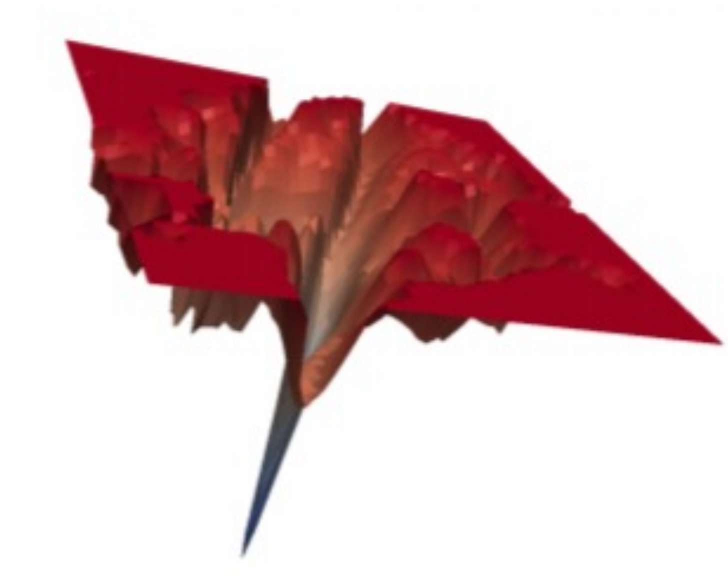
Surrogate Gap Guided Minimization Improves Sharpness-Aware Training

Juntang Zhuang[1], Boqing Gong[2], Liangzhe Yuan[2], Yin Cui[2], Hartwig Adam[2],
Nicha C. Dvornek[1], Sekhar Tatikonda[1], James S. Duncan[1], Ting Liu[2]
1: Yale University
2: Google Research

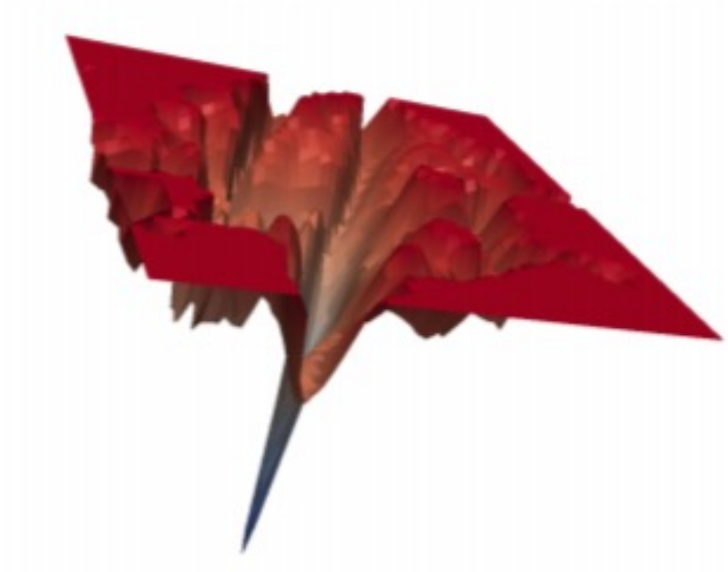
Generalization and curvature of loss surface

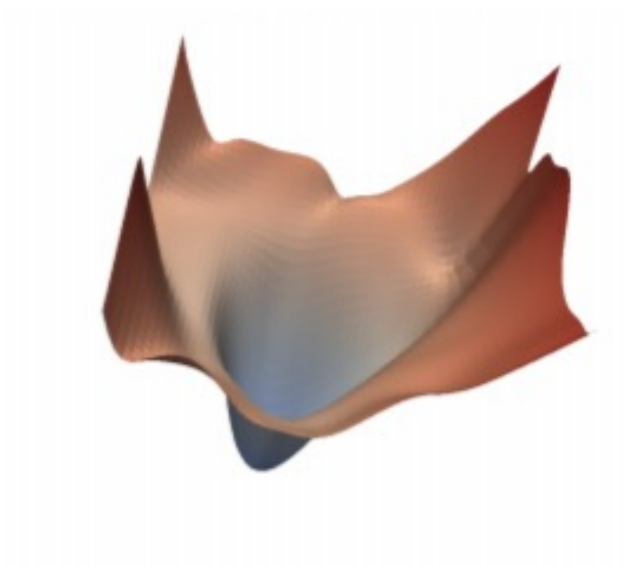


- Minimizing the training loss → sharp local minima (left) → poor generalization

[1] Chaudhari, Pratik, et al. "Entropy-sgd: Biasing gradient descent into wide valleys." *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019): 124018.
[2] Jiang, Yiding, et al. "Fantastic generalization measures and where to find them." *arXiv preprint arXiv:1912.02178* (2019).
[3] Niladri S Chatterji, et al.The intriguing role of module criticality in the generalization of deep networks. ICLR 2020

