Inconsistency-Aware Minimization: Improving Generalization with Unlabeled Data

Hee-Sung Kim Hyeonseong Kim Sungyoon Lee
Department of Computer Science, Hanyang University, Seoul, Korea | ICML 2026 | muwonijr@hanyang.ac.kr

ICML International Conference On Machine Learning

Can We Regularize Models for Better Generalization Using Unlabeled Data?

Flatness-aware methods such as SAM can improve generalization. However, existing generalization measures are difficult to use when labels are limited:

- **Sharpness** (SAM): relies on a label-dependent loss.
- **Disagreement**: requires two models and is non-differentiable.
- **Inconsistency**: need multiple models on different data splits.

Goal: a practical regularizer that is **single-model**, **label-free**, and **differentiable**.

Local Inconsistency $S_\rho(\theta)$

How much can a model's predictions change under the worst small perturbation of its parameters?

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

- **High** S_ρ means small parameter changes cause large output shifts.
- It is computed from **inputs $\boldsymbol{x}$ only** — **no labels required**.
- It is **differentiable**, so we can use it as a training regularizer.

Measure	Single model	Label free	Diff. able
Sharpness (SAM) $\max_{\|\delta\|\le\rho} CE\big(f_{\theta+\delta}(x), y\big)$	✓	×	✓
Disagreement (Jiang'22) $\mathbb{E}_x[1(\hat{y}_\theta(x) \neq \hat{y}_{\theta'}(x))]$	×	✓	×
Inconsistency (Johnson'23) $\mathbb{E}_{Z_n}\mathbb{E}_{\theta,\theta'}\mathbb{E}_x\,[\mathrm{KL}(f(x;\theta)\,\|\,f(x;\theta'))]$	×	✓	✓
Local Inconsistency (Ours)	✓	✓	✓

Practical Unlabeled Regularization.

Local Inconsistency is a **single-model, label-free, and differentiable** measure of worst-case shifts in a model's output distribution under local parameter perturbations. **IAM** incorporates this label-free signal into training to improve generalization.

Inconsistency-Aware Minimization (IAM)

IAM-D — Direct regularization. Add S_ρ to the loss:

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

IAM-S — SAM-like perturbation.
Train at the most-inconsistent point:

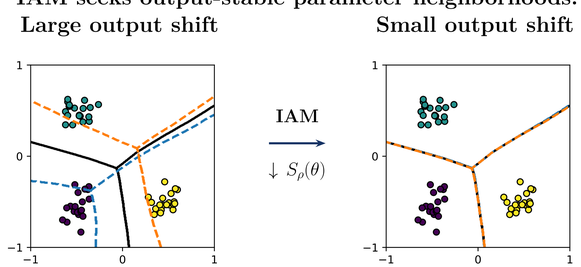
$$L_{\text{IAM-S}}(\theta) \;=\; L(\theta+\delta^*)$$

where $\delta^* = \arg\max_{\|\delta\|\le\rho} \widehat{\mathrm{KL}}_{\mathcal{B}}\big(f_\theta\,\|\,f_{\theta+\delta}\big)$.

Both methods use an inner maximization, but optimize different signals:

- **SAM:** $\delta_{\text{SAM}} \propto \nabla_\theta \widehat{L}_{\text{task}}(\theta)$ — **task-loss signal**
- **IAM:** $\delta_{\text{IAM}} \propto \nabla_\delta \widehat{\mathbb{E}}_{\boldsymbol{x}}\mathrm{KL}(f_\theta\,\|\,f_{\theta+\delta})$ — **unlabeled inputs**

IAM seeks output-stable parameter neighborhoods.



The two panels use perturbations of the same magnitude. Here, decision-boundary movement visualizes the underlying change in the predictive distribution. IAM prefers the stable neighborhood by minimizing the worst-case output shift $S_\rho(\theta)$ using unlabeled $\boldsymbol{x}$ alone.

Why It Works: Local Inconsistency → FIM → Loss Hessian

A 2nd-order expansion of the KL is a **Fisher Information Matrix (FIM)** quadratic, so the worst-case shift is its **top eigenvalue**:

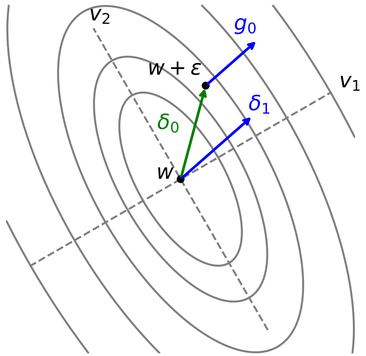
$$S_\rho(\theta) \;\approx\; \tfrac{1}{2}\rho^2\,\lambda_{\max}\big(F(\theta)\big) \qquad \delta^* = \rho\,\boldsymbol{v}_{\max}.$$

For cross-entropy, $H(\theta) \approx G(\theta) = F(\theta)$ (Gauss–Newton)
⇒ **controlling S_ρ can suppress the top loss curvature** — using only unlabeled $\boldsymbol{x}$.

Estimating S_ρ (label-free). K-step normalized ascent:

$$\delta_{k+1} = \rho\,\frac{g_k}{\|g_k\|}, \qquad g_k = \nabla_\delta\,\mathrm{KL}\big(f_\theta\,\|\,f_{\theta+\delta}\big).$$

These steps **approximate power iteration** on $F(\theta)$; K=1 gives **SAM-like computational cost**.



Supervised: IAM matches / beats SAM

Test error (%):

	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**

Without ever using labels in S_ρ, IAM is on par with or better than SAM — IAM-S beats SAM by **0.81%** on CIFAR-100. In training, IAM-D **suppresses the rise of S_ρ after the LR decay**, exactly where SGD overfits and its inconsistency rebounds.

Semi-Supervised: a regime SAM cannot handle

IAM-D plugs into FixMatch — CIFAR-10, 250 labels (test error %):

Method	Error (%)
FixMatch	6.26
FixMatch + SAM	**9.90** (worse)
FixMatch + IAM-D	**5.30** (best)

SAM measures flatness over the *tiny* labeled set ⇒ misleading curvature, **hurting** FixMatch (6.26 → 9.90). **IAM-D** computes S_ρ over *all* unlabeled data ⇒ true flatness and a new **best (5.30)**. This is a regime that SAM cannot structurally reach.

Self-Supervised: it works here too

Computing S_ρ on **SimCLR projection-head outputs** needs **no labels at all**. Plugging IAM-D into SimCLR (ResNet-18 / CIFAR-10):

- **Higher** linear-probe accuracy on the downstream task.
- **Faster** convergence — it also drives the SimCLR loss down more quickly.

Controlling output sensitivity helps representation learning *itself*, even with no supervision.

Take-Home

Yes — we can get SAM-like generalization gains from *unlabeled* data.

Local inconsistency is a single-model, label-free, differentiable measure; minimizing it (IAM) steers training toward flat minima **without labels**, matching SAM in supervised learning and surpassing it wherever labels are scarce or absent.