arXiv:2305.05392v2 [cs.LG] 1 Jul 2023

尖锐度感知最小化 仅靠尖锐度感知最小化即可提升对抗鲁棒性 s

Zeming Wei * 1 Jingyu Zhu * 1 Yihao Zhang * 1

摘要

尖锐度感知最小化（SAM）是一种通过正则化损失尖锐度来提升泛化能力的有效方法。本文探讨了 SAM 在对抗鲁棒性方面的作用。我们发现，与标准训练相比，仅使用 SAM 就能在不牺牲干净准确率的情况下实现更强的对抗鲁棒性，这是一项意料之外的优势。我们还讨论了 SAM 与一种广受欢迎的提升深度神经网络（DNN）对抗鲁棒性的方法——对抗训练（AT）之间的关系。具体而言，我们表明，SAM 与 AT 在扰动强度方面存在差异，因而带来不同的准确率与鲁棒性权衡。我们在一个简化模型中提供了支持这些结论的理论依据。最后，鉴于 AT 会导致干净准确率下降并带来计算开销，我们认为，在某些要求下，SAM 可作为 AT 的轻量级替代方案。代码可在htt ps://github.com/weizemin g/SAM_AT.

1. 引言

尖锐度感知最小化（SAM）（Foret 等人，2020年）是一种新颖的训练框架，通过同时最小化损失值和损失尖锐度来提升模型的泛化能力。SAM 的目标是最小化参数周围的尖锐度，可表示为

$$\max_{\|\epsilon\|\leq\rho} L(w+\epsilon)+\lambda\|w\|_2^2, \tag{1}$$

其中， L 是损失函数， w 是模型的参数， $\|w\|_2^2$ 是正则化项， ρ 控制权重扰动的幅度。直观来说，较大的 ρ 会带来更强的权重扰动，并促使模型寻找更平坦的损失景观。迄今为止，SAM 已经成为一种能够提升机器学习模型自然准确率表现的强大工具。

本文旨在从对抗鲁棒性的角度探究 SAM。具体而言，我们研究 SAM 抵御对抗样本的鲁棒性。对抗样本是经过微小扰动、会误导模型产生错误预测的自然样本（Szegedy 等人，2013年；Goodfellow 等人，2014）。对抗样本的发现引发了人们对关键领域应用安全性的严重担忧（Ma 等人，2020年），并吸引了大量关于如何防御此类样本的研究。目前，在各种防御方法中，对抗训练（AT）（Madry 等人，2017年）已被证明是提升深度神经网络（DNN）的对抗鲁棒性最有效的方法（Athalye 等人，2018年）。然而，尽管对抗鲁棒性有所提升，对抗训练仍存在一些缺陷，例如自然准确率下降（Tsipras 等，2018年）、计算开销（Shafahi 等人，2019年）、类别公平性（Xu 等人，2021年；Wei 等人，2023a）以及缺乏形式化保证（Wang 等人，2021年；Zhang 等人，2023年）。

令人惊讶的是，我们发现，采用 SAM 训练的模型表现出显著高于采用标准方法训练的模型的对抗鲁棒性，这是一个意想不到的优势。此外，与对抗训练（AT）相比，SAM 具有计算成本更低且不会降低自然准确率的优势。基于上述讨论，本文提出了两个研究问题（RQs）：

- RQ1：与标准训练相比，为什么 SAM 能提高对抗鲁棒性?
- RQ2：SAM 能否作为对抗训练的一种轻量级替代方案?

为回答上述两个问题，我们首先从对抗鲁棒性的角度全面理解 SAM。具体而言，我们指出 SAM 与 Tsipras 等，2018年）的对抗训练（AT）存在内在联系：二者在训练阶段都会对自然样本进行对抗数据增强，以消除非鲁棒特征。因此，SAM 和 AT 都能有效提升训练后模型的鲁棒性，从而

*贡献相同 [1]北京大学。通信作者：Zeming Wei < weizeming@stu.pku.edu.cn>。

2^{nd} AdvML Frontiers 研讨会，举办于 40^{th} 国际机器学习会议，檀香山，夏威夷，美国。PMLR 202，2023年。版权所有 2023年，作者所有。

提升鲁棒泛化能力。不过，我们也注意到，SAM 与 AT 之间仍存在一些差异。例如，SAM 隐式地添加对抗扰动，而 AT 则显式地施加扰动。此外，SAM 和 AT 训练过程中的扰动（攻击）强度也有所不同，因此在自然准确率与鲁棒准确率之间会产生不同的权衡。

此外，我们在简化的数据模型中以理论证据验证了所提出的经验理解。遵循基于鲁棒特征和非鲁棒特征分解的数据分布（Tsipras 等，2018年），我们表明，SAM 和对抗训练都可以通过对鲁棒特征赋予更高权重来提升训练模型的鲁棒性。此外，我们还表明，SAM 需要更大的扰动预算才能达到与对抗训练相当的鲁棒性，这验证了我们的假设：SAM 的扰动强度低于对抗训练。

最后，我们在基准数据集上进行了实验，以验证我们的理解。我们发现，与标准训练的模型相比，采用 SAM 训练的模型在对抗鲁棒性方面确实显著更优，同时还表现出更高的自然准确率。总而言之，我们的经验和理论理解可以回答 RQ1。

值得注意的是，SAM 与对抗训练之间的鲁棒性仍存在较大差距。然而，对抗训练的自然准确率始终低于标准训练，更不用说 SAM 了。同时，SAM 在计算成本方面也优于对抗训练。因此，我们最终通过以下结论回答 RQ2：在满足以下要求的情况下，SAM 可以被视为一种用于提升对抗鲁棒性的轻量级对抗训练替代方案：（1）不损失自然准确率；（2）不显著增加计算成本。

总而言之，本文的主要贡献如下：

1. 我们指出，与标准训练相比，单独使用 SAM 能够在不牺牲干净准确率的情况下，显著增强对抗鲁棒性，这是一个出乎意料的益处。

2. 我们从经验和理论两方面进行解释，以阐明 SAM 如何增强对抗鲁棒性。特别是，我们讨论了 SAM 与对抗训练之间的关系，并证明二者通过消除非鲁棒特征来提升对抗鲁棒性。然而，二者的扰动强度不同，从而导致自然准确率与鲁棒准确率之间存在不同的权衡。

