$

Key contrast vs. SAM

- SAM: perturbation $\propto \nabla_\theta L(y, f_\theta(x))$ (labels)
- IAM: perturbation $\propto \nabla_\delta \text{KL}$ (no labels)



(a) $f(x; \theta + \delta_1)$ **(b)** $f(x; \theta + \varepsilon)$

Figure 1: A synthetic classification example. The black, blue, orange lines correspond to decision boundaries of the NN with trained parameter values, and parameter values perturbed by $\pm\delta_1$ or random direction ε.

Results — Supervised Learning

Table 1: Test error (mean ± stderr) with supervised setting

	ImageNet	CIFAR-10	CIFAR-100
SGD	22.66	3.68	19.17
SAM	21.80	3.31	17.63
IAM-D	**21.36**	**3.28**	17.16
IAM-S	21.72	**3.28**	**16.82**

Training dynamics — CIFAR-100

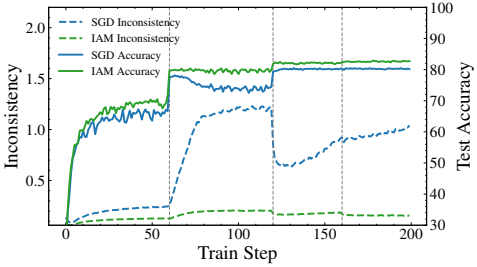


Figure 2: IAM keeps S_ρ **low** after LR decay and shows less overfitting.

Semi-Supervised Learning — IAM-D as a Plug-in Regularizer

Table 2: Test error (mean ± stderr) with semi-supervised setting on CIFAR-10 and CIFAR-100

	CIFAR-10		CIFAR-100	
labels	250	4000	2500	10000
SGD	63.82	22.45	68.91	45.94
SAM	63.91	19.95	69.53	43.30
IAM-D	**61.77**	**15.07**	**66.98**	**40.02**
FixMatch	6.26	4.10	32.84	22.93
+ **IAM-D**	**5.30**	**3.88**	**28.95**	**21.99**

Why does SAM fail on FixMatch?

SAM (label-only)
flatness over **tiny labeled set**
$\Rightarrow$ may misleading curvature

IAM-D (label-free)
S_ρ over **all unlabeled data**
$\Rightarrow$ true geometric flatness

CIFAR-10, 250 labels:

FixMatch + SAM: **9.90%**

FixMatch + **IAM-D:** **5.30%**

Self-Supervised Learning — SimCLR + IAM-D

Setup

- Local inconsistency computed on **projection-head outputs**

Observations

- **Higher** linear-probe accuracy
- **Faster** convergence

Controlling S_ρ helps even **without explicit labels**.

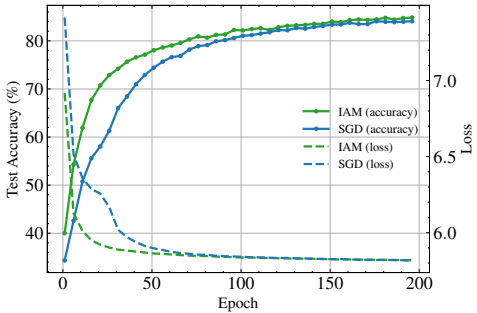


Figure 3: Linear-probe accuracy & SimCLR loss: IAM-D vs. SGD (ResNet-18, CIFAR-10).

Conclusion

Contributions

- Introduced **Local Inconsistency** $S_\rho(\theta)$
  a **single-model**, **label-free**, **differentiable** generalization measure.
- Theoretical bridge: $S_\rho \rightarrow$ FIM $\rightarrow$ Hessian; FIM-based generalization bound.
- Proposed **IAM-D / IAM-S**
  competitive with SAM/ASAM in supervised, and uniquely **useful in Semi-/Self-SL**.

Take-home

Controlling **local inconsistency** from **unlabeled data** alone improves generalization.

Questions?

Hee-Sung Kim Hyeonseong Kim Sungyoon Lee

muwonijr@hanyang.ac.kr