arXiv:2302.02713v5 [cs.LG] 6 Nov 2023

趋平坦贝叶斯神经网络

Van-Anh Nguyen[1] **Tung-Long Vuong**[1,2] **Hoang Phan**[2,3] **Thanh-Toan Do**[1] **Dinh Phung** [1,2] **Trung Le** [1]

[1]莫纳什大学数据科学与人工智能系，澳大利亚[2]VinAI，越南[3]纽约大学，美国{van-anh.nguyen, tung-long.vuong, toan.do, dinh.phung, trunglm}@monash.edu hvp2011@nyu.edu

摘要

贝叶斯神经网络（BNN）通过对模型参数施加先验分布，并基于观测数据推断后验分布，为深度学习模型提供概率解释。从后验分布中采样得到的模型可用于生成集成预测并量化预测不确定性。众所周知，尖锐度较低的深度学习模型具有更强的泛化能力。然而，现有的后验推断方法在公式设计上并未考虑尖锐度/平坦度，这可能导致从这些方法中采样得到的模型具有较高的尖锐度。本文针对尖锐度感知后验，提出了相关理论、贝叶斯设定以及变分推断方法。具体而言，从我们提出的尖锐度感知后验中采样得到的模型，以及用于估计该尖锐度感知后验的最优近似后验，都具有更好的平坦度，因此可能具备更强的泛化能力。我们利用最先进的贝叶斯神经网络中的尖锐度感知后验开展实验，结果表明，趋平坦对应方法在所有关注的指标上均优于其基线方法。

1 引言

贝叶斯神经网络（BNN）提供了一种以概率方式解释深度学习模型的方法。具体而言，我们为模型参数设定先验分布，再根据观测数据推断模型参数的后验分布。这使我们不仅能够进行预测，还能量化预测不确定性，这对许多现实应用都很有用。为了从复杂的后验分布中采样深度学习模型，通常会使用哈密顿蒙特卡洛（HMC） [41], 随机梯度哈密顿蒙特卡洛（SGHMC） [10], 随机梯度朗之万动力学（SGLD） [59], 和斯坦变分梯度下降（SVGD） [36]等先进的粒子采样方法。然而，这些方法的计算开销可能很大，尤其是在需要采样许多模型以构建更优集成时。

为减轻这一计算负担，并能够从后验分布中采样多个深度学习模型，变分推断方法采用近似后验来估计真实后验。这些方法使用属于足够丰富的分布族的近似后验，这些分布族既经济又便于采样。然而，变分推断领域的早期先驱工作，例如 [21, 5, 33], , 假设近似后验是完全因子化的分布，也称为平均场变分推断。这种方法未能考虑神经网络随机权重之间强烈的统计依赖关系，因而限制了其捕捉真实后验复杂结构并估计真实模型不确定性的能力。为了解决这一问题，后续研究尝试提供表达能力更强的后验近似

第37届神经信息处理系统大会（NeurIPS 2023）。

表达能力 [62, 52, 53, 55, 20, 45, 56, 30, 48]。这些方法旨在提高后验近似的准确性，并实现更有效的不确定性量化。
在标准深度网络训练中，人们发现，平坦极小值点能够提升模型的泛化能力。这是因为它们能够找到更宽的局部极小值点，从而更能适应训练集和测试集之间的变化。包括 [27, 47, 15], 在内的多项研究为这一原则提供了证据。然而，现有贝叶斯神经网络（BNN）所用的后验分布，在模型构建时并未考虑由其生成的模型的尖锐度/平坦度。因此，采样得到的模型可能位于尖锐度高、平坦度低的区域，导致泛化能力较差。此外，在变分推断方法中，使用近似后验来估计这些非尖锐度感知后验，可能导致从相应的最优近似后验中采样得到的模型缺乏对尖锐度/平坦度的感知，进而导致泛化能力较差。

本文旨在提出一种用于学习贝叶斯神经网络的尖锐度感知后验分布，从而采样得到平坦度较高、泛化能力更强的模型。为此，我们针对所提出的后验分布设计了贝叶斯设定和变分推断方法。通过估计最优近似后验，我们可以生成更平坦的模型，从而提升泛化能力。我们的方法3.1如下：我们证明，标准后验分布是一个优化问题的最优解；该问题通过平衡由近似后验采样得到的模型所产生的经验损失（用于拟合训练集）与鼓励近似后验保持简单的 Kullback-Leibler（KL）散度来实现这一点。基于这一见解，我们在定理 3.2 中以整个数据标签分布上的总体损失替代近似后验产生的经验损失，以提升泛化能力。受尖锐度感知最小化的启发 [16]，我们在定理 3.2中推导出总体损失的一个上界，进而在定理 3.3中构建尖锐度感知后验分布。最后，我们为尖锐度感知后验分布设计了贝叶斯设定和变分方法。总而言之，本文的贡献可概括如下：

- 我们针对尖锐度感知后验提出并发展了相关理论、贝叶斯设定以及变分推断方法。该后验使我们能够采样一组平坦模型，从而提升模型的泛化能力。我们注意到，SAM [16] 仅考虑单个模型的尖锐度，而我们是首个研究模型分布尖锐度概念及理论的工作，$\mathbb{Q}$ 即模型分布3.2 的支撑集包含无限多个模型，因此定理的证明极具挑战性，同时也十分精妙且复杂 $\mathbb{Q}$。
- 我们利用尖锐度感知后验，结合最先进且广为人知的贝叶斯神经网络开展了广泛实验，包括具有近似高斯分布的贝叶斯神经网络 *[33]*, 采用随机梯度朗之万动力学（*SGLD*）的贝叶斯神经网络 *[59]*, *MC-Dropout* *[18]*, 贝叶斯深度集成 *[35]*, 和*SWAG* *[39]* 以证明，平坦化方法在所有相关指标上始终优于相应方法，包括集成准确率、期望校准误差（ECE）和负对数似然（NLL）。

2 相关工作

2.1 贝叶斯神经网络

马尔可夫链蒙特卡洛（MCMC）： 这种方法允许我们从后验分布中采样多个模型，并因哈密顿蒙特卡洛（HMC） [41]而广为人知，常用于神经网络的推断。然而，HMC 需要估计全梯度，这对于神经网络而言计算开销很大。为了使 HMC 框架切实可用，随机梯度哈密顿蒙特卡洛（SGHMC） [10]允许在贝叶斯推断中使用随机梯度，这对于提升可扩展性和探索解空间都至关重要。另一种方法是随机梯度朗之万动力学（SGLD） [59] ，它在随机梯度设定下采用一阶朗之万动力学。此外，斯坦变分梯度下降（SVGD） [36] 通过维护一组粒子，逐步逼近后验分布。从理论上讲，当步长趋近于零时，SGHMC、SGLD 和 SVGD 都渐近地从后验分布中采样。

变分推断：该方法通过最大化变分下界，利用某个族中的近似后验分布来估计真实后验分布。 [21] 建议拟合

神经网络权重上的高斯变分后验近似，后来在 [32, 33, 5], 中得到了推广，该方法使用重参数化技巧来训练深度潜变量模型。为获得表达能力更强的后验近似，已有许多深入研究。值得注意的是，[38] 将权重矩阵作为整体，通过矩阵变量高斯分布 [22] 并基于这种参数化对后验分布进行近似。之后的若干研究考察了这种分布，以探索变分高斯后验的不同结构化表示，例如克罗内克因子化 [60, 52, 53], k 绑定分布 [55],非中心化或秩一参数化[20, 14]。表示高斯后验真实协方差矩阵的另一种方法是采用低秩近似 [45, 56, 30, 39]。

Dropout 变分推断：这种方法利用 Dropout 来表征近似后验。通常，[18] 和 [33] 基于这一原理提出贝叶斯 Dropout 推断方法，例如蒙特卡洛 Dropout 和变分 Dropout。Concrete Dropout [19] 将这一思路扩展为优化 Dropout 概率。变分结构化 Dropout [43] 则采用 Householder 变换，在变分 Dropout 方法中学习乘性高斯噪声的结构化表示。

2.2 平坦极小值

研究发现，平坦极小值点有助于提升神经网络的泛化能力。这是因为它们能使模型找到更宽的局部极小值，从而增强模型对训练集和测试集之间分布偏移的鲁棒性 [27, 47, 15, 44]。许多研究（尤其是 [23, 42,12, 17]）从理论和实证角度探讨了泛化能力与极小值宽度之间的关系。此外，[46, 9, 29, 25, 16, 44]提出了多种寻求平坦极小值的方法。通常，[29, 26, 58] 会研究批量大小、学习率、梯度协方差和 Dropout 等不同训练因素对所找到的极小值平坦度的影响。此外，还有一些方法通过在损失函数中添加正则化项来寻求宽的局部极小值 [46, 62, 61,9]。这类正则化项包括 Softmax 输出低熵惩罚 [46] 和蒸馏损失 [62, 61]。

SAM 是一种通过寻找平坦区域来最小化当前模型周围最坏情况损失的方法。与以往方法相比，它具有良好的可扩展性和有效性，因此近期备受关注 [16, 57]。SAM 已广泛应用于各种领域和任务，例如元学习双层优化 [1], 、联邦学习 [51], 和多任务学习 [50], ，并在这些任务中取得了更快的收敛速率，且提出了泛化界。SAM 还在视觉模型 [11], 、语言模型 [3], 、领域泛化 [8], 和多任务学习 [50]中展现了泛化能力。一些研究者尝试利用 SAM 的几何特性 [34, 31], 、进一步最小化替代间隙 [63], 以及缩短其训练时间 [13, 37]来改进 SAM。关于 SAM 的行为，[28]通过实证研究比较了 SAM [16] 与 SWA [24], [40] 得到的尖锐度差异；[44] 则表明，SAM 是具有正态后验分布的标准贝叶斯推断的一种最优贝叶斯松弛，而 [4, 49]证明分布鲁棒性是 SAM 的概率扩展。

3 提出的框架

接下来，我们将介绍所提出的锐度感知后验分布的技术细节。具体而言，第 3.1 节介绍了锐度感知后验分布的问题设定和动机。第3.2 节专门阐述我们的理论推导，而第 3.3节则用于说明锐度感知后验分布的贝叶斯设定和变分推断方法。

3.1 问题设定与动机

我们的目标是开发尖锐度感知贝叶斯神经网络（SA-BNN）。考虑一族神经网络 $f_\theta(x)$，其具有 $\theta \in \Theta$，以及一个训练集 $\mathcal{S} = \{(x_1, y_1) ...,(x_n, y_n)\}$ 其中 $(x_i, y_i) \sim \mathcal{D}$。我们希望学习一个后验分布 $\mathbb{Q}_S^{SA}$，其密度函数为 $q^{SA}(\theta|\mathcal{S})$，使得任何模型 $\theta \sim \mathbb{Q}_S^{SA}$在训练集 $\mathcal{S}$上进行预测时都能感知尖锐度。

我们从标准后验出发

$$q(\theta \mid \mathcal{S}) \propto \prod_{i=1}^{n} p(y_i \mid x_i, \mathcal{S}, \theta) p(\theta),$$

其中先验分布 $\mathbb{P}$ 的密度函数为 $p(\theta)$，似然的形式为

$$p\left(y \mid x, \mathcal{S}, \theta\right) \propto \exp\left\{-\frac{\lambda}{|\mathcal{S}|}\ell\left(f_\theta(x), y\right)\right\} = \exp\left\{-\frac{\lambda}{n}\ell\left(f_\theta(x), y\right)\right\}$$

采用损失函数 ℓ。标准后验分布 $\mathbb{Q}_{\mathcal{S}}$ 的密度函数定义为

$$q(\theta \mid \mathcal{S}) \propto \exp\left\{-\frac{\lambda}{n}\sum_{i=1}^{n}\ell\left(f_\theta\left(x_i\right), y_i\right)\right\} p(\theta), \tag{1}$$

其中，$\lambda \geq 0$ 是正则化参数。

我们将总体损失和经验损失定义如下：

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(f_\theta(x), y\right)\right].$$

$$\mathcal{L}_{\mathcal{S}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{S}}\left[\ell\left(f_\theta(x), y\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\ell\left(f_\theta\left(x_i\right), y_i\right).$$