Generalization and curvature of loss surface





- Minimizing the training loss → sharp local minima (left) → poor generalization

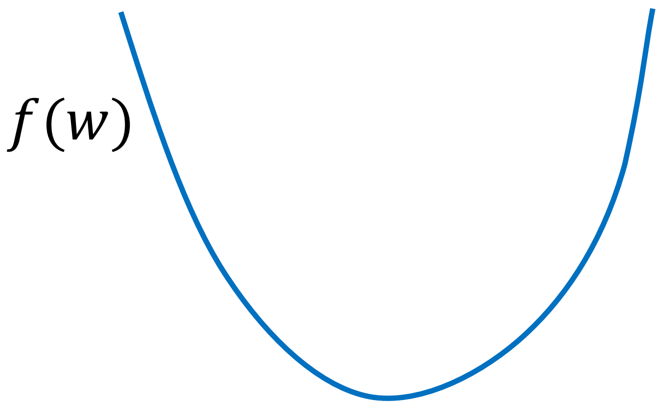
- Flatter local minima (right) → better generalization [1,2,3]

[1] Chaudhari, Pratik, et al. "Entropy-sgd: Biasing gradient descent into wide valleys." *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019): 124018.
[2] Jiang, Yiding, et al. "Fantastic generalization measures and where to find them." *arXiv preprint arXiv:1912.02178* (2019).
[3] Niladri S Chatterji, et al.The intriguing role of module criticality in the generalization of deep networks. ICLR 2020

Surrogate Gap: an equivalent measure of curvature at local minima

$f(w)$

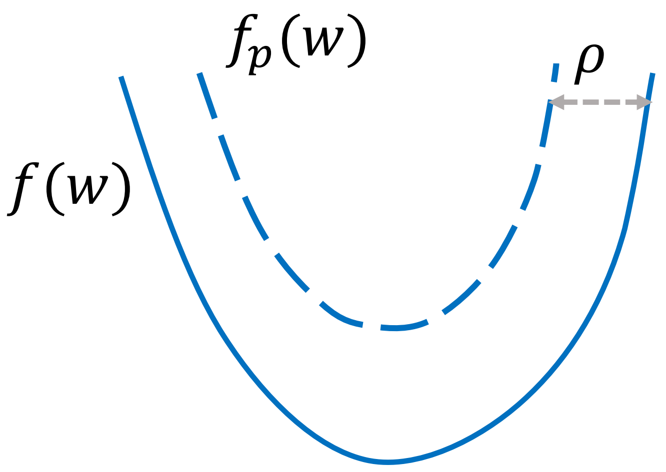
Surrogate Gap: an equivalent measure of curvature at local minima



$f(w)$: loss function

$f(w)$

Surrogate Gap: an equivalent measure of curvature at local minima



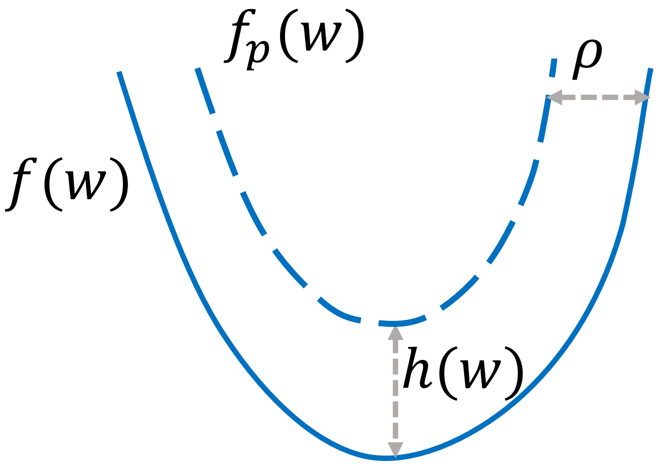
$f(w)$: loss function

$f_p(w)$: the maximum loss within the neighborhood (with a fixed radius ρ) centered around w