3. 我们在基准数据集上进行了实验，以验证我们提出的见解。我们还建议，在某些要求下，可以将 SAM 视为对抗训练的一种轻量级替代方案。

2. 背景与相关工作

2.1. 尖锐度感知最小化（SAM）

为了处理传统机器学习算法中的泛化不佳问题（Hochreiter& Schmidhuber，1994年；1997年），两者分别尝试寻找平坦极小值，并惩罚损失景观中的尖锐度，从而获得良好的泛化效果（Keskar 等人，2016年；Neyshabur 等人，2017年；Dziugaite & Roy，2017年）。受此启发，一系列研究着重于利用损失景观中的平坦性或尖锐度这一概念，以确保更好的泛化，例如：Entropy-SGD（Chaudhari 等人，2019年）和随机权重平均（SWA）（Izmailov 等人，2018年）。尖锐度感知最小化（SAM）（Foret 等人，2020年）也属于这一类，它同时最小化损失值和损失尖锐度，如（1）所述。

从理论上讲，SAM 的良好泛化能力由以下事实得到保证：以高概率，下面的不等式成立：

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + h\left(\|\boldsymbol{w}\|_2^2/\rho^2\right), \quad (2)$$

其中集合 $\mathcal{S}$ 由分布 $\mathscr{D}$, h 生成：$\mathbb{R}_+ \to \mathbb{R}_+$ 是严格递增函数。

SAM 也在语言模型（Bahri 等人，2021年）和流体动力学（Jetly 等人，2022年）等其他研究领域有许多应用，展现了 SAM 在不同领域的可扩展性。此外，还涌现出许多对 SAM 算法的改进，例如自适应 SAM（ASAM）（Kwon 等人，2021年）、高效 SAM（ESAM）（Du 等人，2021年）、LookSAM（Liu 等人，2022年）、稀疏 SAM（SSAM）（Mi 等人，2022年）、Fisher SAM（Kim 等人，2022年）和 FSAM（Zhong 等人，2022年），这些方法在 SAM 的基础上进行了一些改进，进一步提升了模型的泛化能力。

2.2. 对抗鲁棒性

自对抗样本被发现以来，对抗鲁棒性和对抗训练已成为热门研究课题（Szegedy 等人，2013年；Goodfellow 等人，2014），这一发现揭示了：通过向正常样本添加微小扰动构造的对抗样本，可以轻易欺骗 DNN，使其做出错误决策。恶意攻击者可以通过构造对抗样本发动对抗攻击（Chen 等人，2023b；Wei 等人，2023b），这给 DNN 的部署带来了严重的安全隐患。迄今为止，研究人员已提出大量防御方法（Papernot 等人，2016年；Xie 等人，2019年；Bai 等人，2019年；Mo 等人，2022年；Chen 等人，2023a），其中对抗训练

(对抗训练) (Madry等人，2017年；Wang 等人，2019年) 一直被认为是最有前景的对抗攻击防御方法，可表示为

$$\min_{w} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\delta\|\leq\epsilon} L(w; x+\delta, y), \tag{3}$$

其中， $\mathcal{D}$ 是数据分布， ϵ 是扰动幅度， w 是模型参数， L 是损失函数（例如交叉熵损失）。对于内层最大化过程，通常使用投影梯度下降（PGD）攻击来生成对抗样本：

$$x^{t+1} = \Pi_{\mathcal{B}(x,\epsilon)}(x^t + \alpha \cdot \text{sign}(\nabla_x \ell(\theta; x^t, y))), \tag{4}$$

其中， Π 将对抗样本投影到扰动范围 $\mathcal{B}(x, \epsilon) = \{x' : \|x'-x\|_p \leq \epsilon\}$，而 α 表示梯度上升的步长。

尽管对抗训练能有效提升对抗鲁棒性，但它也暴露出若干缺陷，例如计算开销过大（Shafahi 等人，2019年）、类别间公平性问题（Xu 等人，2021年；Wei 等人，2023a）等。其中，自然准确率下降（Tsipras 等，2018年；Wang与Wang，2023年）已成为主要问题。研究证明，鲁棒性与自然准确率之间存在内在权衡（Tsipras 等，2018年），这可以解释为什么对抗训练会显著降低标准准确率。

在对抗鲁棒性的背景下，也有一些工作尝试在对抗训练中引入平坦损失景观（Wu 等，2020年；Yu 等，2022年a；b）。其中最具代表性的是对抗权重扰动（AWP）（Wu 等，2020年），该方法同时在样本和特征空间上添加扰动，以便在对抗训练中应用尖锐度感知最小化。然而，AWP 也会导致自然准确率下降。此外，平坦损失景观为何能够带来更好的鲁棒性，尚未得到充分解释。

据我们所知，我们首次揭示了 SAM 与对抗训练（AT）之间的内在联系，并发现 SAM 能够通过隐式地赋予鲁棒特征更高的权重来提升对抗鲁棒性。

3. 经验理解

在本节中，我们介绍所提出的关于 SAM 与对抗训练之间关系的经验性理解，这有助于解释 SAM 如何提升对抗鲁棒性。

回顾一下，SAM 的目标是同时最小化泛化误差和损失尖锐度。尖锐度项可表示为 $\max_{||\boldsymbol{\epsilon}||<\rho}[L_S(\boldsymbol{w}+\boldsymbol{\epsilon}) - L_S(\boldsymbol{w})]$，且损失项为 $L_S(\boldsymbol{w})$。将这两项结合起来，我们得到 SAM 的目标函数为

$$\min_{\boldsymbol{w}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{||\boldsymbol{\epsilon}||<\rho} L_S(\boldsymbol{w}+\boldsymbol{\epsilon}; x, y). \tag{5}$$

另外，回顾一下，对抗训练的目标是

$$\min_{\boldsymbol{w}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\delta\|\leq\epsilon} L_s(\boldsymbol{w}; x+\delta, y). \tag{6}$$

为了说明二者的关系，我们首先强调，这两种技术都涉及添加扰动，将其作为一种数据增强方式，以消除非鲁棒特征（Ilyas 等人，2019年）。不过，对抗训练会显式地将这些扰动添加到输入样本中，而 SAM 则侧重于扰动参数，这可视为在特征空间上进行的一种隐式数据增强。因此，两种技术都会对特征施加扰动，但作用于不同的空间。

更具体、形式化地说，我们可以通过模型中的一个中间线性层来推导这一理解，该层从输入 x 中提取特征 z： $z = Wx$。在对抗训练中，我们直接在输入空间中添加扰动，得到$x \leftarrow x + \delta$。然而，在 SAM 中，扰动并非直接施加于输入空间，而是按照 $W \leftarrow W + \delta$ 施加于参数空间。这会分别得到输入扰动对应的 $Wx + W\delta$ 和参数扰动对应的 $Wx + \delta x$。这两种扰动都可视为一种数据增强，前者更显式，后者则更隐式。