基本上，总体损失定义为整个数据标签分布 $\mathcal{D}$上的期望损失，而经验损失则定义为特定训练集 $\mathcal{S}$上的经验损失。

公式（1）中的标准后验分布可改写为

$$q(\theta \mid \mathcal{S}) \propto \exp\left\{-\lambda\mathcal{L}_{\mathcal{S}}(\theta)\right\} p(\theta). \tag{2}$$

给定模型参数 $\theta \in \Theta$上的一个分布 $\mathbb{Q}$，其密度函数为 $q(\theta)$，我们将该模型分布 $\mathbb{Q}$ 上的经验损失和总体损失定义为

$$\mathcal{L}_{\mathcal{S}}(\mathbb{Q}) = \int_{\Theta}\mathcal{L}_{\mathcal{S}}(\theta)\, d\mathbb{Q}(\theta) = \int_{\Theta}\mathcal{L}_{\mathcal{S}}(\theta)\, q(\theta)\, d\theta.$$

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) = \int_{\Theta}\mathcal{L}_{\mathcal{D}}(\theta)\, d\mathbb{Q}(\theta) = \int_{\Theta}\mathcal{L}_{\mathcal{D}}(\theta)\, q(\theta)\, d\theta.$$

具体而言，模型分布 $\mathbb{Q}$ 上的总体损失定义为从该分布中采样的模型所产生的总体损失的期望，而模型分布 $\mathbb{Q}$ 上的经验损失定义为从该分布中采样的模型所产生的经验损失的期望。

3.2 我们的理论推导

现在我们给出尖锐度感知后验的理论推导，其证明可见补充材料。受标准后验的吉布斯形式 $\mathbb{Q}_{\mathcal{S}}$（式2）的启发，我们建立以下定理，以将标准后验 $\mathbb{Q}_{\mathcal{S}}$ 与密度 $q(\theta \mid \mathcal{S})$ 和经验损失 $\mathcal{L}_{\mathcal{S}}(\mathbb{Q})$ [7, 2]联系起来。

定理 3.1。考虑以下优化问题

$$\min_{\mathbb{Q}<<\mathbb{P}}\left\{\lambda\mathcal{L}_S(\mathbb{Q}) + KL(\mathbb{Q}, \mathbb{P})\right\}, \tag{3}$$

其中，我们在相对于 $\mathbb{P}$ 绝对连续的 $\mathbb{Q}$ 上进行搜索，并且 $KL(\cdot, \cdot)$ 是 *Kullback-Leibler* 散度。该优化问题具有闭式最优解 $\mathbb{Q}^*$，其密度为

$$q^*(\theta) \propto \exp\left\{-\lambda\mathcal{L}_{\mathcal{S}}(\theta)\right\} p(\theta),$$

这恰好是密度为 $q(\theta \mid \mathcal{S})$ 的标准后验 $\mathbb{Q}_{\mathcal{S}}$。

定理 3.1 揭示了我们需要找到后验分布 $\mathbb{Q}_{\mathcal{S}}$，在优化其经验损失 $\mathcal{L}_S(\mathbb{Q})$ 与通过 $KL(\mathbb{Q}, \mathbb{P})$保持简洁之间取得平衡。然而，仅最小化经验损失 $\mathcal{L}_S(\mathbb{Q})$ 只能确保对 $\mathcal{S}$中的训练样本做出正确预测，因此可能会出现过拟合。因此，理想的做法是用总体损失替代经验损失，以应对过拟合。

为缓解过拟合，在（8）中，我们用总体损失替代经验损失，并求解以下优化问题（OP）：

$$\min_{\mathbb{Q}<<\mathbb{P}}\left\{\lambda\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) + KL(\mathbb{Q}, \mathbb{P})\right\}. \tag{4}$$

值得注意的是，求解式 (4) 中的优化问题（OP）通常难以处理。为了使其可处理，我们找到了其上界，该上界与模型上的分布的尖锐度有关，如下述定理所示。 $\mathbb{Q}$

定理 3.2。 假设 Θ 是一个紧集。在一些温和条件下，给定任意 $\delta \in [0;1]$,，对于 $\mathcal{S} \sim \mathcal{D}^n$ 的选择，以至少 $1-\delta$ 的概率，对任意分布 $\mathbb{Q}$，我们有

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) \leq \mathbb{E}_{\theta \sim \mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta')\right] + f\left(\max_{\theta \in \Theta}\|\theta\|^2, n\right),$$

*whe*对于 f，它是关于第一个变量的非递减函数，并且随着训练规模的增大而趋近于0 e $^{n\,a}$趋近于 ∞。

我们注意到，定理 3.2 的证明并非尖锐度感知最小化的简单推广，因为我们需要处理分布 $\mathbb{Q}$上的一般损失和经验损失。为了明确模型 $\mathbb{Q}$上关于分布的尖锐度，我们将该不等式的上界改写为

$$\mathbb{E}_{\theta \sim \mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta') - \mathcal{L}_{\mathcal{S}}(\theta)\right] + \mathcal{L}_{\mathcal{S}}(\mathbb{Q}) + f\left(\max_{\theta \in \Theta}\|\theta\|^2, n\right),$$

其中，第一项 $\mathbb{E}_{\theta \sim \mathbb{Q}}\left[\max_{\theta':\ \|\theta'-\theta\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta') - \mathcal{L}_{\mathcal{S}}(\theta)\right]$ 可视为模型空间 $\mathbb{Q}$ 上关于分布的尖锐度，而最后一项 $f\left(\max_{\theta \in \Theta}\|\theta\|^2,\ n\right)$ 是一个常数。

此外，受定理 3.2的启发，我们提出求解以下优化问题（OP），它构成了所需优化问题在（4）的一个上界

$$\min_{\mathbb{Q}<<\mathbb{P}}\left\{\lambda \mathbb{E}_{\theta \sim \mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta')\right] + KL(\mathbb{Q}, \mathbb{P})\right\}. \tag{5}$$

下面的定理刻画了（5）中的优化问题的最优解。

定理 3.3。 （5）中的优化问题的最优解是尖锐度感知后验分布 $\mathbb{Q}_S^{SA}$，其密度函数为 q^{SA}（$\theta|\mathcal{S}$）：

$$q^{SA}(\theta|\mathcal{S}) \propto \exp\left\{-\lambda \max_{\theta':\|\theta'-\theta\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta')\right\} p(\theta) = \exp\left\{-\lambda \mathcal{L}_{\mathcal{S}}(s(\theta))\right\} p(\theta),$$

其中，我们定义了 s（θ） $= \underset{\theta':\|\theta'-\theta\| \leq \rho}{argmax}$ $\mathcal{L}_{\mathcal{S}}$ （θ'）。

定理 3.3 给出了尖锐度感知后验分布的闭式表达 $\mathbb{Q}_S^{SA}$，其密度函数为 q^{SA}（$\theta|\mathcal{S}$）。基于这一刻画，接下来我们将介绍尖锐度感知贝叶斯设定，并阐明其变分方法。

3.3 尖锐度感知贝叶斯设定及其变分方法

贝叶斯设定： 为了推广针对尖锐度感知后验分布 $\mathbb{Q}_S^{SA}$ 的贝叶斯设定，我们考察尖锐度感知似然

$$p^{SA}(y \mid x, \mathcal{S}, \theta) \propto \exp\left\{-\frac{\lambda}{|\mathcal{S}|}\ell\left(f_{s(\theta)}(x), y\right)\right\} = \exp\left\{-\frac{\lambda}{n}\ell\left(f_{s(\theta)}(x), y\right)\right\},$$

其中， $s(\theta) = \underset{\theta':\|\theta'-\theta\| \leq \rho}{\text{argmax}}$ $\mathcal{L}_{\mathcal{S}}(\theta')$。

借助这一预定义的尖锐度感知似然，我们可以利用密度函数 $q^{SA}(\theta|\mathcal{S})$ 恢复尖锐度感知后验分布 $\mathbb{Q}_S^{SA}$:

$$q^{SA}(\theta|\mathcal{S}) \propto \prod_{i=1}^{n} p^{SA}(y_i \mid x_i, \mathcal{S}, \theta)\, p(\theta).$$

尖锐度感知后验分布的变分推断： 现在，我们针对尖锐度感知后验分布构建变分推断。令 $X = [x_1, ..., x_n]$ and

$Y=[y_1,...,y_n]$。考虑近似后验族 $\{q_\phi(\theta):\phi\in\Phi\}$，我们有

$$\begin{aligned}\log p^{SA}(Y\mid X,\mathcal{S}) &= \int_\Theta q_\phi(\theta)\log p^{SA}(Y\mid X,\mathcal{S})\,d\theta \\ &= \int_\Theta q_\phi(\theta)\log\frac{p^{SA}(Y\mid\theta,X,\mathcal{S})\,p(\theta)}{q_\phi(\theta)}\frac{q_\phi(\theta)}{q^{SA}(\theta|\mathcal{S})}d\theta \\ &= \mathbb{E}_{q_\phi(\theta)}\left[\sum_{i=1}^n\log p^{SA}(y_i\mid x_i,\mathcal{S},\theta)\right]-KL(q_\phi,p)+KL\left(q_\phi,q^{SA}\right).\end{aligned}$$

显然，为了最大限度地缩小间隙 $KL\left(q_\phi,\ q^{SA}\right)$，我们需要最大化以下下界：

$$\max_{q_\phi}\left\{\mathbb{E}_{q_\phi(\theta)}\left[\sum_{i=1}^n\log p^{SA}(y_i\mid x_i,\mathcal{S},\theta)\right]-KL(q_\phi,p)\right\},$$

其可等价地改写为

$$\begin{aligned}&\min_{q_\phi}\left\{\lambda\mathbb{E}_{q_\phi(\theta)}\left[\mathcal{L}_\mathcal{S}(s(\theta))\right]+KL(q_\phi,p)\right\}\text{ or}\\ &\min_{q_\phi}\left\{\lambda\mathbb{E}_{q_\phi(\theta)}\left[\max_{\theta':\|\theta'-\theta\|\le\rho}\mathcal{L}_\mathcal{S}(\theta')\right]+KL(q_\phi,p)\right\}.\end{aligned}\tag{6}$$

采用高斯近似后验的变分方法推导： 受基于几何的 SAM 方法 [34, 31], 的启发，我们通过以下距离将几何纳入 SA 变分方法，以定义尖锐度对应的球体：$\|\theta'-\theta\|_{\text{diag}(T_\theta)}=\sqrt{(\theta'-\theta)^T\text{diag}(T_\theta)^{-1}(\theta'-\theta)}$ 作为

$$\min_{q_\phi}\left\{\lambda\mathbb{E}_{q_\phi(\theta)}\left[\max_{\theta':\|\theta'-\theta\|_{\text{diag}(T_\theta)}\le\rho}\mathcal{L}_\mathcal{S}(\theta')\right]+KL(q_\phi,p)\right\}.$$

为进一步阐明这一点，我们将 SA 后验分布应用于贝叶斯神经网络，其中我们对其权重矩阵 $W_i\sim\mathcal{N}\left(\mu_i、\ \sigma_i^2\mathbb{I}\right)$ 、 $i=1$、 ...、 L[1] 施加高斯分布。参数 ϕ 由 μ_i、 σ_i、 $i=1$、 ...、 L组成。对于 $\theta=W_{1:L}\sim q_\phi$，利用重参数化技巧 $W_i=\mu_i+\text{diag}(\sigma_i)\epsilon_i$、 $\epsilon_i\sim\mathcal{N}(0,\mathbb{I})$，并通过搜索 $\theta^{'}=W_1^{'}{}_{:L}$ 以及 $W_i^{'}=\mu_i^{'}+\text{diag}(\sigma_i)\epsilon_i$、 $\epsilon_i\sim\mathcal{N}(0,\mathbb{I})$，约束条件 $\|\theta-\theta'\|_{\text{diag}(T_\theta)}=\|\mu-\mu'\|_{\text{diag}(T_\theta)}$ 中的 $\mu=\mu_{1:L}$ 和 $\mu^{'}=\mu_1^{'}{}_{:L}$。因此，式（6）中的优化问题可写为

$$\min_{\mu,\sigma}\left\{\lambda\mathbb{E}_\epsilon\left[\max_{\|\mu'-\mu\|_{\text{diag}(T_{\mu,\sigma})}\le\rho}\mathcal{L}_\mathcal{S}\left(\left[\mu_i^{'}+\text{diag}(\sigma_i)\epsilon_i\right]_{i=1}^L\right)\right]\right\},\tag{7}$$

其中 $\sigma=\sigma_{1:L}$、 $\epsilon=\epsilon_{1:L}$，并且我们将 $\text{diag}(T_\theta)=\text{diag}(T_{\mu,\ \sigma})$ 定义为几何距离中的 y.