$$f_p(w) \approx f(w + \rho \frac{\nabla f(w)}{||\nabla f(w)||})$$

$f(w)$

Surrogate Gap: an equivalent measure of curvature at local minima



$f(w)$: loss function

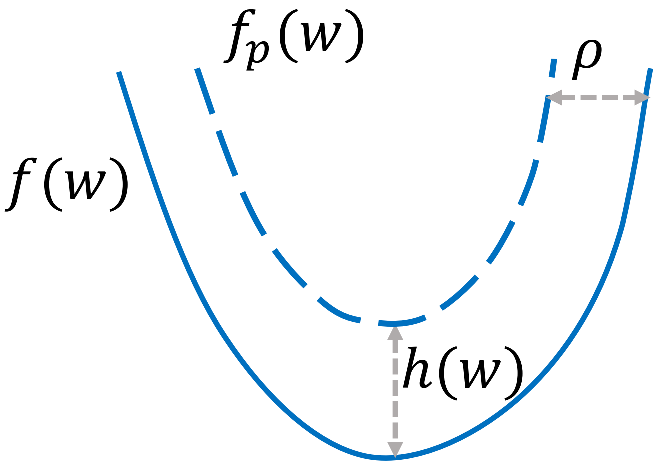
$f_p(w)$: the maximum loss within the neighborhood (with a fixed radius ρ) centered around w

$$f_p(w) \approx f(w + \rho \frac{\nabla f(w)}{||\nabla f(w)||})$$

$h(w)$: $f_p(w) - f(w)$ The “surrogate gap”

$f(w)$

Surrogate Gap: an equivalent measure of curvature at local minima



$f(w)$: loss function

$f_p(w)$: the maximum loss within the neighborhood (with a fixed radius ρ) centered around w

$$f_p(w) \approx f(w + \rho \frac{\nabla f(w)}{||\nabla f(w)||})$$

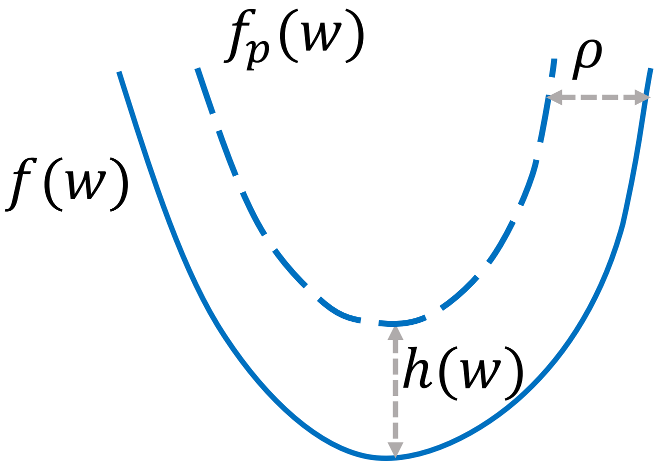
$h(w)$: $f_p(w) - f(w)$ The “surrogate gap”

At a local minimum:

$$h(w) \approx \frac{1}{2} \, |\sigma_{max}| \, \rho^2 \propto |\sigma_{max}|$$

σ_{max} : dominate eigenvalue of the Hessian

Surrogate Gap: an equivalent measure of curvature at local minima



$f(w)$: loss function

$f_p(w)$: the maximum loss within the neighborhood (with a fixed radius ρ) centered around w

$$f_p(w) \approx f\left(w + \rho \frac{\nabla f(w)}{||\nabla f(w)||}\right)$$

$h(w)$: $f_p(w) - f(w)$ The "surrogate gap"

At a local minimum:

