FAM: Relative Flatness Aware Minimization

Linara Adilova[1] linara.adilova@rub.de
Amr Abourayya[1,2] amr.abourayya@uk-essen.de
Jianning Li[2] jianning.li@uk-essen.de
Amin Dada[2] amin.dada@uk-essen.de
Henning Petzka[3] petzka@cs.rwth-aachen.de
Jan Egger[2,4] jan.egger@uk-essen.de
Jens Kleesiek[2] jens.kleesiek@uk-essen.de
Michael Kamp[1,2,5] michael.kamp@uk-essen.de

Relative Flatness

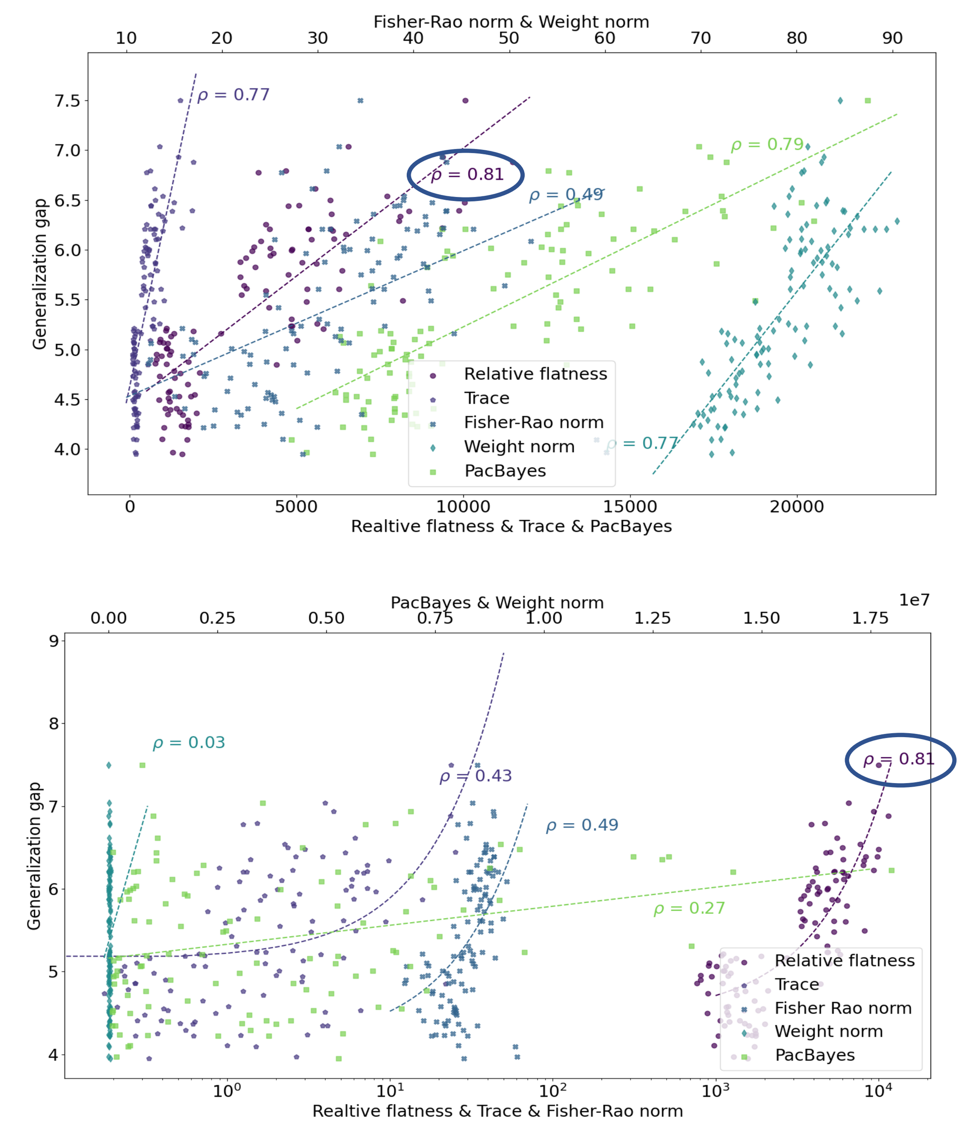
Taylor approximation on feature robustness derives a **reparameterization-invariant flatness measure** using the weights of one chosen layer

$$\kappa^{\phi}_{Tr}(\mathbf{w}) := \sum_{s,s'=1}^{d} \langle \mathbf{w}_s, \mathbf{w}_{s'} \rangle \cdot Tr(H_{s,s'}(\mathbf{w}, \phi(S)))$$

where $H_{s,s'}(\mathbf{w}, \phi(S)) = \left[\frac{\partial^2 \mathcal{E}_{emp}(\mathbf{w}, \phi(S))}{\partial w_{s,t} \partial w_{s',t'}} \right]_{1 \le t, t' \le m}$

$$\mathbf{w}_s = (w_{s,t})_t \in \mathbf{R}^{1 \times m}$$

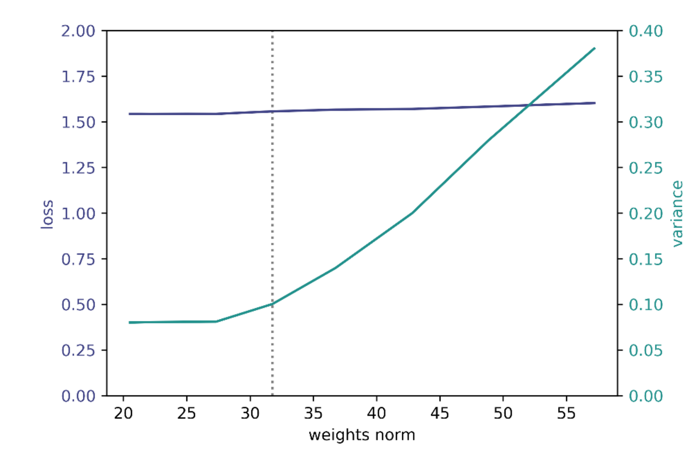
Comparing Flatness Measures (CIFAR-10)