为求解式（7）中的优化问题，我们从标准高斯分布中采样 $\epsilon=\epsilon_{1:L}$，采用单步梯度上升来求得 $\mu^{'}$，并使用 $\mu^{'}$ 处的梯度来更新 μ。具体而言，我们求得 μ' [6]（第 9 章）如下：

$$\mu'=\mu+\rho\frac{\text{diag}(T_{\mu,\sigma})\nabla_\mu\mathcal{L}_\mathcal{S}\left(\left[\mu_i+\text{diag}(\sigma_i)\epsilon_i\right]_{i=1}^L\right)}{\left\|\text{diag}(T_{\mu,\sigma})\nabla_\mu\mathcal{L}_\mathcal{S}\left(\left[\mu_i+\text{diag}(\sigma_i)\epsilon_i\right]_{i=1}^L\right)\right\|}.$$

$\text{diag}(T_{\mu,\sigma})$ 的定义说明了模型权重的重要性等级，即鼓励重要性较高的权重具有更高的尖锐度，这可通过减小损失相对于该权重的绝对偏导数来实现。我们考虑 $\text{diag}(T_{\mu,\sigma})=\mathbb{I}$（即标准 *SA* 贝叶斯神经网络）和 $\text{diag}(T_{\mu,\sigma})=\text{diag}\left(\frac{|\mu|}{\sigma}\right)$（即 几何 *SA* 贝叶斯神经网络）。这里注意，$\vdots$表示逐元素除法。

[1]我们将偏置并入权重矩阵。

表 1：CIFAR-100数据集上的分类得分。每项实验使用不同的随机种子重复三次，并报告均值和标准差。

方法	PreResNet-164 准确率 ↑	PreResNet-164 负对数似然 ↓	PreResNet-164 期望校准误差 ↓	WideResNet28x10 准确率 ↑	WideResNet28x10 负对数似然 ↓	WideResNet28x10 期望校准误差 ↓
变分推断						
蒙特卡洛 Dropout	79.50 ± 0.37	0.9162 ± 0.0103	0.0993 ± 0.0033	82.30 ± 0.19	0.6500 ± 0.0049	0.0574 ± 0.0028
F-MC-Dropout	**81.06 ± 0.44**	**0.7027 ± 0.0049**	**0.0514 ± 0.0047**	**83.24 ± 0.11**	**0.6144 ± 0.0068**	**0.0250 ± 0.0027**
深度集成	82.08 ± 0.42	0.7189 ± 0.0108	0.0334 ± 0.0064	83.04 ± 0.15	0.6958 ± 0.0335	0.0483 ± 0.0017
F-深度集成	**82.54 ± 0.10**	**0.6286 ± 0.0022**	**0.0143 ± 0.0041**	**84.52 ± 0.03**	**0.5644 ± 0.0106**	**0.0191 ± 0.0039**
马尔可夫链蒙特卡洛						
SGLD	80.13 ± 0.01	0.7604 ± 0.0010	0.1161 ± 0.0031	81.38 ± 0.10	0.7123 ± 0.0204	0.0958 ± 0.0004
F-SGLD	**80.82 ± 0.02**	**0.7276 ± 0.0012**	**0.1085 ± 0.0008**	**82.12 ± 0.16**	**0.6722 ± 0.0112**	**0.0820 ± 0.0021**
样本						
SWAG-Diag	80.18 ± 0.50	0.6837 ± 0.0186	**0.0239 ± 0.0047**	82.40 ± 0.09	0.6150 ± 0.0029	0.0322 ± 0.0018
F-SWAG-Diag	**81.01 ± 0.29**	**0.6645 ± 0.0050**	0.0242 ± 0.0039	**83.50 ± 0.29**	**0.5763 ± 0.0120**	**0.0151 ± 0.0020**
SWAG	79.90 ± 0.50	**0.6595 ± 0.0019**	0.0587 ± 0.0048	82.23 ± 0.19	0.6078 ± 0.0006	**0.0113 ± 0.0020**
F-SWAG	**80.93 ± 0.27**	0.6704 ± 0.0049	**0.0350 ± 0.0025**	**83.57 ± 0.26**	**0.5757 ± 0.0136**	0.0196 ± 0.0015

表2：CIFAR-10 数据集上的分类得分。每项实验使用不同的随机种子重复三次，并报告均值和标准差。

方法	PreResNet-164 准确率 ↑	PreResNet-164 负对数似然 ↓	PreResNet-164 期望校准误差 ↓	WideResNet28x10 准确率 ↑	WideResNet28x10 负对数似然 ↓	WideResNet28x10 期望校准误差 ↓
变分推断						
蒙特卡洛 Dropout	96.18 ± 0.02	0.1270 ± 0.0030	0.0162 ± 0.0007	96.39 ± 0.09	0.1094 ± 0.0021	**0.0094 ± 0.0014**
F-MC-Dropout	**96.39 ± 0.18**	**0.1137 ± 0.0024**	**0.0118 ± 0.0006**	**97.10 ± 0.12**	**0.0966 ± 0.0047**	0.0095 ± 0.0008
深度集成	96.39 ± 0.09	0.1277 ± 0.0030	0.0108 ± 0.0015	96.96 ± 0.10	0.1031 ± 0.0076	0.0087 ± 0.0018
F-深度集成	**96.70 ± 0.04**	**0.1031 ± 0.0016**	**0.0057 ± 0.0031**	**97.11 ± 0.10**	**0.0851 ± 0.0011**	**0.0059 ± 0.0012**
马尔可夫链蒙特卡洛						
SGLD	94.79 ± 0.10	0.2089 ± 0.0021	0.0711 ± 0.0061	95.87 ± 0.08	0.1573 ± 0.0190	0.0463 ± 0.0050
F-SGLD	**95.04 ± 0.06**	**0.1912 ± 0.0080**	**0.0601 ± 0.0002**	**96.43 ± 0.05**	**0.1336 ± 0.004**	**0.0385 ± 0.0003**
样本						
SWAG-Diag	96.03 ± 0.10	0.1251 ± 0.0029	0.0082 ± 0.0008	96.41 ± 0.05	0.1077 ± 0.0009	0.0047 ± 0.0013
F-SWAG-Diag	**96.23 ± 0.01**	**0.1108 ± 0.0013**	**0.0043 ± 0.0005**	**97.05 ± 0.08**	**0.0888 ± 0.0052**	**0.0043 ± 0.0004**
SWAG	96.03 ± 0.02	0.1232 ± 0.0022	**0.0053 ± 0.0004**	96.32 ± 0.08	0.1122 ± 0.0009	0.0088 ± 0.0006
F-SWAG	**96.25 ± 0.03**	**0.11062 ± 0.0014**	0.0056 ± 0.0002	**97.09 ± 0.14**	**0.0883 ± 0.0004**	**0.0036 ± 0.0008**

最后，式（6）中的目标函数表明，我们旨在找到一个近似后验分布，使得从中采样的任何模型都能感知尖锐度，同时也倾向于更简单的近似后验分布。这种偏好可以根据我们为这些分布所采用的设定来估计。在贝叶斯设定和变分推断的框架下，我们提出的具备尖锐度感知能力的后验分布可以整合到基于 MCMC 和变分推断的贝叶斯神经网络中。补充材料详细介绍了如何推导变分方法，以及如何将尖锐度感知能力融入实验中使用的贝叶斯神经网络，包括具有近似高斯分布的贝叶斯神经网络 [33],采用随机梯度朗之万动力学（SGLD）的贝叶斯神经网络 [59], 蒙特卡洛 Dropout [18], 贝叶斯深度集成 [35],以及 SWAG [39]。

4 实验

在本节中，我们进行了多项实验，以展示基于尖锐度感知的方法在贝叶斯神经网络上的有效性，包括具有近似高斯分布的贝叶斯神经网络 [33]（即，对模型使用重参数化技巧的 SGVB，以及对表示使用重参数化技巧的 SGVB-LRT）、采用随机梯度朗之万动力学（SGLD）的贝叶斯神经网络 [59], MC-Dropout [18], 贝叶斯深度集成 [35], 以及 SWAG [39]。实验在三个基准数据集上进行：CIFAR-10、CIFAR-100 和 ImageNet ILSVRC-2012，并报告准确率、负对数似然（NLL）和期望校准误差（ECE），以评估我们的方法相较于基线方法的校准能力和不确定性。数据集和实现的详细信息见补充材料[2]。

[2]实现代码见提供于https://github.com/anh-ntv/flat_bnn.git

表3：CIFAR 数据集上高斯后验近似的分类得分。每项实验使用不同的随机种子重复三次，并报告均值和标准差。

方法	Resnet10 准确率 ↑	Resnet10 负对数似然 ↓	Resnet10 期望校准误差 ↓	Resnet18 准确率 ↑	Resnet18 负对数似然 ↓	Resnet18 期望校准误差 ↓
CIFAR-100数据集上的实验						
SGVB	61.75 ± 0.75	1.534 ± 0.03	0.0676 ± 0.01	68.95 ± 1.20	1.140 ± 0.21	0.063 ± 0.04
F-SGVB-LRT	62.25 ± 0.57	1.4001 ± 0.04	0.0642 ± 0.01	70.00 ± 1.42	1.127 ± 0.25	**0.022 ± 0.05**
+ 几何	**62.54 ± 0.67**	**1.3704 ± 0.01**	**0.0301 ± 0.03**	**70.12 ± 1.02**	**1.121 ± 0.23**	0.036 ± 0.06
SGVB	54.40 ± 0.98	1.968 ± 0.05	0.214 ± 0.00	60.91 ± 2.31	1.746 ± 0.15	0.246 ± 0.03
F-SGVB	54.53 ± 0.33	1.967 ± 0.00	0.212 ± 0.00	61.54 ± 2.23	1.695 ± 0.15	0.242 ± 0.03
+ 几何	**55.53 ± 0.65**	**1.906 ± 0.02**	**0.207 ± 0.00**	**62.58 ± 0.53**	**1.612 ± 0.03**	**0.224 ± 0.00**
在 CIFAR-10 数据集上的实验						
SGVB	84.98 ± 1.87	0.422 ± 0.10	0.043 ± 0.04	89.10 ± 1.32	0.344 ± 0.02	0.033 ± 0.02
F-SGVB-LRT	86.32 ± 1.34	0.409 ± 0.03	**0.017 ± 0.06**	90.00 ± 1.10	0.291 ± 0.02	0.019 ± 0.01
+ 几何	**86.44 ± 1.12**	**0.403 ± 0.06**	0.025 ± 0.03	**90.31 ± 1.11**	**0.262 ± 0.01**	**0.014 ± 0.02**
SGVB	80.52 ± 2.10	0.781 ± 0.23	0.237 ± 0.06	86.74 ± 1.25	0.541 ± 0.01	0.181 ± 0.02
F-SGVB	80.60 ± 1.88	0.776 ± 0.13	0.223 ± 0.05	**87.01 ± 0.91**	0.534 ± 0.01	0.183 ± 0.01
+ 几何	**82.05 ± 0.47**	**0.704 ± 0.01**	**0.206 ± 0.00**	86.80 ± 1.30	**0.531 ± 0.01**	**0.175 ± 0.01**