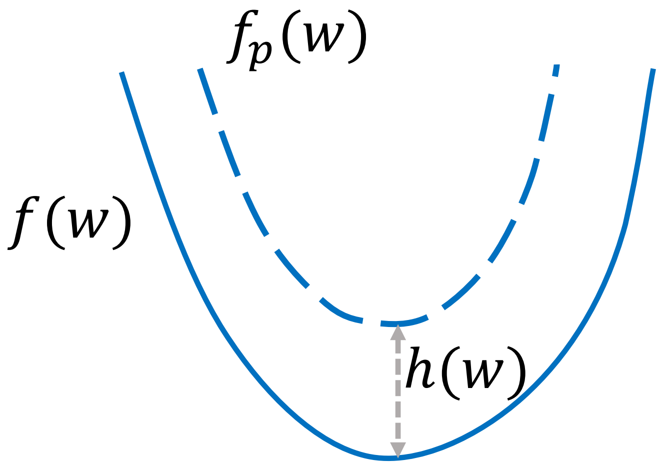
$$h(w) \approx \frac{1}{2} |\sigma_{max}| \rho^2 \propto |\sigma_{max}|$$

σ_{max} : dominate eigenvalue of the Hessian

Larger h, sharper surface, worse generalization

$f(w)$

Sharpness-Aware Minimization (SAM)

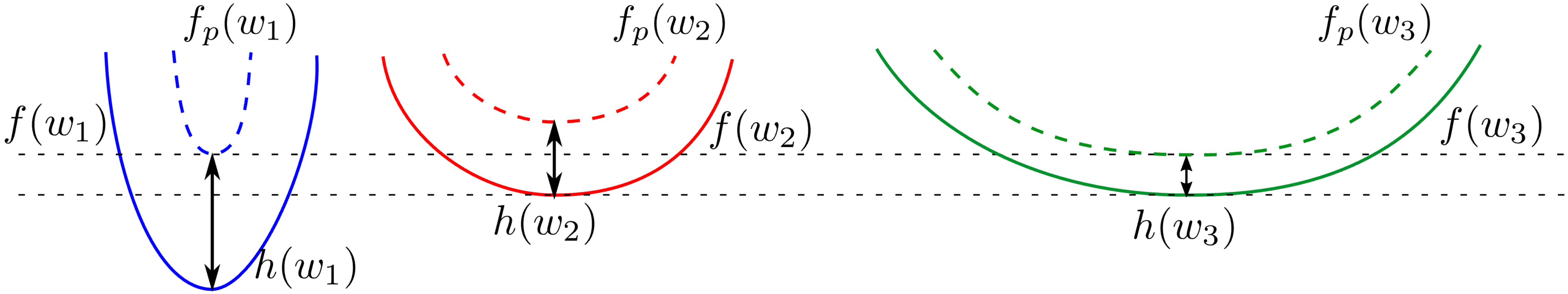


Vanilla Training: $\min f(w)$

SAM: $\min f_p(w)$

[1] Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." *arXiv preprint arXiv:2010.01412* (2020).

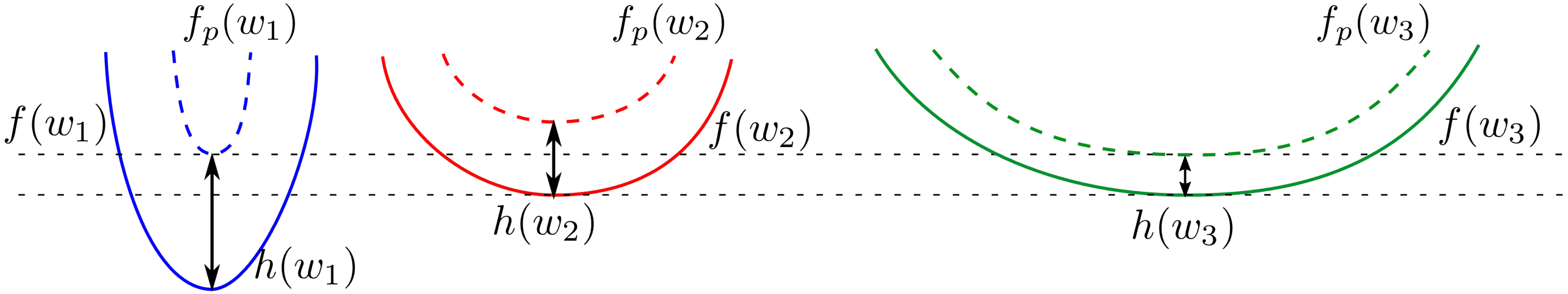
Potential Caveat of Sharpness-Aware Minimization (SAM)