此外，我们还讨论了对抗训练和 SAM 的攻击（扰动）强度。对于 SAM 而言，其扰动相对更为温和，因为扰动被注入特征空间。然而，这种微小扰动仍有助于提升鲁棒性，因为它能够隐式地消除非鲁棒特征。另一方面，为了通过破坏非鲁棒特征来实现最佳鲁棒性，对抗训练会对输入空间施加更大且更直接的扰动，从而获得更好的鲁棒性，但会损失自然准确率。

因此，SAM 与对抗训练之间的差异和联系可以被视为鲁棒性与准确率之间的权衡（Zhang 等人，2019年）。总而言之，SAM 将小幅扰动隐式地施加到特征空间中，以维持良好的自然准确率表现；而对抗训练则采用直接的数据增强幅度，这可能导致自然准确率严重下降。我们将在下一节中为这些论断提供更多理论证据。

4. 理论分析

本节将对 SAM 进行理论分析，并探讨 SAM 与对抗训练之间的关系。遵循鲁棒/非鲁棒特征分解（Tsipras 等，

2018年），我们引入一个简单的二分类模型，并在其中展示 SAM 与对抗训练隐含的本质相似性和差异。我们首先介绍数据分布和假设空间，然后分别说明 SAM 与对抗训练在该模型中的作用，最后讨论二者之间的关系。

4.1. 二分类模型

考虑一个二分类任务，其输入-标签对 $(\boldsymbol{x}, y)$ 从 $\boldsymbol{x} \in \{-1,+1\} \times \mathbb{R}^d$ 和 $y \in \{-1,+1\}$ 中采样，分布 $\mathcal{D}$ 定义如下。

$$\begin{aligned} y &\overset{\text{u.a.r}}{\sim} \{-1,+1\}, x_1 = \begin{cases} +y, & \text{w.p. } p, \\ -y, & \text{w.p. } 1-p, \end{cases} \\ x_2, \ldots, x_{d+1} &\overset{i.i.d}{\sim} \mathcal{N}(\eta y, 1), \end{aligned} \tag{7}$$

其中 $p \in (0.5,1)$ 是特征 x_1的准确率，常数$\eta > 0$ 是一个很小的正数。在该模型中，x_1 被称为鲁棒特征，因为任何微小扰动都无法改变其符号。然而，鲁棒特征并不完美，因为 $p < 1$。相应地，特征 x_2，$\cdots$，x_{d+1}有助于识别 y，这是由于其符号的一致性，因此它们有助于提高自然准确率。然而，由于 η 是一个很小的正数，这些特征很容易被扰动到相反的一侧（符号发生改变），因此被称为非鲁棒特征（Ilyas 等人，2019年）。

现在考虑一个线性分类器模型，它通过计算 $f_{\boldsymbol{w}}(\boldsymbol{x}) = \text{sgn}(\boldsymbol{w} \cdot \boldsymbol{x})$ 来预测数据点的标签，并优化参数 $w_1, w_2, \cdots, w_n$，以最大化$\mathbb{E}_{x.y\sim\mathcal{D}}\mathbf{1}(f_{\boldsymbol{w}}(x) = y)$。在该模型中，鉴于 $x_i (i = 2, \cdots, n)$ 的等价性，我们可以通过归一化设置 $w_2 = \cdots = w_n = 1$，且不失一般性。因此，数值 w_1 与模型的鲁棒性密切相关。具体而言，较大的 w_1 表示模型对鲁棒特征 x_1 赋予了更大的权重，而对非鲁棒特征 $x_2, \cdots, x_{d+1}$赋予较小的权重，从而提升鲁棒性。

接下来，我们分别讨论标准训练（ST）、对抗训练（AT）和SAM训练下的模型。为使描述清晰，我们将损失函数 $\mathcal{L}(\boldsymbol{x}, y, w)$ 记为$1 - \Pr(f_w(\boldsymbol{x}) = y)$；对于给定的 $\epsilon > 0$，我们将 SAM 的损失函数 $\mathcal{L}^{SAM}$ 定义为 $\max_{|\delta|\le\epsilon} \mathcal{L}(\boldsymbol{x}, y, w + \delta)$。

4.2. 标准训练（ST）

我们首先通过以下定理证明，在该模型中，标准训练存在一个最优参数 w_1^*：

定理 4.1（标准训练）。在上述模型中，在标准训练下，最优参数值为

$$w_1^* = \frac{\ln p - \ln(1-p)}{2\eta}. \tag{8}$$

因此，w_1^*可以视为使用该模型进行标准训练后返回的参数 w_1。

4.3. 对抗训练（AT）

现在考虑应用对抗训练的情况。此时，损失函数不再是标准损失函数，而是期望对抗损失

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{||\boldsymbol{\delta}||_\infty \le \epsilon} \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}, y; w) \right]. \tag{9}$$

与标准训练类似，对抗训练也会得到一个最优参数 w_1^{AT}，具体如下定理所述：

定理 4.2（对抗训练）。在上述分类问题中，在扰动界为 $\epsilon < \eta$的对抗训练下，对抗最优参数值

$$w_1^{AT} = \frac{\ln p - \ln(1-p)}{2(\eta - \epsilon)}. \tag{10}$$

我们可以看到，w_1 已乘以$\frac{\eta}{\eta-\epsilon}$，这增加了分类器对鲁棒特征 x_1 的依赖。这表明，与标准训练的模型相比，对抗训练模型更加关注鲁棒性，从而提升了对抗鲁棒性。

4.4. 尖锐度感知最小化（SAM）

现在我们来考虑 SAM 的情况。回顾一下，SAM 的优化目标是

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{|\delta|\le\epsilon} \mathcal{L}(x, y; w + \delta) \right]. \tag{11}$$

我们首先证明使用 SAM 训练得到的参数 w_1 也大于 w_1^*，以此说明 SAM 能提升对抗鲁棒性。具体表述如下：

定理 4.3（尖锐度感知最小化）。在上述分类问题中，SAM 训练的最优参数 w_1^{SAM}满足

$$w_1^{SAM} > w_1^*. \tag{12}$$

根据定理4.2 和 4.3 我们可以看到，w_1^{AT}和w_1^{SAM}都大于 w_1^*，这表明 SAM 和对抗训练都能提升训练后模型的鲁棒性。然而，这种定性关系不足以量化 SAM 相较于对抗训练能达到多大程度的鲁棒性，因此我们尝试进一步定量估计以下定理中的w_1^{SAM}：