表4：ImageNet 数据集上的分类得分

模型	Densenet-161 准确率 ↑	Densenet-161 负对数似然 ↓	Densenet-161 期望校准误差 ↓	ResNet-152 准确率 ↑	ResNet-152 负对数似然 ↓	ResNet-152 期望校准误差 ↓
SWAG-Diag	78.59	0.8559	0.0459	78.96	0.8584	0.0566
F-SWAG-Diag	**78.71**	**0.8267**	**0.0194**	**79.20**	**0.8065**	**0.0199**
SWAG	78.59	0.8303	0.0204	79.08	0.8205	0.0279
F-SWAG	**78.70**	**0.8262**	**0.0185**	**79.17**	**0.8078**	**0.0208**
SGLD	78.50	0.8317	**0.0157**	79.00	0.8165	0.0220
F-SGLD	**78.64**	**0.8236**	0.0166	**79.16**	**0.8050**	**0.0167**

4.1 实验结果

4.1.1 预测性能

我们的实验结果见表 1、2、3 （针对 CIFAR-100 和 CIFAR-10 数据集）以及表 4 （针对 ImageNet 数据集），结果表明所有实验均取得了显著改进。值得注意的是，准确率、负对数似然和期望校准误差之间存在权衡。尽管如此，与整体改进相比，我们的方法在这些指标之间实现了良好的平衡。

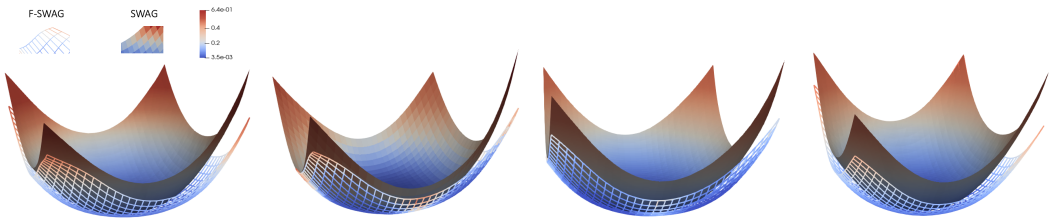


图 1：比较使用 SWAG 和 F-SWAG 方法在 CIFAR-100数据集上训练的 PreResNet-164 的损失景观。为便于可视化，我们分别从 SWAG 和 F-SWAG 中抽取两个模型，然后绘制损失景观。可以观察到，我们的 F-SWAG 的损失景观更加平坦，这支持了我们关于其采样模型更加平坦的论点。



图2：PreResNet164 在 CIFAR-100 上的可靠性图。将置信度划分为 20 个区间，并绘制每个区间内置信度与准确率之间的差值。当差值为零时，即为黑色虚线所示的最佳情况。F-SWAG 的曲线更接近零线，这表明我们的 F-SWAG 能更好地校准不确定性。

表5：使用 CIFAR-10 训练的 PreResNet-164 模型在 CIFAR-10-C 上的分类得分。每种损坏类型的完整结果见补充材料。

	期望校准误差 ↓				准确率 ↑			
损坏类型	SWAG-D	F-SWAG-D	SWAG	F-SWAG	SWAG-D	F-SWAG-D	SWAG	F-SWAG
噪声	0.0729	0.0701	0.0958	0.0078	74.26	75.59	74.02	75.08
Blur	0.0121	0.0090	0.0202	0.0273	91.13	90.55	91.03	90.93
天气	0.018	0.0142	0.0272	0.0240	89.47	89.18	89.42	89.11
数字及其他	0.0277	0.0229	0.0384	0.0209	87.03	86.94	86.93	87.19
平均值	0.0328	**0.0290**	0.0454	**0.0200**	85.47	**85.56**	85.35	**85.58**

4.2 尖锐度感知后验分布的有效性

不确定性估计的校准：我们评估了各设置的期望校准误差，并在表1、2和4中将其与基线方法进行了比较。该分数衡量模型准确率与置信度之间的最大差异。为进一步阐明这一点，我们在图2中展示了 PreResNet-164 在 CIFAR-100 上的可靠性图，以了解模型根据置信度阈值进行预测的效果。实验详情见补充材料。

分布外预测：上述实验展示了锐度感知贝叶斯神经网络（贝叶斯神经网络）的有效性，尤其是与非平坦方法相比时。本节将评估扩展至分布外场景。具体而言，我们使用在 CIFAR-10 数据集上训练的贝叶斯神经网络模型，评估其在 CIFAR-10-C 数据集上的表现。CIFAR-10-C 是对 CIFAR-10 的扩展，旨在评估机器学习模型对输入数据中常见损坏和扰动的鲁棒性。这些损坏包括各种形式的噪声、模糊、天气条件和数字失真。我们从平坦后验分布中采样 30 个模型并进行集成，再与非平坦模型进行比较。表5展示了各损坏组的平均结果以及整个数据集上的平均结果，各种损坏形式的详细结果见补充材料。值得注意的是，在平均期望校准误差和准确率指标方面，平坦贝叶斯神经网络模型始终优于非平坦模型。这一发现进一步证明了尖锐度感知后验的泛化能力。

4.3 消融研究

在图 1 中，我们绘制了从我们提出的尖锐度感知后验中采样的模型与从非尖锐度感知后验中采样的模型的损失景观。具体而言，在相同的超参数设置下，我们分别从两种方法 F-SWAG 和 SWAG 的后验分布中随机选取四个模型进行比较。正如观察到的那样，我们的方法不仅提升了集成推理的泛化能力，这一点可由第 4.1 节中的分类结果和第 4.2 节中的尖锐度结果加以证明，而且单个采样模型本身也更加平坦。

我们测量并可视化模型的尖锐度。为此，我们从近似后验中采样五个模型，然后计算这些模型尖锐度的平均值。对于一个模型θ，其尖锐度评估为 $\max_{||\epsilon||_2 \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta + \epsilon) - \mathcal{L}_{\mathcal{S}}(\theta)$，以衡量损失值的变化

在 θ附近。我们计算了在CIFAR-100数据集上采用SWAG和F-SWAG训练的PreResNet-164网络的尖锐度分数，并在补充材料中将其可视化。如其中所示，与相应的基线方法相比，尖锐度感知版本产生了更小的尖锐度分数，这表明我们的模型进入了更平坦的区域。

5 结论

本文介绍了贝叶斯设定下的相关理论，并讨论了贝叶斯神经网络（BNN）中针对尖锐度感知后验分布的变分推断。尖锐度感知后验分布能够使从中采样得到的模型以及最优近似后验估计具有更高的平坦度，从而降低模型对噪声的敏感性，并提升其泛化能力。我们将尖锐度感知后验分布应用于最先进的贝叶斯神经网络，并开展了大量实验。主要结果表明，与基线方法相比，从所提出的后验分布中采样得到的模型在集成准确率、期望校准误差（ECE）和负对数似然（NLL）方面表现更佳。这说明，趋平坦对应方法更善于捕捉神经网络中真实的权重分布，并提供准确的概率预测。此外，我们还进行了消融研究，以展示平坦后验分布在不确定性估计、损失景观和分布外预测等方面的效果。总体而言，尖锐度感知后验分布为提升贝叶斯神经网络的泛化性能提供了一种有前景的方法。

Ac致谢。本研究部分得到 ARC DP23 项目资助（项目编号 DP230101176），并获 b y the 空军科学研究办公室，资助编号 FA2386-23-1-4044。

参考文献

[1] Momin Abbas、Quan Xiao、Lisha Chen、Pin-Yu Chen 和 Tianyi Chen。Sharp-maml：尖锐度感知的模型无关元学习。arXiv预印本 arXiv:2206.03996，2022年。3[2] Pierre Alquier、James Ridgway 和 Nicolas Chopin。关于吉布斯后验的变分近似的性质。机器学习研究期刊，17(1):8374–8414，2016年。4[3] Dara Bahri、Hossein Mobahi 和 Yi Tay。尖锐度感知最小化提升语言模型泛化能力。载于第60届计算语言学协会年会论文集（第1卷：长篇论文），第7360–7371页，都柏林，爱尔兰，2022年5月。计算语言学协会。3[4] Jose Blanchet 和 Karthyek Murthy。通过最优传输量化分布模型风险。运筹学数学，44(2):565–600，2019年。3[5] Charles Blundell、Julien Cornebise、Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。载于国际机器学习会议论文集，第1613–1622页。PMLR，2015年。1, 3[6] Stephen Boyd 和 Lieven Vandenberghe。凸优化。剑桥大学出版社，3月。6[7] Olivier Catoni。PAC-贝叶斯监督分类：统计学习热力学。arXiv预印本 arXiv:0712.0248，2007年。4[8] Junbum Cha、Sanghyuk Chun、Kyungjae Lee、Han-Cheol Cho、Seunghyun Park、Yunsung Lee 和 Sungrae Park。SWAD：通过寻找平坦极小值实现领域泛化。神经信息处理系统进展，34:22405–22418，2021年。3[9] Pratik Chaudhari、Anna Choroma´nska、Stefano Soatto、Yann LeCun、Carlo Baldassi、Christian Borgs、Jennifer T. Chayes、Levent Sagun 和 Riccardo Zecchina。Entropy-SGD：将梯度下降引导至宽谷。统计力学杂志：理论与实验，2019，2017年。3[10] Tianqi Chen、Emily Fox 和 Carlos Guestrin。随机梯度哈密顿蒙特卡洛。载于国际机器学习会议论文集，第1683–1691页。PMLR，2014年。1, 2[11] Xiangning Chen、Cho-Jui Hsieh 和 Boqing Gong。视觉变换器如何在未经预训练或强数据增强的情况下优于残差网络。arXiv预印本 arXiv:2106.01548，2021年。3[12] Laurent Dinh、Razvan Pascanu、Samy Bengio 和 Yoshua Bengio。尖锐极小值也能在深度网络中实现泛化。载于国际机器学习会议论文集，第1019–1028页。PMLR，2017年。3[13] Jiawei Du、Daquan Zhou、Jiashi Feng、Vincent YF Tan 和 Joey Tianyi Zhou。免费的尖锐度感知训练。arXiv预印本 arXiv:2205.14083，2022年。3[14] Michael Dusenberry、Ghassen Jerfel、Yeming Wen、Yian Ma、Jasper Snoek、Katherine Heller、Balaji Lakshminarayanan 和 Dustin Tran。利用秩一因子实现高效且可扩展的贝叶斯神经网络。载于国际机器学习会议论文集，第2782–2792页。PMLR，2020年。3[15] Gintare Karolina Dziugaite 和 Daniel M. Roy。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。载于 UAI。AUAI出版社，2017年。2, 3[16] Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。尖锐度感知最小化：高效提升泛化能力。载于国际学习表征会议，2021年。2, 3, 23[17] Stanislav Fort 和 Surya Ganguli。神经网络损失景观局部几何的涌现特性。arXiv预印本 arXiv:1910.05929，2019年。3