Left to right:

$$Sharpness\ (w_1) > Sharpness\ (w_2) > Sharpness\ (w_3)$$
$$h(w_1) \quad > \quad h(w_2) \quad > \quad h(w_3)$$
$$f_p(w_1) \quad < \quad f_p(w_2) \quad > \quad f_p(w_3)$$

Potential Caveat of Sharpness-Aware Minimization (SAM)

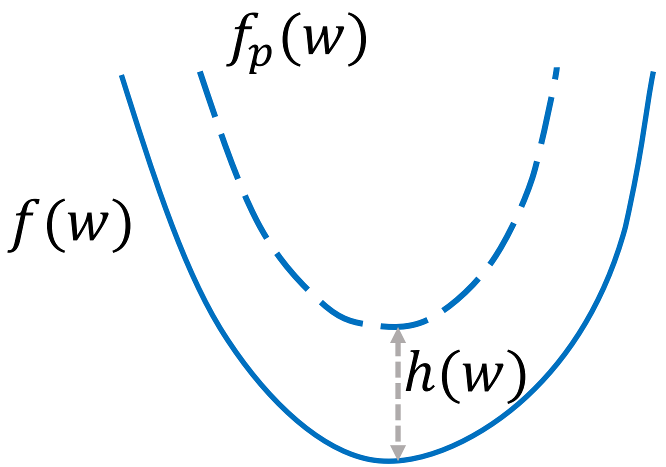


Left to right:

$$Sharpness\ (w_1) > Sharpness\ (w_2) > Sharpness\ (w_3)$$
$$h(w_1) \quad > \quad h(w_2) \quad > \quad h(w_3)$$
$$f_p(w_1) \quad < \quad f_p(w_2) \quad > \quad f_p(w_3)$$

f_p might disagree with sharpness, surrogate gap h agrees with sharpness (Lemma 3.1, 3.2, 3.3)

Surrogate Gap Guided Sharpness-Aware Minimization (GSAM)