定理 4.4。在上述分类问题中，记 SAM 训练的最佳参数为 w_1^{SAM}。假设ϵ 较小，则有 $w_1^{SAM} \approx w_1^* + \frac{2}{3} w_1^* \epsilon^2$。

表1。在 CIFAR-100 数据集上评估自然准确率和鲁棒准确率。

方法	自然准确率	ℓ_∞-鲁棒准确率		ℓ_2-鲁棒准确率	
		$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 16/255$	$\epsilon = 32/255$
ST	76.9	13.6	1.7	44.5	21.2
SAM ($\rho = 0.1$)	78.0	19.6	3.0	51.5	27.2
SAM ($\rho = 0.2$)	78.5	23.1	4.2	54.2	31.3
SAM ($\rho = 0.4$)	**78.7**	**28.3**	**6.5**	**57.0**	**36.2**
对抗训练（ℓ_∞-$\epsilon = 1/255$）	**73.1**	**60.4**	46.6	**67.4**	**61.5**
对抗训练（ℓ_∞-$\epsilon = 2/255$）	70.1	60.3	50.6	65.7	60.6
对抗训练（ℓ_∞-$\epsilon = 4/255$）	66.2	59.3	**52.0**	62.8	58.9
对抗训练 (ℓ_∞-$\epsilon = 8/255$)	60.4	55.1	50.4	57.0	54.3
对抗训练 (ℓ_2-$\epsilon = 16/255$)	**74.8**	52.8	31.4	66.3	57.7
对抗训练 (ℓ_2-$\epsilon = 32/255$)	73.2	57.4	40.6	**67.1**	**61.0**
对抗训练 (ℓ_2-$\epsilon = 64/255$)	70.7	58.1	45.9	66.1	60.7
对抗训练 (ℓ_2-$\epsilon = 128/255$)	67.4	**58.2**	**48.7**	63.9	60.4

4.5. SAM 与对抗训练之间的关系

我们进一步讨论对抗训练与 SAM 之间不同的攻击（扰动）强度。回顾第3节中的经验理解，SAM 的扰动比对抗训练更温和，这可以解释为：在鲁棒性与准确率的权衡中，SAM 更侧重自然准确率，而较少侧重鲁棒性。因此，为了达到相同的鲁棒性水平（该水平通过对特征的依赖程度来衡量 x_1，即的大小 w_1），SAM 需要大得多的扰动范围，而对抗训练只需对 x 施加较小的扰动即可。从理论上讲，以下定理验证了我们的论述：

定理4.5。分别将对抗训练和 SAM 的扰动范围记为 ϵ ϵ_{AT} 和 ϵ_{SAM}。那么，当两种方法返回相同的参数 w_1 时， ϵ_{AT} 和 ϵ_{SAM} 之间有如下关系：

$$2 + \frac{3}{\epsilon_{SAM}^2} \approx \frac{2\eta}{\epsilon_{AT}} \tag{13}$$

根据定理4.5，我们可以看出对抗训练和 SAM 的扰动强度不同。由定理4.5 可轻易推得 ϵ_{SAM} 大于 ϵ_{AT}，当13成立时，因为我们假设 η 是一个较小的正数，定理 ϵ 4.4 中较小，且定理4.2 中 $\epsilon_{AT} < \eta$。因此，要在鲁棒特征上获得相同的权重 w_1， $x_1\epsilon_{AT}$只需选择远小于 ϵ_{SAM}。另一方面，在相同的扰动界 $\epsilon_{AT} = \epsilon_{SAM}$下，对抗训练所得模型的参数 w_1 大于 SAM，因此它更注重鲁棒性，但自然准确率下降得更多。

所有证明均可在附录 A中找到。总而言之，我们可以得出结论：对抗训练利用显式且强烈的扰动来去除非鲁棒特征，而 SAM 则利用温和的隐式扰动。

这与我们在第 3 节中的经验理解一致，我们也将在下一节通过实验进一步验证这些观点。

5. 实验

本节中，我们展示实验结果，以验证我们提出的理解。

5.1. 实验设置

为展示 SAM 在提升对抗鲁棒性方面的有效性，我们将使用标准 SGD 优化器训练的模型与使用 SAM 训练的模型进行比较。我们还讨论对抗训练。不过，我们将对抗训练（AT）所获得的鲁棒性视为上界，而非 SAM 的基线。

在我们的实验中，我们使用交叉熵损失在 CIFAR-10 和 CIAR-100 数据集（He 等人，2016年）上训练 PreActResNet-18（PRN-18）模型，共训练 100 个 epoch。学习率初始值为 0.1，并分别在第 75 和第 90 个 epoch 时缩小为原来的十分之一。对于优化器，权重衰减设为 `5e-4`，动量设为 0.9。

对于 SAM，我们从范围 $\{0.1, 0.2, 0.4\}$ 中选择扰动超参数 ρ。对于对抗训练（AT），我们同时考虑 ℓ_2 和 ℓ_∞ 鲁棒性，并针对两种范数分别使用不同的扰动界训练 4 个模型。

在鲁棒性评估方面，我们考察 ℓ_∞-范数扰动界 $\epsilon \in \{1/255, 2/255\}$ 和ℓ_2-范数扰动界 $\epsilon \in \{16/255, 32/255\}$ 下的鲁棒性。评估采用 10 步 PGD攻击。

表 2. CIFAR-10 数据集上的自然准确率和鲁棒准确率评估。

方法	自然准确率	ℓ_∞-鲁棒准确率		ℓ_2-鲁棒准确率	
		$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 16/255$	$\epsilon = 32/255$
ST	94.6	39.6	8.9	76.1	51.7
SAM ($\rho = 0.1$)	**95.6**	45.1	9.4	81.0	56.3
SAM ($\rho = 0.2$)	95.5	48.9	10.2	82.9	58.8
SAM ($\rho = 0.4$)	94.7	**56.1**	**15.6**	**84.0**	**64.4**
对抗训练 (ℓ_∞-$\epsilon = 1/255$)	**93.7**	86.4	75.5	**90.5**	86.4
对抗训练 (ℓ_∞-$\epsilon = 2/255$)	92.8	**87.4**	79.6	90.3	**86.9**
对抗训练 (ℓ_∞-$\epsilon = 4/255$)	90.9	86.4	**81.3**	88.3	85.7
对抗训练 (ℓ_∞-$\epsilon = 8/255$)	84.2	81.5	78.4	82.5	80.8
对抗训练 (ℓ_2-$\epsilon = 16/255$)	**94.5**	82.2	61.7	90.3	84.5
对抗训练 (ℓ_2-$\epsilon = 32/255$)	93.7	84.9	70.9	**91.0**	86.7
对抗训练 (ℓ_2-$\epsilon = 64/255$)	92.7	85.6	75.2	90.7	**87.5**
对抗训练 (ℓ_2-$\epsilon = 128/255$)	90.2	**85.7**	**78.4**	89.6	87.1