[18] Yarin Gal 和 Zoubin Ghahramani。将 Dropout 作为贝叶斯近似：在深度学习中表示模型不确定性。载于 Maria Florina Balcan 和 Kilian Q. Weinberger 编，《第33届国际机器学习会议论文集》，机器学习研究论文集第48卷，第1050–1059页，美国纽约州纽约市，2016年6月20日至22日。PMLR。2, 3, 7, 20[19] Yarin Gal、Jiri Hron 和 Alex Kendall。Concrete Dropout。《神经信息处理系统进展》，30，2017年。3[20] Soumya Ghosh、Jiayu Yao 和 Finale Doshi-Velez。基于马蹄先验的贝叶斯神经网络结构化变分学习。载于《国际机器学习会议》，第1744–1753页。PMLR，2018年。2, 3[21] Alex Graves。神经网络的实用变分推断。《神经信息处理系统进展》，24，2011年。1, 2[22] Arjun K Gupta 和 Daya K Nagar。《矩阵变元分布》。Chapman and Hall/CRC，2018年。3[23] Sepp Hochreiter 和 J¨urgen Schmidhuber。通过发现平坦极小值简化神经网络。载于 NIPS，第529–536页。MIT出版社，1994年。3[24] Pavel Izmailov、Dmitrii Podoprikhin、Timur Garipov、Dmitry Vetrov 和 Andrew Gordon Wilson。权重平均带来更宽的最优解和更好的泛化。arXiv 预印本 arXiv:1803.05407，2018年。3[25] Pavel Izmailov、Dmitrii Podoprikhin、Timur Garipov、Dmitry P. Vetrov 和 Andrew Gordon Wilson。权重平均带来更宽的最优解和更好的泛化。载于 UAI，第876–885页。AUAI出版社，2018年。3[26] Stanislaw Jastrzebski、Zachary Kenton、Devansh Arpit、Nicolas Ballas、Asja Fischer、Yoshua Bengio 和 Amos J. Storkey。影响随机梯度下降中极小值的三个因素。ArXiv，abs/1711.04623，2017年。3[27] Yiding Jiang、Behnam Neyshabur、Hossein Mobahi、Dilip Krishnan 和 Samy Bengio。绝妙的泛化度量及其寻找之处。载于国际学习表征会议。OpenReview.net，2020年。2, 3[28] Jean Kaddour、Linqing Liu、Ricardo Silva 和 Matt Kusner。平坦极小值优化器何时有效？载于 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，《神经信息处理系统进展》，2022年。3[29] Nitish Shirish Keskar、Dheevatsa Mudigere、Jorge Nocedal、Mikhail Smelyanskiy 和 Ping Tak Peter Tang。深度学习的大批量训练：泛化间隙与尖锐极小值。载于国际学习表征会议。OpenReview.net，2017年。3[30] Mohammad Khan、Didrik Nielsen、Voot Tangkaratt、Wu Lin、Yarin Gal 和 Akash Srivastava。通过 Adam 中的权重扰动实现快速且可扩展的贝叶斯深度学习。载于《国际机器学习会议》，第2611–2620页。PMLR，2018年。2, 3[31] Minyoung Kim、Da Li、Shell X Hu 和 Timothy Hospedales。Fisher SAM：信息几何与尖锐度感知最小化。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，《第39届国际机器学习会议论文集》，机器学习研究论文集第162卷，第11148–11161页。PMLR，2022年7月17日至23日。3, 6[32] Diederik P Kingma 和 Max Welling。自动编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114，2013年。3[33] Durk P Kingma、Tim Salimans 和 Max Welling。变分 Dropout 与局部重参数化技巧。《神经信息处理系统进展》，28，2015年。1, 2, 3, 7[34] Jungmin Kwon、Jeongseop Kim、Hyunseo Park 和 In Kwon Choi。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。3, 6

[35] Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。使用深度集成进行简单且可扩展的预测不确定性估计。《神经信息处理系统进展》，30，2017年。2, 7, 20[36] Qiang Liu 和 Dilin Wang。斯坦变分梯度下降：一种通用贝叶斯推断算法。《神经信息处理系统进展》，29，2016年。1, 2[37] Yong Liu、Siqi Mai、Xiangning Chen、Cho-Jui Hsieh 和 Yang You。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。3[38] Christos Louizos 和 Max Welling。用于变分贝叶斯神经网络的乘性归一化流。载于《国际机器学习会议》，第2218–2227页。PMLR，2017年。3[39] Wesley J. Maddox、Timur Garipov、Pavel Izmailov、Dmitry Vetrov 和 Andrew Gordon Wilson。《深度学习中贝叶斯不确定性的简单基线》。Curran Associates Inc.，美国纽约州雷德胡克，2019年。2, 3, 7, 21, 22, 23[40] Thomas M¨ollenhoff 和 Mohammad Emtiyaz Khan。SAM 作为贝叶斯方法的最优松弛。arXiv 预印本 arXiv:2210.01620，2022年。3, 21, 23[41] Radford M. Neal。《神经网络的贝叶斯学习》。施普林格出版社，柏林，海德堡，1996年。1, 2[42] Behnam Neyshabur、Srinadh Bhojanapalli、David McAllester 和 Nati Srebro。探索深度学习中的泛化。《神经信息处理系统进展》，30，2017年。3[43] Son Nguyen、Duong Nguyen、Khai Nguyen、Khoat Than、Hung Bui 和 Nhat Ho。贝叶斯神经网络的结构化丢弃变分推断。《神经信息处理系统进展》，34:15188–15202，2021年。3[44] Van-Anh Nguyen、Trung Le、Anh Bui、Thanh-Toan Do 和 Dinh Phung。最优传输模型的分布鲁棒性。载于《神经信息处理系统进展》，2023年。3[45] Victor M-H Ong、David J Nott 和 Michael S Smith。具有因子协方差结构的高斯变分近似。《计算与图形统计学杂志》，27(3):465–478，2018年。2, 3[46] Gabriel Pereyra、George Tucker、Jan Chorowski、Lukasz Kaiser 和 Geoffrey E. Hinton。通过惩罚高置信度输出分布来正则化神经网络。载于国际学习表征会议（研讨会）。OpenReview.net，2017年。3[47] Henning Petzka、Michael Kamp、Linara Adilova、Cristian Sminchisescu 和 Mario Boley。相对平坦性与泛化。载于神经信息处理系统大会，第18420–18432页，2021年。2, 3[48] Cuong Pham、C. Cuong Nguyen、Trung Le、Phung Dinh、Gustavo Carneiro 和 Thanh-Toan Do。互学习中贝叶斯神经网络的模型与特征多样性。载于《神经信息处理系统进展》，2023年。2[49] Hoang Phan、Trung Le、Trung Phung、Anh Tuan Bui、Nhat Ho 和 Dinh Phung。通过分布鲁棒性实现全局-局部正则化。载于 Francisco Ruiz、Jennifer Dy 和 Jan-Willem van de Meent 编，《第26届国际人工智能与统计学会议论文集》，机器学习研究论文集第206卷，第7644–7664页。PMLR，2023年4月25日至27日。3[50] Hoang Phan、Lam Tran、Ngoc N Tran、Nhat Ho、Dinh Phung 和 Trung Le。通过寻找基于任务的平坦区域改进多任务学习。arXiv 预印本 arXiv:2211.13723，2022年。3[51] Zhe Qu、Xingyu Li、Rui Duan、Yao Liu、Bo Tang 和 Zhuo Lu。通过尖锐度感知最小化实现广义联邦学习。arXiv 预印本 arXiv:2206.02618，2022年。3

[52] Hippolyt Ritter、Aleksandar Botev 和 David Barber。《神经网络的可扩展拉普拉斯近似》。载于第六届国际学习表征会议（ICLR 2018）会议论文集，第 6 卷。国际表征学习会议，2018年。2, 3[53] Simone Rossi、Sebastien Marmin 和 Maurizio Filippone。《用于贝叶斯深度学习的 Walsh-Hadamard 变分推断》。神经信息处理系统进展，33：9674–9686，2020年。2, 3[54] Shai Shalev-Shwartz 和 Shai Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。16[55] Jakub Swiatkowski、Kevin Roth、Bastiaan Veeling、Linh Tran、Joshua Dillon、Jasper Snoek、Stephan Mandt、Tim Salimans、Rodolphe Jenatton 和 Sebastian Nowozin。《k-tied 正态分布：贝叶斯神经网络中高斯均值场后验的紧凑参数化》。载于国际机器学习会议，第 9289–9299 页。PMLR，2020年。2, 3[56] Marcin Tomczak、Siddharth Swaroop 和 Richard Turner。《用于神经网络的高效低秩高斯变分推断》。神经信息处理系统进展，33：4610–4622，2020年。2, 3[57] Tuan Truong、Hoang-Phi Nguyen、Tung Pham、Minh-Tuan Tran、Mehrtash Harandi、Dinh Phung 和 Trung Le。《RSAM：通过黎曼锐度感知最小化在流形上学习》，2023。3[58] Colin Wei、Sham Kakade 和 Tengyu Ma。《Dropout 的隐式与显式正则化效应》。载于国际机器学习会议，第 10181–10192 页。PMLR，2020年。3[59] Max Welling 和 Yee W Teh。《通过随机梯度朗之万动力学进行贝叶斯学习》。载于第 28 届国际机器学习会议（ICML-11）论文集，第 681–688 页，2011年。1, 2, 7, 21[60] Guodong Zhang、Shengyang Sun、David Duvenaud 和 Roger Grosse。《作为变分推断的噪声自然梯度》。载于国际机器学习会议，第 5852–5861 页。PMLR，2018年。3[61] Linfeng Zhang、Jiebo Song、Anni Gao、Jingwei Chen、Chenglong Bao 和 Kaisheng Ma。《做自己的老师：通过自蒸馏提升卷积神经网络的性能》。载于 IEEE/CVF 国际计算机视觉会议论文集，第 3713–3722 页，2019年。3[62] Ying Zhang、Tao Xiang、Timothy M. Hospedales 和 Huchuan Lu。《深度互学习》。2018 IEEE/CVF 计算机视觉与模式识别会议，第 4320–4328 页，2018年。2, 3[63] Juntang Zhuang、Boqing Gong、Liangzhe Yuan、Yin Cui、Hartwig Adam、Nicha Dvornek、Sekhar Tatikonda、James Duncan 和 Ting Liu。《代理间隙最小化改进尖锐度感知训练》。arXiv 预印本 arXiv:2203.08065，2022年。3

A 理论推导

A.1 所有证明

定理 A.1。（正文论文中的定理 *3.1*）考虑以下优化问题 m

$$\min_{\mathbb{Q}<<\mathbb{P}} \left\{\lambda \mathcal{L}_S\left(\mathbb{Q}\right) + KL\left(\mathbb{Q},\mathbb{P}\right)\right\}, \tag{8}$$

其中，我们在满足相对于 $\mathbb{P}$ 绝对连续的 $\mathbb{Q}$ 上进行搜索，且 $KL\left(\cdot,\ \cdot\right)$ 是 *Kullback-Leibler* 散度。该优化问题有一个闭式最优解 $\mathbb{Q}^*$，其密度为

$$q^*\left(\theta\right) \propto \exp\left\{-\lambda \mathcal{L}_{\mathcal{S}}\left(\theta\right)\right\} p(\theta),$$

这恰好是标准后验分布 $\mathbb{Q}_{\mathcal{S}}$，其密度为 $q\left(\theta \mid \mathcal{S}\right)$。

证明。我们有

$$\lambda \mathcal{L}_S\left(\mathbb{Q}\right) + KL\left(\mathbb{Q},\mathbb{P}\right) = \lambda \int \mathcal{L}_S\left(\theta\right) q\left(\theta\right) d\theta + \int q(\theta) \log \frac{q\left(\theta\right)}{p\left(\theta\right)} d\theta.$$

拉格朗日函数如下

$$L\left(q,\alpha\right) = \lambda \int \mathcal{L}_S\left(\theta\right) q\left(\theta\right) d\theta + \int q(\theta) \log \frac{q\left(\theta\right)}{p\left(\theta\right)} d\theta + \alpha \left(\int q(\theta) d\theta - 1\right).$$

对 $q(\theta)$ 求导并令其等于 0，得到

$$\lambda \mathcal{L}_S\left(\theta\right) + \log q\left(\theta\right) + 1 - \log p\left(\theta\right) + \alpha = 0.$$

$$q\left(\theta\right) = \exp\left\{-\lambda \mathcal{L}_S\left(\theta\right)\right\} p\left(\theta\right) \exp\left\{-\alpha - 1\right\}.$$

$$q\left(\theta\right) \propto \exp\left\{-\lambda \mathcal{L}_S\left(\theta\right)\right\} p\left(\theta\right).$$

□

引理 A.2. 假设数据空间 $\mathcal{X}$、标签空间 $\mathcal{Y}$ 和模型空间 Θ 都是紧集。存在连续模 $\omega:\ \mathbb{R}^+ \to \mathbb{R}^+$，满足 $\lim_{t\to 0^+} \omega(t) = 0$，使得$\left|\ell\left(f_\theta\left(x\right)\ \ y\right) - \ell\left(f_{\theta'}\left(x\right)\ \ y\right)\right| \le \omega\left(\left\|\theta - \theta'\right\|\right)\ \ \forall x \in \mathcal{X}, y \in \mathcal{Y}$。