Vanilla Training: $\min f(w)$

SAM: $\min f_p(w)$

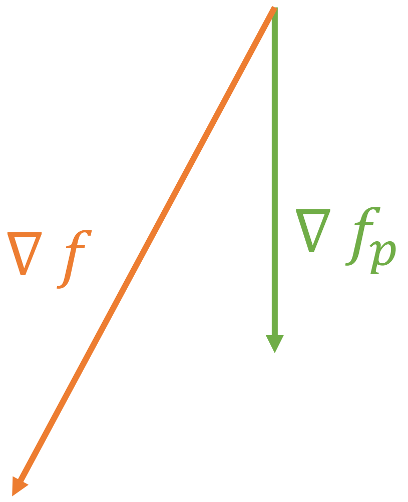
GSAM (ours): $\min(\ f_p(w)\ , h(w)\ )$

Minimize training loss

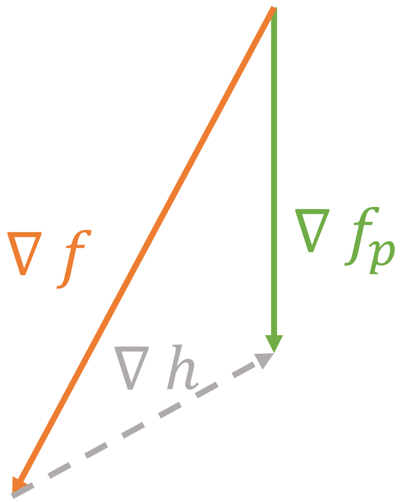
Minimize generalization gap

Simultaneously minimize two objectives that might ***compete***

GSAM (ours): $\min(\, f_p(w)\,, h(w)\,)$



Simultaneously minimize two objectives that might ***compete***

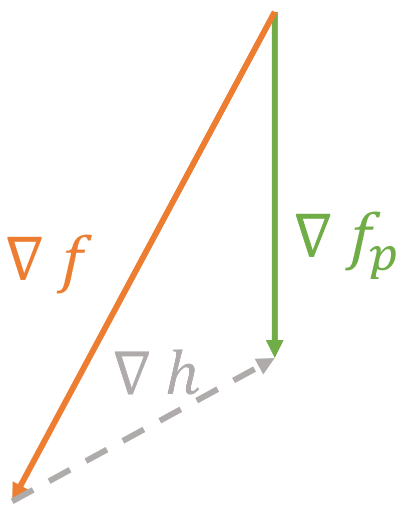


GSAM (ours): $\min(\, f_p(w)\,, h(w)\,)$

$$\nabla h = \nabla f_p - \nabla f$$

Simultaneously minimize two objectives that might ***compete***

GSAM (ours): $\min(\ f_p(w)\,, h(w)\ )$

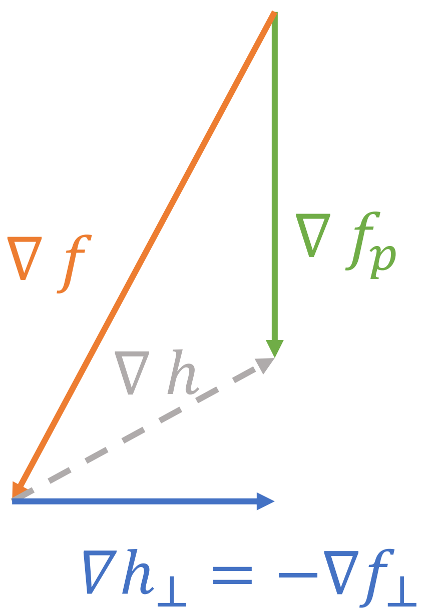


$$\nabla h = \nabla f_p - \nabla f$$

∇h might have ***negative*** inner product with ∇f_p

$\rightarrow$ minimizing h would ***increase*** f_p

Simultaneously minimize two objectives that might ***compete***



GSAM (ours): $\min(f_p(w) , h(w))$

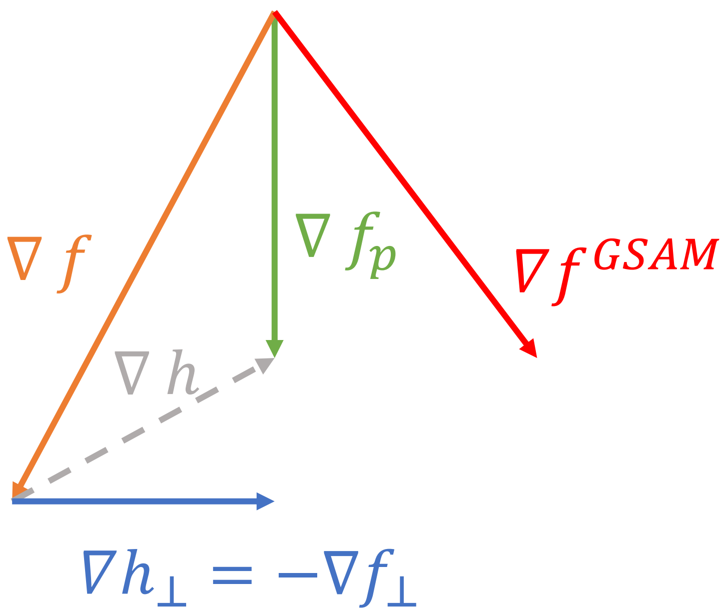
$\nabla h = \nabla f_p - \nabla f$

∇h might have ***negative*** inner product with ∇f_p
$\rightarrow$ minimizing h would ***increase*** f_p

Decompose ∇h onto parallel and vertical to ∇f_p,
update in $\nabla h_\perp$ ***not affects*** f_p.

Simultaneously minimize two objectives that might ***compete***

GSAM (ours): $\min(\, f_p(w)\,, h(w)\,)$



$\nabla h = \nabla f_p - \nabla f$

∇h might have ***negative*** inner product with ∇f_p
$\rightarrow$ minimizing h would ***increase*** f_p

Decompose ∇h onto parallel and vertical to ∇f_p,
update in $\nabla h_{\perp}$ ***not affects*** f_p.