对于所有模型，我们都独立重复实验三次，并报告平均结果。由于标准差较小（低于 0.5%），且不会影响我们的结论，因此省略标准差。

5.2. 准确率与鲁棒性评估

在 CIFAR-100 和 CIFAR-10 数据集上进行的实验结果分别列于表1 和表2中。

我们首先讨论 SAM 的自然准确率和鲁棒准确率表现。从表中可以看出，与采用标准训练（ST）训练的模型相比，所有采用 SAM 训练的模型都表现出明显更高的自然准确率和鲁棒性。具体而言，使用 SAM 时，ρ 的取值越大，鲁棒性越高。对于 CIFAR-100 数据集，使用 $\rho = 0.4$ 训练的模型甚至比 ST 达到了高出数倍的鲁棒准确率，且其自然准确率仍高于 ST。相比于自然准确率的提升（约 2%），鲁棒性的增幅更为显著（平均超过 10%）。同样，对于 CIFAR-10 数据集，采用 SAM 训练的模型在干净准确率方面也优于 ST，且鲁棒性显著高于 ST。因此，我们可以得出结论：采用相对较大的权重扰动界 ρ 的 SAM，是一种有望提升模型对抗鲁棒准确率、同时不牺牲自然准确率的技术。

对于对抗训练模型，尽管 SAM 与对抗训练所获得的鲁棒性之间仍存在很大差距，但所有对抗训练模型的自然准确率都低于标准训练，更不用说 SAM 了。尤其是对于 ℓ_∞-对抗训练，即使使用扰动界 $\epsilon = 1/255$ 进行训练，也会降低自然准确率。

在 CIFAR-100 和 CIFAR-10 数据集上，分别下降了 3.8% 和 0.9%。此外还需注意，在对抗训练中使用的扰动界 ϵ 越大，相应模型的自然准确率就越低。因此，相较于对抗训练，使用 SAM 的一个关键优势是不会降低干净准确率。此外还需注意，在对抗训练中求解 PGD 过程会带来显著的计算开销。具体而言，由于我们使用 10 步 PGD，所有对抗训练实验的计算成本都是 ST 的 10 倍，而 SAM 只需多出 1 倍。

基于上述讨论，我们得出结论：与标准训练方法相比，经过 SAM 训练的模型能够显著提升鲁棒性，同时不会降低任何自然准确率。此外，SAM 的另一项优势在于，它不需要大量的计算资源。因此，我们指出，SAM 可以作为对抗训练的一种轻量级替代方案，在不降低自然准确率且不会带来显著训练开销的情况下提升鲁棒性。

6. 结论

本文表明，仅使用尖锐度感知最小化（SAM）即可提升对抗鲁棒性，并揭示了 SAM 与对抗训练（AT）之间的基本联系。我们通过实证和理论分析证明，SAM 和 AT 都会向特征添加扰动，以实现更强的鲁棒泛化能力。然而，SAM 以隐式方式添加适度扰动，而 AT 则以显式方式添加强扰动。因此，两者在准确率和鲁棒性之间形成了不同的权衡。我们还在基准数据集上开展实验，以验证所提出观点的有效性。最后，我们指出，在特定要求下，SAM 可作为 AT 的轻量级替代方案。

参考文献

Athalye、A.、Carlini、N. 和 Wagner、D. 混淆梯度会带来虚假的安全感：规避针对对抗样本的防御措施。载于《国际机器学习会议论文集》，第 274–283 页。PMLR，2018年。

Bahri, D., Mobahi, H., and Tay, Y. 尖锐度感知最小化提升语言模型泛化能力。arXiv 预印本 arXiv:2110.08529, 2021年。

Bai, Y., Feng, Y., Wang, Y., Dai, T., Xia, S.-T., and Jiang, Y. 面向对抗样本的基于希尔伯特空间的生成式防御。发表于国际计算机视觉大会，2019年。

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. 熵随机梯度下降：将梯度下降偏置到宽阔的谷底。《统计力学杂志：理论与实验》，2019年(12):124018，2019年。

Chen, H., Dong, Y., Wang, Z., Yang,X., Duan, C., Su, H., and Zhu, J. 通过单个扩散模型实现鲁棒分类，2023a。

Chen, H., Zhang, Y., Dong, Y., and Zhu, J. 重新思考基于迁移的对抗攻击中的模型集成，2023b。

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. 高效的尖锐度感知最小化方法，可改进神经网络的训练。arXiv预印本 arXiv:2110.03141，2021年。

Dziugaite, G. K. and Roy, D. M. 为参数数量远超训练数据的深度（随机）神经网络计算非平凡泛化界。arXiv预印本 arXiv:1703.11008，2017年。

Foret, P.、Kleiner, A.、Mobahi, H. 和 Neyshabur, B. 尖锐度感知最小化：高效提升泛化能力。arXiv 预印本 arXiv:2010.01412，2020年。

Goodfellow, I. J., Shlens, J. 和 Szegedy, C. 解释并利用对抗样本。arXiv 预印本 arXiv:1412.6572，2014 年。

He, K., Zhang, X., Ren, S. 和 Sun, J. 深度残差网络中的恒等映射。载于欧洲计算机视觉会议，第 630–645 页。施普林格，2016年。

Hochreiter、S. 和 Schmidhuber、J. 通过发现平坦极小值简化神经网络。《神经信息处理系统进展》，第 7 卷，1994 年。

Hochreiter, S. 和 Schmidhuber, J. 平坦极小值。《神经计算》，9(1):1–42，1997年。

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. 和 Madry, A. 对抗样本不是错误，而是特征。载于《神经信息处理系统》，2019年。Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. 和 Wilson, A. G. 权重平均能够带来更宽的最优解和更好的泛化。arXiv 预印本 arXiv:1803.05407，2018年。