- **Relative** flatness: the further away from origin, the flatter minimum must be for same generalization. For symmetries, this can be empirically observed.
- Good generalization can be achieved by **regularizing the geometry of the minimum**
- Regularizing via relative flatness provides **generalization bound**

$$\mathcal{E}_{gen}(\mathbf{W}, S) \lesssim |S|^{-\frac{2}{4+m}} \left((2m)^{-1} \kappa^{\phi}_{Tr}(\mathbf{w}^l) + C_1 + \frac{C2}{\sqrt{\delta}} \right)$$

Loss and width of minimum (estimated variance via SWAG) vs. distance to origin (weights norm)



Relative Flatness Aware Minimization

For a training set S and a differentiable loss function $\ell(S, \mathbf{W})$ the regularized objective is $\ell(S, \mathbf{W}) + \kappa^{\phi}_{Tr}(\mathbf{w}^l)$ for selected layer l.

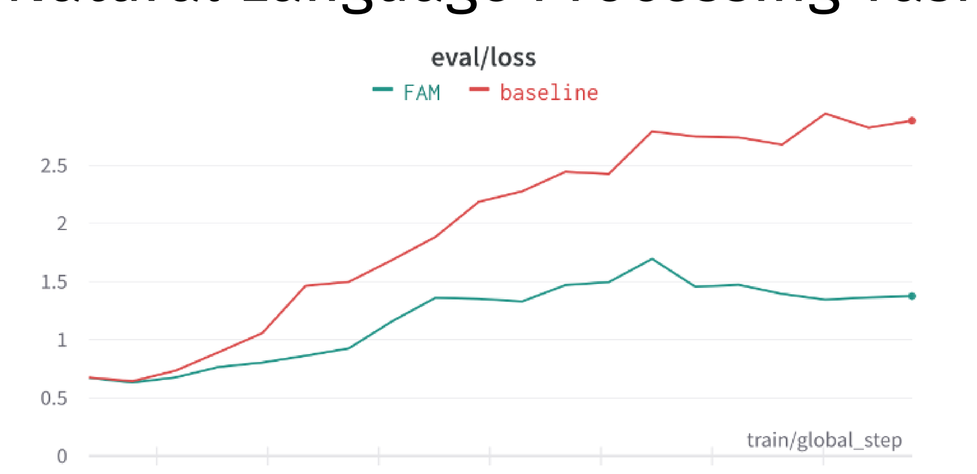
The overall complexity of computing the regularizer is in $\mathcal{O}\left(|\mathbf{W}| + d^2 m^2\right)$, where d, m is dimensionality of the layer l selected for regularization, i.e., $\mathbf{w}^l \in \mathbb{R}^{d \times m}$.

Evaluation

Image Classification Datasets

	Baseline	SAM	FAM
CIFAR10	95.53 ± 0.0001	95.61 ± 0.001	$\mathbf{95.62 \pm 0.002}$
CIFAR100	84.48 ± 0.12	85.72 ± 0.08	$\mathbf{87.2 \pm 0.05}$
SVHN	97.72 ± 0.02	$\mathbf{97.84 \pm 0.05}$	97.81 ± 0.07
FashionMNIST	94.57 ± 0.28	$\mathbf{94.99 \pm 0.02}$	94.6 ± 0.04

Natural Language Processing Task



Skull Reconstruction Problem

methods	evaluation set 1						evaluation set 2					
	DSC	DSC (100)	HD	HD (100)	$HD95$	$HD95$ (100)	DSC	DSC (100)	HD	HD (100)	$HD95$	$HD95$ (100)
baseline	0.6464	0.6569	7.0130	7.1787	2.0635	2.0422	0.6413	0.6489	7.1421	7.1939	2.0924	2.1371
FAM, $\lambda = 0.0006$	0.7155	0.6817	6.5531	6.7772	1.8202	**1.8281**	0.7156	0.6762	6.5542	7.0115	1.8178	1.9088
FAM, $\lambda = 0.002$	0.7173	**0.7175**	**6.4813**	6.5478	**1.8175**	**1.8281**	0.7175	**0.7176**	**6.4813**	6.5478	**1.8148**	**1.8281**
FAM, $\lambda = 0.02$	**0.7176**	0.7168	6.5221	**6.5271**	1.8210	1.8344	**0.7176**	0.7168	6.5221	**6.5271**	1.8210	1.8344
FAM, $\lambda = 0.1$	**0.7176**	0.7169	6.5085	**6.5222**	1.8210	1.8345	**0.7176**	0.7169	6.5085	**6.5222**	1.8210	1.8345
FAM, $\lambda = 0.7$	**0.7177**	0.7169	6.5202	6.5389	1.8210	1.8359	**0.7177**	0.7169	6.5202	6.5389	1.8210	1.8359

[1] Ruhr University Bochum, Germany
[2] Institute for AI in Medicine (IKIM), Germany
[3] RTWH Aachen, Germany
[4] TU Graz, Austria
[5] Monash University, Australia