$$\nabla f^{GSAM} = \nabla f_p + \alpha\, \nabla \mathrm{h}_{\perp} = \nabla f_p - \alpha\, \nabla f_{\perp}$$

min f_p

min h, not affect f_p

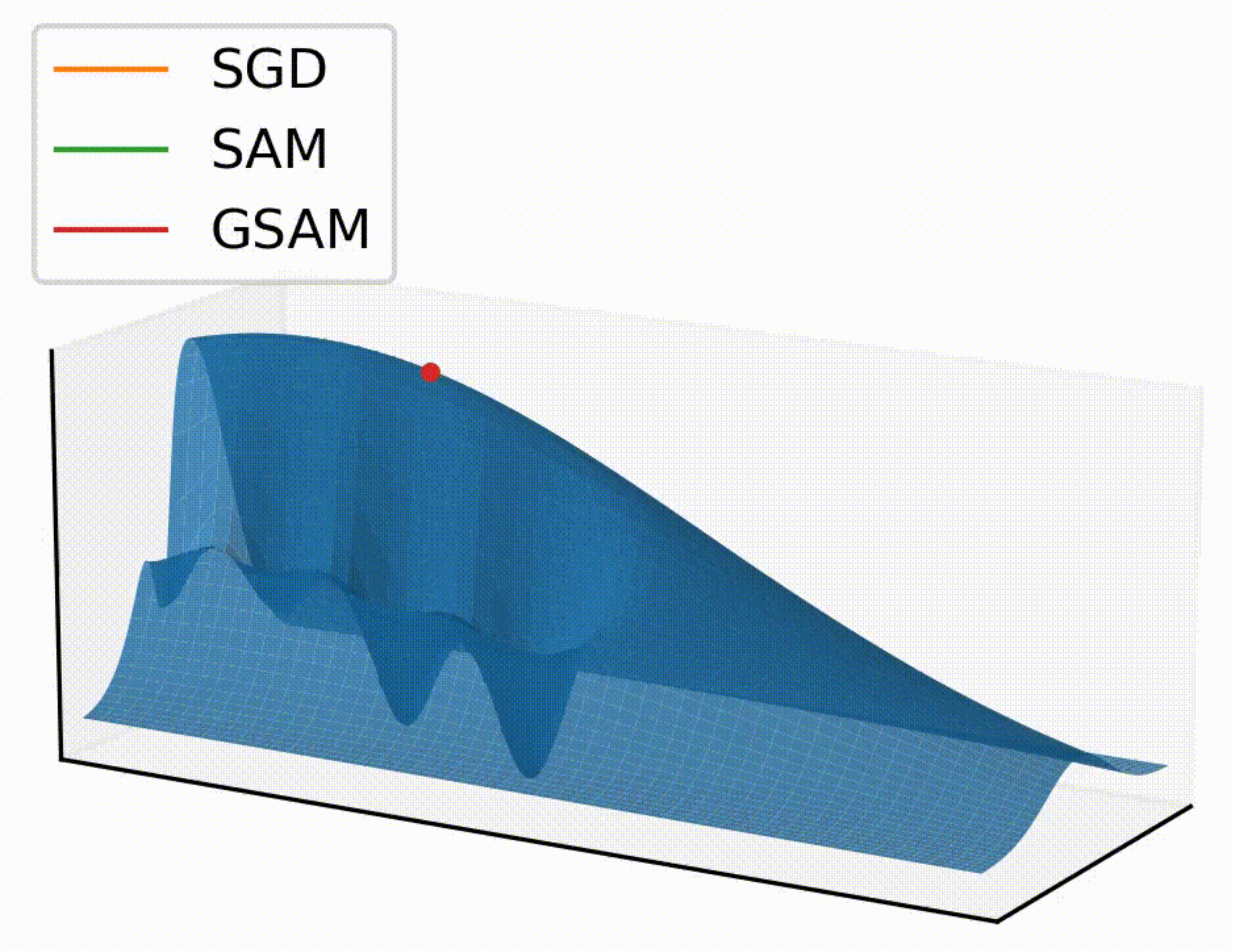

SGD
SAM
GSAM

Table 1: Top-1 Accuracy (%) on ImageNet datasets for ResNets, ViTs and MLP-Mixers trained with Vanilla SGD or AdamW, SAM, and GSAM optimizers.