Jetly、V.、Ibayashi、H. 和 Nakano、A. 瞬间溅起：用于高效液体飞溅模拟的尖锐度感知最小化。2022年。

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. 和 Tang, P. T. P. 大批量深度学习训练：泛化差距与尖锐极小值。arXiv 预印本 arXiv:1609.04836，2016年。

Kim, M., Li, D., Hu, S. X. 和 Hospedales, T. Fisher SAM：信息几何与尖锐度感知最小化。载于《国际机器学习会议》，第 11148–11161 页。PMLR，2022年。

Krizhevsky, A., Hinton, G. 等。从微小图像中学习多层特征。2009年。

Kwon, J., Kim, J., Park, H. 和 Choi, I. K. ASAM：用于深度神经网络尺度不变学习的自适应尖锐度感知最小化。载于《国际机器学习会议》，第 5905–5914 页。PMLR，2021年。

Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., and You, Y. 迈向高效且可扩展的尖锐度感知最小化。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12360–12370页，2022年。

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., and Lu, F. 理解针对基于深度学习的医学图像分析系统的对抗攻击。《模式识别》，2020年。

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. 迈向能够抵御对抗攻击的深度学习模型。arXiv预印本 arXiv:1706.06083，2017年。

Mi, P., Shen, L., Ren, T., Zhou, Y., Sun, X., Ji, R., and Tao, D. 强化尖锐度感知最小化：一种稀疏化扰动方法。arXiv预印本 arXiv:2210.05177，2022年。

Mo, Y., Wu, D., Wang, Y., Guo, Y., and Wang, Y. 当对抗训练遇上视觉 Transformer：从训练到架构的实践方案。载于神经信息处理系统大会，2022年。

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Sre- bro, N. 探索深度学习中的泛化。载于《神经信息处理系统进展》，第30卷，2017年。Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. 将蒸馏作为抵御针对深度神经网络的对抗扰动的一种防御手段。载于 SP，2016年。Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. 免费的对抗训练！《神经信息处理系统进展》，第32卷，2019年。Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. 神经网络的有趣特性。arXiv 预印本 arXiv:1312.6199，2013年。Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. 鲁棒性可能与准确率相悖。arXiv 预印本 arXiv:1805.12152，2018年。Wang, H. and Wang, Y. Generalist：解耦自然泛化与鲁棒泛化。载于计算机视觉与模式识别会议，2023年。Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.-J., and Kolter, J. Z. Beta-CROWN：通过逐神经元分割约束进行高效界传播，以实现完整和不完整的神经网络鲁棒性验证，2021年。Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. 对抗训练的收敛性与鲁棒性。载于国际机器学习大会，2019年。Wei, Z., Wang, Y., Guo, Y., and Wang, Y. CFA：类别校准公平对抗训练。载于计算机视觉与模式识别会议，2023a。Wei, Z., Zhang, X., Zhang, Y., and Sun, M. 面向自然语言任务的循环神经网络加权自动机提取与解释，2023b。Wu, D., Xia, S.-T., and Wang, Y. 对抗权重扰动有助于鲁棒泛化。载于神经信息处理系统大会，2020年。Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. 通过特征去噪提升对抗鲁棒性。载于计算机视觉与模式识别会议，2019年。Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. 追求鲁棒还是公平：迈向对抗训练中的公平性。载于国际机器学习大会，2021年。Yu, C., Han, B., Gong, M., Shen, L., Ge, S., Du, B., and Liu, T. 用于对抗训练的鲁棒权重扰动。arXiv 预印本 arXiv:2205.14826，2022年a。Yu, C., Han, B., Gong, M., Shen, L., Ge, S., Du, B., and Liu, T. 用于对抗训练的鲁棒权重扰动。arXiv 预印本 arXiv:2205.14826，2022年b。Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L. 和 Jordan, M. 鲁棒性与准确率之间具有理论依据的权衡。载于《国际机器学习会议》，第 7472–7482 页。PMLR，2019年。Zhang, Y., Wei, Z., Zhang, X. 和 Sun, M. 使用 z3 对前馈神经网络的全局鲁棒性进行形式化建模与验证，2023年。Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B. 和 Tao, D. 使用 Fisher 掩码改进尖锐度感知最小化，以提升语言模型的泛化能力。arXiv 预印本 arXiv:2210.05497，2022年。

A. 证明

A.1. 定理 1 的证明

证明。由于对称性，我们只需计算 $y=1$ 的情况，不失一般性。根据该分布，我们可以轻松推导出 $x_2+\cdots+x_{d+1}\sim\mathcal{N}(\eta d,d)$。

Thus

$$\begin{aligned}&\underset{\boldsymbol{w}}{\text{argmax}}\ \mathbb{E}_{\boldsymbol{x}.y\sim\mathcal{D}}\mathbf{1}(f_{\boldsymbol{w}}(\boldsymbol{x})=y)\\=&\underset{\boldsymbol{w}}{\text{argmax}}\ p\Pr(x_2+\cdots+x_{d+1}>-w_1)+(1-p)\Pr(x_2+\cdots+x_{d+1}>w_1)\\=&\underset{w_1}{\text{argmax}}\ \frac{p}{\sqrt{2\pi d}}\int_{-w_1}^{\infty}e^{-(t-\eta d)^2/2d}dt+\frac{1-p}{\sqrt{2\pi d}}\int_{w_1}^{\infty}e^{-(t-\eta d)^2/2d}dt\\:=&\underset{w_1}{\text{argmax}}\ u(w_1).\end{aligned}\tag{14}$$

因此，最优参数 w_1 可由 $du/dw_1=0.$ 推导得到。其导数为

$$\frac{du}{dw_1}=\frac{p}{\sqrt{2\pi d}}e^{-(w_1+\eta d)^2/2d}-\frac{1-p}{\sqrt{2\pi d}}e^{-(w_1-\eta d)^2/2d}=0.\tag{15}$$

求解后可得最优值

$$w_1^*=\frac{\ln p-\ln(1-p)}{2\eta}.\quad\square\tag{16}$$

A.2. 定理 2 的证明

证明。由于 x_1 已被选为 ±1，对 x_1 的扰动不会对其产生影响，因此我们可以忽略它。因此，为了通过偏置 $\boldsymbol{\delta}$ 攻击分类器，使准确率尽可能低，一个直观的想法是设置 $\boldsymbol{\delta}$，使 $x_i(i=2,\cdots,d+1)$ 的期望最小化，从而降低标准准确率。事实上，期望准确率关于每个 $\delta_i,i=2,\cdots,d+1.$ 都是单调递增的。因此，对于任意 $w>0$，选择 $\delta=(0,-\epsilon,\cdots,-\epsilon)$ 都可以作为最佳对抗攻击向量。在这种情况下，这等于 $x_i'(i=2,\cdots,d+1)\sim\mathcal{N}(\eta-\epsilon,d)$。因此，类似于公式 (14)，我们可以推导出训练准确率，即】【。

$$v(w)=p\Phi((w+(\eta-\epsilon)d)/\sqrt{d})+(1-p)\Phi((-w+(\eta-\epsilon)d)/\sqrt{d}).\tag{17}$$