证明。损失函数 $\ell(f_\theta(x)\ \ y)$ 在紧集 $\mathcal{X} \times \mathcal{Y} \times \Theta$ 上连续，因此在该集上等度连续。对于每个 $\epsilon > 0$，存在 $\delta_x, \delta_y, \delta_\theta > 0$，使得

$$\forall \left\|x' - x\right\| \le \delta_x, \left\|y' - y\right\| \le \delta_y, \left\|\theta' - \theta\right\| \le \delta_\theta,$$

我们有 $\left|\ell\left(f_{\theta'}(x')\ \ y'\right) - \ell\left(f_{\theta'}(x)\ \ y\right)\right| \le \epsilon$。
因此，对于所有 $\left\|\theta' - \theta\right\| \le \delta_\theta$，我们有

$$\left|\ell\left(f_\theta\left(x\right), y\right) - \ell\left(f_{\theta'}\left(x\right), y\right)\right| \le \epsilon, \forall x, y.$$

这意味着族 $\left\{\ell\left(f_\theta(x)\ \ y\right):\ x \in \mathcal{X}, y \in \mathcal{Y}\right\}$ 关于 $\theta \in \Theta$是等度连续的。这意味着存在公共连续模 $\omega:\ \mathbb{R}^+ \to \mathbb{R}^+$，且 $\lim_{t\to 0^+} \omega(t) = 0$。

□

定义 A.3。给定 $\epsilon > 0$，我们称集合 Θ' 是 ϵ-覆盖的，如果对于所有 $\theta \in \Theta$，存在$\theta' \in \Theta'$ 使得 $\left\|\theta' - \theta\right\| \le \epsilon.$。我们将 $\mathcal{N}(\Theta, \epsilon)$ 定义为覆盖 Θ 的最小集合 Θ' 的基数。

引理 A.4. 令 $R = \text{最大值}_{\theta\in\Theta} \left\|\theta\right\|^2 < \infty$，且 k 是 Θ 的维数。我们可以将覆盖数上界限定为

$$\mathcal{N}\left(\Theta, \epsilon\right) \le \left(\frac{2R\sqrt{k}}{\epsilon}\right)^k.$$

证明。 证明见 [54] 第27章。 □

通过选择 $\epsilon = \frac{1}{n^{\frac{1}{2k}}}$ 我们得到

$$\mathcal{N}\left(\Theta, n^{-\frac{1}{2k}}\right) \leq \left(2R\sqrt{k}\right)^k \sqrt{n}.$$

然而，对于一般的数据-标签分布 $\mathcal{D}$，求解优化问题（OP）通常是难以处理的。为了使其变得可处理，我们求出其与尖锐度相关的上界，如下述定理所示。

定理 A.5。 （主论文中的定理 *3.2*）假设 Θ 是一个紧集。给定任意 $\delta \in [0;\ 1]$,，对于 $\mathcal{S} \sim \mathcal{D}^n$的选择，以至少 $1-\delta$ 的概率，对任意分布 $\mathbb{Q}$，有

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) \leq \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')\right] + \mathcal{L}_{\mathcal{S}}(\mathbb{Q}) + \frac{1}{\sqrt{n}} + 2\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right) + \sqrt{\frac{k\left(1+\log\left(1+\frac{2R^2}{\rho^2}\left(1+2\log\left(2R\sqrt{k}\right)+\frac{2}{k}\log n\right)\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}},$$

其中我们假设 $\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) = \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\mathcal{L}_{\mathcal{D}}(\theta)\right] \leq \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}(\theta+\epsilon)\right]\right]$ 满足 $\sigma = \frac{\rho}{k^{1/2}\left(1+\sqrt{\log\left(\frac{N^2 n}{k}\right)}\right)}$，并且 $N = \mathcal{N}\left(\Theta, n^{-\frac{1}{2k}}\right)$、$k$ 是模型的参数数量，$n = |S|$、$R = \max_{\theta\in\Theta}\|\theta\|$，而 $\omega:\ \mathbb{R}^+ \to \mathbb{R}^+$ 是一个函数，满足 $\lim_{t\to 0^+}\omega(t) = 0$。

证明。 给定 $\epsilon = \frac{1}{n^{\frac{1}{2k}}}$，我们将 $\Theta' = \left\{\theta'_1, \ldots, \theta'_N\right\}$（其中 $N = \mathcal{N}\left(\Theta, n^{-\frac{1}{2k}}\right) \leq \left(2R\sqrt{k}\right)^k \sqrt{n}$）记为 Θ 的 ϵ-覆盖集。我们首先考察一个离散分布

$$\mathbb{Q} = \sum_{i=1}^{m} \pi_i \delta_{\theta_i}.$$

不失一般性，我们可以假设 $\left\|\theta'_i - \theta_i\right\| \leq \epsilon, \forall i = 1, \ldots, m$。注意，如果 $m > N$，$\theta'_1, \ldots . \theta'_m$ 可以重复。根据引理 A.2，令 $\omega(\cdot)$ 为 $\ell(f_\theta(x)\ \ y)$ 的连续模，使得 $|\ell(f_\theta(x)\ \ y) - \ell(f_\theta(x)\ \ y)| \leq \omega(\|\theta-\theta'\|)\ \ \forall x, y$，且 $\lim_{t\to 0}\omega(t) = 0$。这意味着

$$\left|\ell\left(f_{\theta_i}(x), y\right) - \ell\left(f_{\theta'_i}(x), y\right)\right| \leq \omega(\epsilon) = \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right), \forall x, y, i = 1, \ldots, m.$$

我们考虑分布 $\bar{\mathbb{Q}} = \sum_{i=1}^m \pi_i \mathcal{N}\left(\theta'_i, \sigma\mathbb{I}\right)$。根据 McAllester PAC-Bayes 界，以关于 $\mathcal{S} \sim \mathcal{D}^n$ 的选择的概率 $1-\delta$，对于任意分布 $\mathbb{P}$，我们有

$$\mathcal{L}_{\mathcal{D}}\left(\bar{\mathbb{Q}}\right) \leq \mathcal{L}_S\left(\bar{\mathbb{Q}}\right) + \sqrt{\frac{KL\left(\bar{\mathbb{Q}}, \bar{\mathbb{P}}\right) + \log\frac{n}{\delta}}{2(n-1)}}.$$

令 $\theta^* = \operatorname{argmax}_{1\leq i\leq m}\left\|\theta'_i\right\|$。我们考虑分布 $\bar{\mathbb{P}} = \mathcal{N}(0, \sigma_{\mathbb{P}})$，其中 $\sigma_{\mathbb{P}}^2 = c\exp\left\{\frac{1-j}{k}\right\}$，$c = \sigma^2\left(1+\exp\left\{\frac{4n}{k}\right\}\right)$，且 $j = \left\lfloor 1 + k\log\frac{c}{\sigma^2 + \frac{\|\theta^*\|^2}{k}}\right\rfloor = \left\lfloor 1 + k\log\frac{\sigma^2(1+\exp\left\{\frac{4n}{k}\right\})}{\sigma^2 + \frac{\|\theta^*\|^2}{k}}\right\rfloor$。由此可得

$$\sigma^2 + \frac{\|\theta^*\|^2}{k} \leq \sigma_{\mathbb{P}} \leq \exp\left\{\frac{1}{k}\right\}\left(\sigma^2 + \frac{\|\theta^*\|^2}{k}\right).$$

我们有

$$KL\left(\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right),\bar{\mathbb{P}}\right)=\frac{1}{2}\left[\frac{k\sigma^2+\left\|\theta_i^{'}\right\|^2}{\sigma_{\mathbb{P}}^2}-k+k\log\left(\frac{\sigma_{\mathbb{P}}^2}{\sigma^2}\right)\right].$$

$$KL\left(\mathcal{N}\left(\theta^*,\sigma\mathbb{I}\right),\bar{\mathbb{P}}\right)=\max_i KL\left(\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right),\bar{\mathbb{P}}\right).$$

$$KL\left(\bar{\mathbb{Q}},\bar{\mathbb{P}}\right)\leq\sum_{i=1}^m\pi_i KL\left(\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right),\bar{\mathbb{P}}\right)\leq KL\left(\mathcal{N}\left(\theta^*,\sigma\mathbb{I}\right),\bar{\mathbb{P}}\right).$$

现在我们对 $KL\left(\mathcal{N}(\theta^*,\ \sigma\mathbb{I})\ \ \bar{\mathbb{P}}\right)$进行界定

$$\begin{aligned}KL\left(\mathcal{N}\left(\theta^*,\sigma\mathbb{I}\right),\bar{\mathbb{P}}\right)=&\frac{1}{2}\left[\frac{k\sigma^2+\left\|\theta^*\right\|^2}{\sigma_{\mathbb{P}}^2}-k+k\log\left(\frac{\sigma_{\mathbb{P}}^2}{\sigma^2}\right)\right]\\ \leq&\frac{1}{2}\left[\frac{k\sigma^2+\left\|\theta^*\right\|^2}{\sigma^2+\frac{\|\theta^*\|^2}{k}}-k+k\log\left(\frac{\exp\left\{\frac{1}{k}\right\}\left(\sigma^2+\frac{\|\theta^*\|^2}{k}\right)}{\sigma^2}\right)\right]\\ \leq&\frac{k}{2}\left(1+\log\left(1+\frac{\left\|\theta^*\right\|^2}{k\sigma^2}\right)\right).\end{aligned}$$

因此，以概率 $1-\delta$，我们得到

$$\mathcal{L}_{\mathcal{D}}\left(\bar{\mathbb{Q}}\right)\leq\mathcal{L}_S\left(\bar{\mathbb{Q}}\right)+\sqrt{\frac{k\left(1+\log\left(1+\frac{\|\theta^*\|^2}{k\sigma^2}\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}}.$$

$$\begin{aligned}\mathbb{E}_{\theta\sim\sum_{i=1}^m\pi_i\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right)}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]\leq&\mathbb{E}_{\theta\sim\sum_{i=1}^m\pi_i\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right)}\left[\mathcal{L}_{\mathcal{S}}\left(\theta\right)\right]+\sqrt{\frac{k\left(1+\log\left(1+\frac{\|\theta^*\|^2}{k\sigma^2}\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}}.\\ \leq&\sum_{i=1}^m\pi_i\mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right)\right]+\sqrt{\frac{k\left(1+\log\left(1+\frac{\|\theta^*\|^2}{k\sigma^2}\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}}.\end{aligned}$$

注意，

$$\begin{aligned}\mathbb{E}_{\theta\sim\mathcal{N}(\theta_i,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]-\mathbb{E}_{\theta\sim\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right)}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]=&\int\left[\mathcal{L}_{\mathcal{D}}\left(\theta_i+\epsilon_i\right)-\mathcal{L}_{\mathcal{D}}\left(\theta_i^{'}+\epsilon_i\right)\right]\mathcal{N}\left(\epsilon_i\mid0,\sigma I\right)d\epsilon_i\\ \leq&\int\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right)\mathcal{N}\left(\epsilon_i\mid0,\sigma I\right)d\epsilon_i=\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right).\end{aligned}$$

$$\mathbb{E}_{\theta\sim\mathcal{N}(\theta_i,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]\leq\mathbb{E}_{\theta\sim\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right)}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]+\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right).$$

$$\sum_{i=1}^m\pi_i\mathbb{E}_{\theta\sim\mathcal{N}(\theta_i,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]\leq\sum_{i=1}^m\pi_i\mathbb{E}_{\theta\sim\mathcal{N}\left(\theta_i^{'},\sigma\mathbb{I}\right)}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]+\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right),$$

因此，我们有

$$\begin{aligned}\mathbb{E}_{\theta\sim\sum_{i=1}^m\pi_i\mathcal{N}(\theta_i,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}\left(\theta\right)\right]\leq&\sum_{i=1}^m\pi_i\mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right)\right]\\ &+\sqrt{\frac{k\left(1+\log\left(1+\frac{\|\theta^*\|^2}{k\sigma^2}\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}}+\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right).\end{aligned}$$