Model	Training	ImageNet-v1	ImageNet-Real	ImageNet-V2	ImageNet-R	ImageNet-C
ResNet						
ResNet50	Vanilla (SGD)	76.0	82.4	63.6	22.2	44.6
	SAM	76.9	83.3	64.4	**23.8**	46.5
	GSAM	**77.2**	**83.9**	**64.6**	23.6	**47.6**
ResNet101	Vanilla (SGD)	77.8	83.9	65.3	24.4	48.5
	SAM	78.6	84.8	66.7	25.9	51.3
	GSAM	**78.9**	**85.2**	**67.3**	**26.3**	**51.8**
ResNet152	Vanilla (SGD)	78.5	84.2	66.3	25.3	50.0
	SAM	79.3	84.9	67.3	25.7	52.2
	GSAM	**80.0**	**85.9**	**68.6**	**27.3**	**54.1**

Vision Transformer						
ViT-S/32	Vanilla (AdamW)	68.4	75.2	54.3	19.0	43.3
	SAM	70.5	77.5	56.9	21.4	46.2
	GSAM	**73.8**	**80.4**	**60.4**	**22.5**	**48.2**
ViT-S/16	Vanilla (AdamW)	74.4	80.4	61.7	20.0	46.5
	SAM	78.1	84.1	65.6	24.7	53.0
	GSAM	**79.5**	**85.3**	**67.3**	**25.3**	**53.3**
ViT-B/32	Vanilla (AdamW)	71.4	77.5	57.5	23.4	44.0
	SAM	73.6	80.3	60.0	24.0	50.7
	GSAM	**76.8**	**82.7**	**63.0**	**25.1**	**51.7**
ViT-B/16	Vanilla (AdamW)	74.6	79.8	61.3	20.1	46.6
	SAM	79.9	85.2	67.5	26.4	**56.5**
	GSAM	**81.0**	**86.5**	**69.2**	**27.1**	55.7

Vision Transformer						
ViT-S/32	Vanilla (AdamW)	68.4	75.2	54.3	19.0	43.3
	SAM	70.5	77.5	56.9	21.4	46.2
	GSAM	**73.8**	**80.4**	**60.4**	**22.5**	**48.2**
ViT-S/16	Vanilla (AdamW)	74.4	80.4	61.7	20.0	46.5
	SAM	78.1	84.1	65.6	24.7	53.0
	GSAM	**79.5**	**85.3**	**67.3**	**25.3**	**53.3**
ViT-B/32	Vanilla (AdamW)	71.4	77.5	57.5	23.4	44.0
	SAM	73.6	80.3	60.0	24.0	50.7
	GSAM	**76.8**	**82.7**	**63.0**	**25.1**	**51.7**
ViT-B/16	Vanilla (AdamW)	74.6	79.8	61.3	20.1	46.6
	SAM	79.9	85.2	67.5	26.4	**56.5**
	GSAM	**81.0**	**86.5**	**69.2**	**27.1**	55.7
MLP-Mixer						
Mixer-S/32	Vanilla (AdamW)	63.9	70.3	49.5	16.9	35.2
	SAM	66.7	73.8	52.4	18.6	39.3
	GSAM	**68.6**	**75.8**	**55.0**	**22.6**	**44.6**
Mixer-S/16	Vanilla (AdamW)	68.8	75.1	54.8	15.9	35.6
	SAM	72.9	79.8	58.9	20.1	42.0
	GSAM	**75.0**	**81.7**	**61.9**	**23.7**	**48.5**
Mixer-S/8	Vanilla (AdamW)	70.2	76.2	56.1	15.4	34.6
	SAM	75.9	82.5	62.3	20.5	42.4
	GSAM	**76.8**	**83.4**	**64.0**	**24.6**	**47.8**
Mixer-B/32	Vanilla (AdamW)	62.5	68.1	47.6	14.6	33.8
	SAM	72.4	79.0	58.0	22.8	46.2
	GSAM	**73.6**	**80.2**	**59.9**	**27.9**	**52.1**
Mixer-B/16	Vanilla (AdamW)	66.4	72.1	50.8	14.5	33.8
	SAM	77.4	83.5	63.9	24.7	48.8
	GSAM	**77.8**	**84.0**	**64.9**	**28.3**	**54.4**

Conclusions:

- Surrogate Gap is an equivalent measure of sharpness
- Minimize both f_p (perturbed training loss) and h (sharpness)
- Gradient decomposition to avoid conflicts of multi-objective optimization
- Code available on project website (https://sites.google.com/view/gsam-iclr22)