这里 $\Phi(\cdot)$ 是标准正态分布的累积分布函数。现在我们只需求解方程$dv/dw=0$。经过简单计算可得

$$p\exp(-(w_1+(\eta-\epsilon)d)^2/2d)/\sqrt{2\pi d}=(1-p)\exp(-(w_1-(\eta-\epsilon)d)^2/2d)/\sqrt{2\pi d}.\tag{18}$$

求解该方程，最终得到最优值

$$w_1^{AT}=\frac{\ln p-\ln(1-p)}{2(\eta-\epsilon)}.\qquad\square$$

A.3. 定理 3 的证明

证明。定义期望干净准确率函数

$$u(w)=\frac{p}{\sqrt{d}}\Phi((w+\eta d)/\sqrt{d})+\frac{(1-p)}{\sqrt{d}}\Phi((-w+\eta d)/\sqrt{d}),\tag{19}$$

其中，$\Phi(\cdot)$ 是标准正态分布的累积分布函数，且 $w\in\mathbb{R}$。其导数为

$$\begin{aligned}\frac{du(w)}{dw}=&\frac{p}{\sqrt{2\pi d}}e^{-(w+\eta d)^2/2d}-\frac{1-p}{\sqrt{2\pi d}}e^{-(w-\eta d)^2/2d}\\=&\frac{1}{\sqrt{2\pi d}}e^{-(w^2+\eta^2d^2)/d}(pe^{-w\eta}-(1-p)e^{w\eta})\end{aligned}\tag{20}$$

根据标准训练的定理 4.1，我们知道 $u(w)$ 只有一个全局最小值 w_1^*。因此，根据（20）我们知道，当 $w > w_1^*$ 且 $\frac{du}{(w)}_w > 0$ 时，$\frac{du}{(w)}_w < 0$；当 $w < w_1^*$ 时，$\frac{du}{(w)}_w < 0$。

在 SAM 算法中，当我们设定 $\lambda = 0$ 且给定 ϵ 时，我们知道

$$w_1^{SAM} = \operatorname*{argmax}_w \min_{\delta \in [-\epsilon, \epsilon]} u(w + \delta). \tag{21}$$

我们很容易知道，

$$\min_{\delta \in [-\epsilon, \epsilon]} u(w_1^{SAM} + \delta) = \min\{u(w_1^{SAM} - \epsilon), u(w_1^{SAM} + \epsilon)\}. \tag{22}$$

如果 $u(w_1^{SAM} - \epsilon) > u(w_1^{SAM} + \epsilon)$，那么 $w_1^{SAM} + \epsilon > w_1^*$。由于 $\frac{du}{(w)}_w$ 连续且局部有界，因此存在 $\delta_0 > 0$ 使得 $u(w_1^{SAM} - \epsilon - \delta_0) > u(w_1^{SAM} + \epsilon)$ 且 $u(w_1^{SAM} + \epsilon - \delta_0) > u(w_1^{SAM} + \epsilon)$。因此，我们有

$$\min\{u(w_1^{SAM} - \epsilon - \delta_0), u(w_1^{SAM} + \epsilon - \delta_0)\} > \min\{u(w_1^{SAM} - \epsilon), u(w_1^{SAM} + \epsilon)\}. \tag{23}$$

因此

$$\min_{\delta \in [-\epsilon, \epsilon]} u((w_1^{SAM} - \delta_0) + \delta) > \min_{\delta \in [-\epsilon, \epsilon]} u(w_1^{SAM} + \delta), \tag{24}$$

这意味着 w_1^{SAM}并非我们想要的最优值。类似地，我们可以证明 $u(w_1^{SAM} - \epsilon) < u(w_1^{SAM} + \epsilon)$ 不成立。因此，$u(w_1^{SAM} - \epsilon) = u(w_1^{SAM} + \epsilon)$。

由此，我们知道，

$$\int_{w_1^{SAM} - \epsilon}^{w_1^*} \frac{du(w)}{dw} dw = -\int_{w_1^*}^{w_1^{SAM} + \epsilon} \frac{du(w)}{dw} dw. \tag{25}$$

使用（20），我们得到

$$\begin{aligned} &\int_{w_1^{SAM} - \epsilon}^{w_1^*} e^{-w^2}(pe^{-w\eta} - (1-p)e^{w\eta}) dw \\ =&-\int_{w_1^*}^{w_1^{SAM} + \epsilon} e^{-w^2}(pe^{-w\eta} - (1-p)e^{w\eta}) dw. \end{aligned} \tag{26}$$

如果 $w_1^{SAM} \le w_1^*$，定义 $h = w_1^{SAM} + \epsilon - w_1^*$。因此，$h \le \epsilon \le -w_1^{SAM} + \epsilon + w_1^*$。在（26）中，我们有

$$\begin{aligned} &\int_{-h}^{0} e^{-(w_1^* + w)^2}(pe^{-(w_1^* + w)\eta} - (1-p)e^{(w_1^* + w)\eta}) dw \\ \le&\int_{w_1^{SAM} - \epsilon}^{w_1^*} e^{-w^2}(pe^{-w\eta} - (1-p)e^{w\eta}) dw \\ =&-\int_{w_1^*}^{w_1^{SAM} + \epsilon} e^{-w^2}(pe^{-w\eta} - (1-p)e^{w\eta}) dw \\ =&-\int_{0}^{h} e^{-(w_1^* + w)^2}(pe^{-(w_1^* + w)\eta} - (1-p)e^{(w_1^* + w)\eta}) dw \\ =&\int_{-h}^{0} e^{-(w_1^* - w)^2}(-pe^{-(w_1^* - w)\eta} + (1-p)e^{(w_1^* - w)\eta}) dw. \end{aligned} \tag{27}$$

由于 $w_1^* > 0$，我们可以得知 $e^{-(w_1^* + w)^2} > e^{-(w_1^* - w)^2} > 0$ 对于 $w \in [-h, 0)$。