根据假设

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) = \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\mathcal{L}_{\mathcal{D}}(\theta)\right] \leq \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}(\theta+\epsilon)\right]\right] = \mathbb{E}_{\theta\sim\sum_{i=1}^{m}\pi_i\mathcal{N}(\theta_i,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{D}}(\theta)\right],$$

我们得到

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) \leq & \sum_{i=1}^{m}\pi_i\mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,\sigma\mathbb{I})}\left[\mathcal{L}_{\mathcal{S}}(\theta_i+\epsilon_i)\right] \\ & + \sqrt{\frac{k\left(1+\log\left(1+\frac{R^2}{k\sigma^2}\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}} + \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right).\end{aligned}$$

由于 $\epsilon_i \sim \mathcal{N}(\mathbf{0},\sigma\mathbb{I})$, $\|\epsilon_i\|^2$ 服从卡方分布。因此，对于任意$i\in[m]$，我们有

$$\mathbb{P}\left(\|\epsilon_i\|^2 - k\sigma^2 \geq 2\sigma^2\sqrt{kt}+2t\sigma^2\right) \leq \exp(-t), \forall t.$$

$$\mathbb{P}\left(\max_{i\in[m]}\|\epsilon_i\|^2 - k\sigma^2 \geq 2\sigma^2\sqrt{kt}+2t\sigma^2\right) \leq N\exp(-t), \forall t.,$$

因为 $\left|\{\theta_1',\ldots,\theta_m'\}\right|$的基数不超过 N。

$$\mathbb{P}\left(\max_{i\in[m]}\|\epsilon_i\|^2 - k\sigma^2 < 2\sigma^2\sqrt{kt}+2t\sigma^2\right) > 1-N\exp(-t), \forall t.$$

令 $t=\log\left(Nn^{1/2}\right)$，以至少$1-\frac{1}{\sqrt{n}}$的概率，对于所有 $i\in[m]$，我们有

$$\|\epsilon_i\|^2 < \sigma^2 k\left(1+\frac{\log\left(N^2n\right)}{k}+2\sqrt{\frac{\log\left(Nn^{1/2}\right)}{k}}\right) \leq \sigma^2 k\left(1+\sqrt{\frac{\log\left(N^2n\right)}{k}}\right)^2.$$

通过选择$\sigma = \frac{\rho}{k^{1/2}\left(1+\sqrt{\frac{\log\left(N^2n\right)}{k}}\right)}$，以至少$1-\frac{1}{\sqrt{n}}$的概率，对于所有 $i\in[m]$，我们有

$$\|\epsilon_i\| < \rho.$$

现在我们推导

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) &\leq \sum_{i=1}^{m} \pi_i \left(1-\frac{1}{\sqrt{n}}\right) \max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right) \\
&+ \frac{1}{\sqrt{n}} + \sqrt{\frac{k\left(1+\log\left(1+\frac{R^2}{k\sigma^2}\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}} + \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right) \\
&\leq \left(1-\frac{1}{\sqrt{n}}\right) \sum_{i=1}^{m} \pi_i \max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right) + \frac{1}{\sqrt{n}} \\
&+ \sqrt{\frac{k\left(1+\log\left(1+\frac{R^2}{\rho^2}\left(1+\sqrt{\frac{\log(N^2n)}{k}}\right)^2\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}} + \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right) \\
&\leq \sum_{i=1}^{m} \pi_i \max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right) + \frac{1}{\sqrt{n}} \\
&+ \sqrt{\frac{k\left(1+\log\left(1+\frac{2R^2}{\rho^2}\left(1+\frac{\log(N^2n)}{k}\right)\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}} + \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right) \\
&\leq \sum_{i=1}^{m} \pi_i \max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right) + \frac{1}{\sqrt{n}} \\
&+ \sqrt{\frac{k\left(1+\log\left(1+\frac{2R^2}{\rho^2}\left(1+2\log\left(2R\sqrt{k}\right)+\frac{2}{k}\log n\right)\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}} + \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right).
\end{aligned}$$

注意，对于所有 $i \in [m]$

$$\max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i^{'}+\epsilon_i\right) \leq \max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i+\epsilon_i\right) + \omega\left(\frac{1}{n^{\frac{1}{2k}}}\right),$$

因此，我们得到

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) &\leq \sum_{i=1}^{m} \pi_i \max_{\|\epsilon_i\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta_i+\epsilon_i\right) + \frac{1}{\sqrt{n}} \\
&+ \sqrt{\frac{k\left(1+\log\left(1+\frac{2R^2}{\rho^2}\left(1+2\log\left(2R\sqrt{k}\right)+\frac{2}{k}\log n\right)\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}} + 2\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right) \\
&\leq \mathcal{L}_{\mathcal{S}}(\mathbb{Q}) + \frac{1}{\sqrt{n}} + 2\omega\left(\frac{1}{n^{\frac{1}{2k}}}\right) \\
&+ \sqrt{\frac{k\left(1+\log\left(1+\frac{2R^2}{\rho^2}\left(1+2\log\left(2R\sqrt{k}\right)+\frac{2}{k}\log n\right)\right)\right)+2\log\frac{n}{\delta}}{4(n-1)}}.
\end{aligned}$$

对于任意分布 $\mathbb{Q}$，我们用其经验分布来近似 $\mathbb{Q}$

$$\mathbb{Q}_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{\theta_i},$$

其弱收敛于 $\mathbb{Q}$，当 $m \to \infty$ 时。利用针对 $\mathbb{Q}_m$ 得到的结果，并在 $m \to \infty$ 时取极限，我们得出结论。

□

定理 A.6。（正文中的定理 *3.3*）中，*OP* 的最优解是尖锐度感知后验分布 $\mathbb{Q}_S^{SA}$，其密度函数为 $q^{SA}(\theta|\mathcal{S})$:

$$\begin{aligned} q^{SA}(\theta|\mathcal{S}) &\propto \exp\left\{-\lambda \max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta'\right)\right\} p\left(\theta\right) \\ &= \exp\left\{-\lambda \mathcal{L}_{\mathcal{S}}\left(s\left(\theta\right)\right)\right\} p\left(\theta\right), \end{aligned}$$

其中，我们定义了 $s\left(\theta\right) = \underset{\theta':\|\theta'-\theta\|\leq\rho}{argmax} \mathcal{L}_{\mathcal{S}}\left(\theta'\right)$。

证明。我们有

$$\lambda \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\max_{\theta':\|\theta'-\theta\|\leq\rho} \mathcal{L}_{\mathcal{S}}\left(\theta'\right)\right] + KL\left(\mathbb{Q},\mathbb{P}\right) = \lambda \int \mathcal{L}_S\left(s\left(\theta\right)\right) q\left(\theta\right) d\theta + \int q(\theta) \log \frac{q\left(\theta\right)}{p\left(\theta\right)} d\theta.$$

拉格朗日函数如下

$$L\left(q,\alpha\right) = \lambda \int \mathcal{L}_S\left(s(\theta)\right) q\left(\theta\right) d\theta + \int q(\theta) \log \frac{q\left(\theta\right)}{p\left(\theta\right)} d\theta + \alpha\left(\int q(\theta) d\theta - 1\right).$$

对 $q(\theta)$ 求导并令其等于 0，可得

$$\begin{aligned} &\lambda \mathcal{L}_S\left(s(\theta)\right) + \log q\left(\theta\right) + 1 - \log p\left(\theta\right) + \alpha = 0. \\ &q\left(\theta\right) = \exp\left\{-\lambda \mathcal{L}_S\left(s(\theta)\right)\right\} p\left(\theta\right) \exp\left\{-\alpha - 1\right\}. \\ &q\left(\theta\right) \propto \exp\left\{-\lambda \mathcal{L}_S\left(s(\theta)\right)\right\} p\left(\theta\right). \end{aligned}$$

□

A.2 基线方法及其对应的平坦版本的技术细节

下面我们介绍如何将实验中使用的基线方法视为变分方法和马尔可夫链蒙特卡洛方法，并融入我们的尖锐度感知技术。

贝叶斯深度集成 [35]: 我们考虑近似后验 $q_\phi = \frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}$ 其中 δ是由若干基础模型构成的均匀分布上的狄拉克 δ 分布 $\theta_{1:K}$。考虑先验分布 $p(\theta) = \mathcal{N}(0,\ \mathbb{I})$，我们得到非平坦版本和平坦版本对应的以下优化问题。

非平坦版本:

$$\min_{\theta_{1:K}} \left\{ \mathbb{E}_{\theta_k\sim q_\phi}\left[\lambda \mathcal{L}_S\left(\theta_k\right)\right] + KL\left(\frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}, \mathcal{N}\left(0,\mathbb{I}\right)\right)\right\},$$

其中 $KL\left(\frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}, \mathcal{N}(0,\ \mathbb{I})\right) = -\frac{1}{K}\sum_{k=1}^{K} \log \mathcal{N}(\theta_k \mid 0,\ \mathbb{I}) +$ 为常数，因此得到 L2 正则化项。

平坦版本:

$$\min_{\theta_{1:K}} \left\{ \mathbb{E}_{\theta_k\sim q_\phi}\left[\lambda \max_{\theta':\|\theta'-\theta_k\|\leq\rho} \mathcal{L}_S\left(\theta'\right)\right] + KL\left(\frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}, \mathcal{N}\left(0,\mathbb{I}\right)\right)\right\}.$$

蒙特卡洛 Dropout [18]: 如 [18],所示，MC-dropout 可以视为一个贝叶斯神经网络，其近似后验分布为 $q_\phi = \delta_\phi$，其中 ϕ 是一个不含任何 Dropout 的全连接基础模型，先验分布为 $p(\theta) = \mathcal{N}(0,\ \mathbb{I})$。$KL(q_\phi,\ p(\theta))$ 可以近似得到，结果是加权 L2 正则化，其中权重与各层的保留概率成正比。主项 $\mathbb{E}_{\theta\sim q_\phi}\left[\lambda \mathcal{L}_S(\theta)\right]$ 可以解释为在最小化损失之前应用 Dropout。对于我们的平坦版本，主项为 $\mathbb{E}_{\theta\sim q_\phi}\left[\lambda \max_{\theta':\ \|\theta'-\theta\|\leq\rho} \mathcal{L}_S\left(\theta'\right)\right]$。

采用随机梯度朗之万动力学（SGLD）的贝叶斯神经网络 [59]： 对于 SGLD，我们直接从后验分布 $q(\theta \mid S)$ 中采样一个或多个粒子模型，用于非平坦版本；对于平坦版本，则从 SA后验分布 $q^{SA}(\theta \mid S)$中采样。对于非平坦版本，更新方式与小批量随机梯度下降类似，不同之处在于我们会向粒子模型添加少量高斯噪声。对于我们的平坦版本，我们首先针对给定的粒子模型 θ^a 计算扰动后的模型 θ，并使用在 θ^a 处计算梯度的小批量随机梯度下降更新，同时加入少量高斯噪声。

SWA [39]： 我们将 SWAG 视为一种马尔可夫链蒙特卡洛方法，通过 SWA 保留粒子模型的轨迹。此外，根据该轨迹确定协方差矩阵，以形成近似高斯后验。对于这种方法对应的平坦版本，我们采用 SWA 从 SA（尖锐度感知）后验中采样。具体而言，我们首先根据当前模型 θ 计算扰动后的模型 θ^a，然后使用在模型 θ^a 处计算的梯度进行小批量随机梯度下降更新。最后，使用 SWA 策略更新最终模型。

B 补充实验

B.1 与 bSAM 方法的比较

我们在 Resnet18 上进行实验，将我们的平坦贝叶斯神经网络与 bSAM 进行比较 [40]，结果如表6所示。bSAM 的作者探讨了 SAM 与贝叶斯神经网络之间的关系，并提出将 SAM 与 Adam 结合起来，在固定方差的同时优化贝叶斯神经网络中的参数均值。结果清楚地表明，在大多数指标得分上，我们的平坦贝叶斯神经网络优于 bSAM。需要指出的是，由于作者未公开代码，我们无法在主文表1和表2所使用的架构上评估 bSAM。因此，我们按照 bSAM 论文中提到的设置运行了我们的方法。