对于函数 $r(v) := (pe^{-v\eta} - (1-p)e^{v\eta})$，它单调递减且有一个零点 w_1^*，因此 $pe^{-(w_1^* + w)\eta} - (1-p)e^{(w_1^* + w)\eta} > 0$ 且 $-pe^{-(w_1^* - w)\eta} + (1-p)e^{(w_1^* - w)\eta} > 0$，对于 $w \in [-h, 0]$。并且

$$\begin{aligned} &pe^{-(w_1^* + w)\eta} - (1-p)e^{(w_1^* + w)\eta} - (-pe^{-(w_1^* - w)\eta} + (1-p)e^{(w_1^* - w)\eta}) \\ =&(pe^{-w_1^*\eta} - (1-p)e^{w_1^*\eta})(e^{w\eta} + e^{-w\eta}) \\ =&0. \end{aligned} \tag{28}$$

因此， $\forall w \in [-h, 0)$,

$$e^{-(w_1^*+w)^2}(pe^{-(w_1^*+w)\eta} - (1-p)e^{(w_1^*+w)\eta}) > e^{-(w_1^*-w)^2}(-pe^{-(w_1^*-w)\eta} + (1-p)e^{(w_1^*-w)\eta}). \tag{29}$$

结合（27），我们得出矛盾。

因此， $w_1^{SAM} > w_1^*$。证明至此结束。 □

A.4. 定理4的证明

证明。 我们从（26）继续证明。由于我们已经证明 $w_1^{SAM} > w_1^*$，因此假设 $h = w_1^* - w_1^{SAM} + \epsilon < \epsilon$。于是，我们得到

$$\begin{aligned} 0 = &\int_{-h}^{0} e^{-(w_1^*+w)^2}(pe^{-(w_1^*+w)\eta} - (1-p)e^{(w_1^*+w)\eta})dw - \int_{-h}^{0} e^{-(w_1^*-w)^2}(-pe^{-(w_1^*-w)\eta} + (1-p)e^{(w_1^*-w)\eta})dw \\ &- \int_{h}^{2\epsilon-h} e^{-(w_1^*+w)^2}(-pe^{-(w_1^*+w)\eta} + (1-p)e^{(w_1^*+w)\eta})dw. \end{aligned} \tag{30}$$

由于我们只关注 h，因此考虑在计算中省略 $o(h^3)$ 项。具体而言， $o(w^2)$ 积分符号 $\int_{-h}^{0}$ 中的项可以省略，而 $o(w)$ 或 $o(h)$ 项对于 w 和 h 积分符号 $\int_{0}^{2\epsilon-2h}$ 中的项也可以省略。[1]结合（28）中的证明和 w_1^*的定义，并舍弃高阶项，我们可以计算右端如下。

$$\begin{aligned} RHS = &\int_{-h}^{0} (e^{-(w_1^*+w)^2} - e^{-(w_1^*-w)^2}) \times (pe^{-(w_1^*+w)\eta} - (1-p)e^{(w_1^*+w)\eta})dw \\ &+ \int_{h}^{2\epsilon-h} e^{-(w_1^*+w)^2}(-pe^{-(w_1^*+w)\eta} + (1-p)e^{(w_1^*+w)\eta})dw \\ \approx &\int_{-h}^{0} e^{-(w_1^*)^2}(1 - 2w_1^*w - 1 - 2w_1^*w + o(w)) \times (pe^{-w_1^*\eta}(1-w\eta) - (1-p)e^{w_1^*\eta}(1+w\eta))dw \\ &- \int_{0}^{2\epsilon-2h} e^{-(w_1^*)^2}(1 - 2w_1^*h) \times (-pe^{-w_1^*\eta}(1-h\eta) + (1-p)e^{w_1^*\eta}(1+h\eta))dw \\ \approx &4e^{-(w_1^*)^2}w_1^* \int_{-h}^{0} w^2\eta(-pe^{-w_1^*\eta} - (1-p)e^{w_1^*\eta})dw - e^{-(w_1^*)^2} \int_{0}^{2\epsilon-2h} \eta h(pe^{-w_1^*\eta} + (1-p)e^{w_1^*\eta})dw \\ \approx &\frac{4}{3}e^{-(w_1^*)^2}w_1^*(pe^{-w_1^*\eta} + (1-p)e^{w_1^*\eta})h^3 - 2e^{-(w_1^*)^2}(\epsilon - h)\eta h(pe^{-w_1^*\eta} + (1-p)e^{w_1^*\eta}) \\ \approx &\frac{2}{3}e^{-(w_1^*)^2}(pe^{-w_1^*\eta} + (1-p)e^{w_1^*\eta})\eta(2w_1^*h^2 - 3(\epsilon - h)). \end{aligned} \tag{31}$$

由于 $RHS = 0$，通过求解（31）中的最后一个等式，我们得到

$$\frac{\epsilon - h}{h} = \frac{2}{3}w_1^*h = o(1). \tag{32}$$

因此，上述计算以及计算中舍弃高阶项的做法是有效的。由于 $h \to \epsilon$，我们有

$$\epsilon - h = \frac{2}{3}w_1^*h^2 \approx \frac{2}{3}w_1^*\epsilon^2. \tag{33}$$

因此，我们最终得出结论： $w_1^{SAM} \approx w_1^* + \frac{2}{3}w_1^*\epsilon^2$。 □

[1]消融的有效性　从计算结果可以看出，舍弃这些高阶项是合理的，结果表明 $\frac{h-\epsilon}{h} \to 0$.

A.5. 定理5的证明

证明。当两种方法得出相同的最优值 w_1' 时，记标准训练的最优参数为 w_1^* =(ln p − ln(1 − p))/2η，则有

$$w_1^*(1+\frac{2}{3}\epsilon_{SAM}^2)\approx w_1', \tag{34}$$

并且

$$\frac{\eta}{\eta-\epsilon_{AT}}w_1^*=w_1'. \tag{35}$$

因此，我们有

$$\frac{\eta}{\eta-\epsilon_{AT}}\approx 1+\frac{2}{3}\epsilon_{SAM}^2 \tag{36}$$

解这个方程，得到关系式

$$\frac{2}{3}\epsilon_{SAM}^2\epsilon_{AT}+\epsilon_{AT}\approx\frac{2}{3}\eta\epsilon_{SAM}^2 \tag{37}$$

将 b 相除　两边同时乘以 $\epsilon_{AT}\epsilon_{SAM}^2$，即可直接推出定理中的关系。证明至此结束。　f.　□