表6：Resnet18 上的分类得分

方法	CIFAR-10 准确率 ↑	CIFAR-10 负对数似然 ↓	CIFAR-10 期望校准误差 ↓	CIFAR-100 准确率 ↑	CIFAR-100 负对数似然 ↓	CIFAR-100 期望校准误差 ↓
bSAM	96.15	0.1200	0.0049	80.22	**0.7000**	0.0310
F-SWAG-Diag	96.56	0.1047	**0.0037**	80.70	0.7012	**0.0227**
F-SWAG	**96.58**	**0.1045**	0.0045	**80.74**	0.7024	0.0243

B.2 分布外预测的完整结果

在主论文的第4.2节中，我们全面分析了模型在多种损坏类别上的表现，包括噪声、模糊、天气条件和数字失真。我们在表 7中列出了每种损坏类型的详细结果，从而更深入地了解这些损坏对模型性能的影响。平均而言，平坦贝叶斯神经网络优于非平坦贝叶斯神经网络，尤其是在期望校准误差方面，优势显著。这些发现进一步凸显了平坦贝叶斯神经网络在提升模型应对各种损坏时的鲁棒性和泛化能力方面的有效性。

B.3 补充消融研究

海森特征值比较 我们报告了采用 CIFAR-100 的 WideResNet28x10 在多种方法下，海森矩阵最大特征值的对数尺度，以及最大特征值与第五大特征值的比值，如表 8所示。这明显表明，我们的方法会将模型更新至曲率较低的极小值点。

计算成本 我们的平坦寻优方法需要计算两次梯度：首先获取扰动后的模型 θ'，随后更新模型。因此，如表9所示，与非平坦方法相比，训练时间几乎翻倍。请注意，深度集成设置使用多个模型分别进行训练以完成预测，我们报告的是每种设置下单个模型的训练时间。

在 KL 深度集成设置中的术语的影响 我们在表A.2 中给出了按照第 10节中平坦版本的公式，使用 SAM 训练深度集成的结果，但未使用 KL（或 L2 正则化）。每项实验均重复三次，并报告均值和标准

表7：使用 CIFAR-10 数据集训练的 PreResNet-164 模型在 CIFAR-10-C 上的分类得分

	期望校准误差 ↓				准确率 ↑			
方法	SWAG-D	F-SWAG-D	SWAG	F-SWAG	SWAG-D	F-SWAG-D	SWAG	F-SWAG
高斯噪声	0.0765	0.0765	0.1032	0.0091	72.01	73.95	71.58	73.43
散粒噪声	0.0661	0.0647	0.0892	0.0075	75.77	77.09	75.44	76.28
斑点噪声	0.0711	0.0686	0.0921	0.0072	75.55	76.71	75.36	76.15
脉冲噪声	0.0779	0.0706	0.0988	0.0077	73.74	74.61	73.71	74.49
失焦模糊	0.0108	0.0071	0.0178	0.0256	92.16	91.55	92.14	91.63
高斯模糊	0.0130	0.0116	0.0214	0.0239	90.79	89.73	90.67	90.22
运动模糊	0.0147	0.0103	0.0233	0.0298	90.33	90.22	90.20	90.52
缩放模糊	0.0099	0.0070	0.0185	0.0301	91.24	90.71	91.12	91.36
snow	0.0298	0.0245	0.0419	0.0208	86.17	86.11	86.10	85.81
Fog	0.0114	0.0075	0.0176	0.0259	91.67	91.22	91.64	91.27
亮度	0.0081	0.0076	0.0129	0.0281	93.47	92.94	93.45	92.98
对比度	0.0110	0.0127	0.0141	0.0306	91.34	90.59	91.36	90.73
弹性变换	0.0244	0.0213	0.0367	0.0220	87.06	86.61	86.98	86.94
像素化	0.0350	0.0269	0.0463	0.0124	85.75	86.11	85.61	85.74
JPEG压缩	0.0605	0.0522	0.0813	0.0093	78.01	78.90	77.57	79.80
飞溅	0.0242	0.0163	0.0341	0.0227	87.94	87.61	87.99	88.13
饱和度	0.0112	0.0080	0.0179	0.0288	92.10	91.79	92.09	91.81
霜	0.0253	0.0174	0.0365	0.0215	86.60	86.48	86.52	86.39
平均值	0.0322	**0.0283**	0.0446	**0.0201**	85.65	**85.71**	85.52	**85.76**

表8：WideResNet28x10 在 CIFAR-100 上训练时海森特征值的对数尺度。 λ_1 是最大特征值， λ_5 是第 5 大特征值

方法	$\lambda_1 \downarrow$	λ_1/λ_5
SWAG	4.17 ± 0.001	1.17 ± 0.012
F-SWAG	$\mathbf{4.08 \pm 0.000}$	1.17 ± 0.020
SGLD	3.34 ± 0.031	1.17 ± 0.009
F-SGLD	$\mathbf{2.83 \pm 0.029}$	1.15 ± 0.010
深度集成	4.64 ± 0.055	1.45 ± 0.020
F-深度集成	$\mathbf{4.01 \pm 0.054}$	1.58 ± 0.032

表9：每轮训练时间比较

网络与数据集	SWAG	F-SWAG	SGLD	F-SGLD	深度集成	F-深度集成
WideResNet28x10 和 CIFAR-100	110s	169s	110s	233s	110s	218s
Densenet-161 & ImageNet	1.75 小时	2.28 小时	1.78 小时	2.49 小时	-	-
ResNet-152 & ImageNet	1.59 小时	2.15 小时	1.64 小时	2.22 小时	-	-

差。根据结果，即使没有 KL 损失，我们的方法仍能取得比非平坦对应方法更好的结果。

C 实验设置

CIFAR： 我们在 CIFAR-10 和 CIFAR-100 上分别使用 PreResNet-164、WideResNet28x10、Resnet10 和 Resnet18 开展实验。这些数据集的图像总数为 60,000，其中 50,000 张用于训练，10,000 张用于测试。对于每个网络-数据集组合，我们将锐度感知贝叶斯方法应用于多种设置，包括 F-SGLD、F-SGVB、F-SWAG-Diag、F-SWAG、F-MC-Dropout 和 F-深度集成。

在主论文表1和表2所列的实验中，我们使用随机梯度下降训练所有模型 300 个周期，学习率设为 0.1，并采用余弦调度。对于 F-SWA 和 F-SWAG 设置，我们从第 161 个周期后开始收集模型，这与 [39]中的方案一致。此外，在所有实验中，我们为 CIFAR-10 设置 $\rho = 0.05$ ，为 CIFAR-100 设置 $\rho = 0.1$ ，但 Resnet10 和 Resnet18 除外，它们的 ρ 设为 0.01。我们使用基本的数据增强来扩充训练集，包括水平翻转、填充四个像素、随机裁剪和归一化。对于

表10：使用 WideResNet28x10 在 CIFAR-100数据集上进行的 F-深度集成变体实验。每项实验使用三个不同的随机种子进行，以计算均值和标准差

	WideResNet28x10		
模型	准确率 ↑	负对数似然 ↓	期望校准误差 ↓
深度集成	83.04 ± 0.15	0.6958 ± 0.0335	0.0483 ± 0.0017
F-深度集成 **(我们的)**	**84.52 ± 0.03**	**0.5644 ± 0.0106**	**0.0191 ± 0.0039**
F-深度集成（不含 L2）	83.80 ± 0.10	0.7026 ± 0.0007	0.0594 ± 0.0005

主论文表3所列的实验，我们对训练集采用与表1和表2中的实验相同的数据增强。不过，模型使用 Adam优化器训练 200 个周期，学习率为 0.001，并采用平台期调度。值得注意的是，SGVB 和 SGVB-LRT 在其他设置下表现不佳，因此很难扩展这种方法。

对于深度集成、随机梯度朗之万动力学、SGVB 和 SGVB-LRT 方法的基线，我们按照平坦版本的超参数和流程复现实验结果。请注意，对于深度集成方法，我们训练了三个相互独立的模型。推理时，对于所有设置，我们都对从后验分布中采样得到的 30 个模型样本进行集成。为确保方法的稳定性，我们使用不同的随机种子，将每组实验重复三次，并报告均值和标准差。

ImageNet： 这是一个包含 1000 个类别、大规模且具有挑战性的数据集。我们使用 Densenet-161 和 ResNet-152 架构，在 F-SWAG-Diag、F-SWAG 和 F-SGLD 上进行实验。对于所有设置，我们都使用在 ImageNet 数据集上预训练的权重初始化模型，这些权重来自 *torchvision* 包，随后使用 $\rho = 0.05$ 微调 10 个 epoch。在微调过程开始时，我们开始每个 epoch 收集 4 个模型，并按照与 CIFAR 数据集实验一致的方案对其进行评估。
SWAG-Diag、SWAG 和蒙特卡洛 Dropout 方法的性能指标引自Maddox等人的原始论文。[39], 但对于 CIFAR-100数据集上 PreResNet-164 的蒙特卡洛 Dropout 结果，由于原论文未提供该结果，我们进行了复现。bSAM 的性能数据取自 [40]。

需要注意的是，这些实验的目的并非取得最先进性能，而是展示将尖锐度感知后验与特定贝叶斯神经网络结合时的实用价值。实现代码见https://github.com/anh-ntv/flat_bnn.git。

训练超参数：表11 提供了训练和测试阶段的设置。请注意，SWAG-Diag 方法采用与 SWAG 相同的设置。通常，使用默认$\rho = 0.05$ 即可在所有实验中取得良好性能。不过，对于 CIFAR-100 数据集，建议在 $\rho = 0.1$ 中使用 [16]。模型评估时，我们使用最后一个训练轮次的检查点，不考虑验证集的性能。

表 11：训练平坦和非平坦版本的贝叶斯神经网络时使用的超参数。除 SGVB 和 SGVB-LRT 的实验使用输入分辨率 32×32 外，所有模型均使用输入分辨率 224×224 和余弦学习率衰减进行训练

模型	方法	初始化权重	轮数	初始学习率	权重衰减	ρ	# 样本
		CIFAR-100					
PreResNet-164	SWAG						30
	MC-Drop	从头训练	300	0.1	3e-4	0.1	30
	深度集成						3
WideResNet28x10	SWAG						30
	MC-Drop	从头训练	300	0.1	5e-4	0.1	30
	深度集成						3
Resnet10	SGVB					5e-3	
	F-SGVB + 几何					5e-4	
	SGVB-LRT	从头训练	200	0.001	5e-4	5e-3	30
	F-SGVB-LRT + 几何					5e-4	
Resnet18	SGVB					5e-3	
	F-SGVB + 几何					5e-4	
	SGVB-LRT	从头训练	200	0.001	5e-4	5e-3	30
	F-SGVB-LRT + 几何					5e-4	
	SWAG	从头训练	300	0.1	5e-4	0.1	30
		CIFAR-10					
PreResNet-164	SWAG						30
	MC-Drop	从头训练	300	0.1	3e-4	0.05	30
	深度集成						3
WideResNet28x10	SWAG						30
	MC-Drop	从头训练	300	0.1	5e-4	0.05	30
	深度集成						3
Resnet10	SGVB					5e-3	
	F-SGVB + 几何					5e-4	
	SGVB-LRT	从头训练	200	0.001	5e-4	5e-3	30
	F-SGVB-LRT + 几何					5e-4	
Resnet18	SGVB					5e-3	
	F-SGVB + 几何					5e-4	
	SGVB-LRT	从头训练	200	0.001	5e-4	5e-3	30
	F-SGVB-LRT + 几何					5e-4	
	SWAG	从头训练	300	0.1	5e-4	0.1	30
		ImageNet					
Densenet-161	所有方法	预训练	10	0.001	1e-4	0.05	30
ResNet-152	所有方法	预训练	10	0.001	1e-4	0.05	30