arXiv:2602.06269v1 [cs.LG] 6 Feb 2026

PurSAMERE：通过尖锐度感知期望重构误差最小化实现可靠对抗净化

Vinh Hoang[1,2],Sebastian Krumscheid[3],Holger Rauhut[4],Ra ´ul Tempone[5]

[1]德国亚琛工业大学数学系[2]于利希研究中心可持续氢经济研究所（IHE），德国¨卡尔斯鲁厄理工学院科学计算中心与应用数学和数值数学研究所，德国[3]慕尼黑大学数学系及慕尼黑机器学习中心，德国[4]阿卜杜拉国王科技大学计算机、电气与数学科学及工程学院，沙特阿拉伯

vi.hoang@fz-juelich.de, sebastian.krumscheid@kit.de,
rauhut@math.lmu.de, raul.tempone@阿卜杜拉国王科技大学.edu.sa

2026年2月9日

摘要

我们提出一种新颖的确定性净化方法，通过将潜在的对抗样本映射到附近一个接近数据分布众数的样本，提升对抗鲁棒性；分类器在这些样本上的表现更可靠。我们将该方法设计为确定性的，以确保测试准确率可靠，并防止在攻击者完全了解系统及其随机性的情况下，随机净化方法所观察到的有效鲁棒性下降。我们采用通过最小化受噪声污染数据的期望重建误差而训练的分数模型，从而学习输入数据分布的结构特征。给定一个潜在的对抗输入，该方法会在其局部邻域内搜索一个净化样本，使其在噪声扰动下的期望重建误差最小，然后将该净化样本输入分类器。在净化过程中，我们使用尖锐度感知最小化，引导净化样本趋向期望重构误差景观中的平坦区域，从而增强鲁棒性。我们进一步表明，随着噪声水平降低，最小化期望重建误差会使净化样本偏向高斯平滑密度的局部最大值点；在对分数模型施加额外局部假设的条件下，我们证明在小噪声极限下可以恢复一个局部最大值点。实验结果表明，在强确定性白盒攻击下，与最先进的方法相比，本方法显著提升了对抗鲁棒性。

1 引言

神经网络在计算机视觉、自然语言处理和强化学习等多个领域取得了显著成功。然而，众所周知，它们容易受到对抗攻击：对输入数据施加微小扰动，就可能导致错误预测 [22, 7], ，这给评估神经网络在现实应用中的可靠性带来了挑战。为解决这一问题，增强神经网络对抗鲁棒性的常见方法有两种：对抗训练和对抗净化。对抗训练通过扰动原始数据来生成对抗样本，并将其加入训练数据集 [14]。对抗净化则旨在将输入数据送入分类器之前，去除其中的对抗扰动。对抗净化主要分为两类：基于流形的方法和基于密度的方法。基于流形的净化方法假设干净数据位于嵌入高维输入空间中的低维流形上。净化过程包括将对抗样本投影到该流形上，以恢复干净数据。Samangouei 等人 [17] 提出了 Defense-GAN，该方法使用生成对抗网络（GAN） [6] 学习数据流形，并通过寻找 GAN 流形上距离对抗样本最近的点来净化对抗样本。Shi 等人 [18] 提出了一种基于自编码器的在线对抗净化方法，通过最小化重构误差，从对抗样本中重建干净数据。基于密度的净化方法旨在将对抗样本移向数据密度较高的区域，因为分类器在这些区域预计更为鲁棒。Song 等人 [21] 提出使用具有可处理似然的生成模型 [23, 16] 净化对抗样本，即将其移向数据密度的局部最大值点。在 [26], ，Yoon等人使用基于分数的生成模型，通过从随机起点出发的简化朗之万动力学来净化对抗样本。近期，研究者提出了基于扩散的净化方法，该方法使用扩散模型对对抗样本进行去噪，并在 [15] 中进一步提出，随后在 [12]中得到改进。

对抗净化可能会造成梯度混淆，这种现象在严重依赖随机性的净化方法中尤为明显。Athalye等人 [2] 证明，梯度混淆可能导致测试准确率不可靠。本文旨在开发一种确定性净化方法，不仅能增强对抗鲁棒性，还能在强白盒对抗攻击下提供可靠的鲁棒性。

自适应攻击是一种专为基于净化的防御设计的对抗攻击，旨在解决梯度混淆问题 [2]。对于涉及随机性的净化方法，可将变换期望（EoT）技术与自适应攻击结合，以提高其有效性。更强的攻击是白盒确定性自适应攻击，在这种攻击中，攻击者不仅完全了解分类器和净化模型，也完全了解随机生成器。Li等人 [12] 表明，在结合 EoT 的强自适应攻击下，随机净化方法的对抗准确率会大幅下降。此外，Liu 等人 [13] 通过确定性白盒攻击进行了数值测试，并表明随机净化方法可能会损失大部分对抗准确率。这些结果表明，只有在强自适应攻击和确定性白盒攻击下均能保持鲁棒性，净化方法才可被视为有效。

本文提出了一种新颖的确定性净化方法，称为通过期望重构误差的尖锐度感知最小化进行净化（Purification via Sharpness-Aware Minimization of the Expected Reconstruction Error，PurSAMERE）。我们的方法通过寻找一个邻近样本，使受噪声污染数据的期望重构误差最小化，从而净化对抗样本。该误差使用基于分数的生成模型进行评估。

所提出的净化方法倾向于选择位于数据密度呈凹形的区域内的净化样本。我们证明，在某些假设下，当噪声协方差趋近于零时，净化样本在渐近极限下会收敛到数据密度函数的局部最大值。我们方法背后的直觉是：对于数据密度函数值较高的输入，分类器往往更加准确；并且局部密度最大值对于原始输入的扰动具有稳定性。此外，我们在净化过程中使用尖锐度感知最小化（SAM）[4]，通过推动净化样本趋向期望重构误差平坦的区域来增强其鲁棒性。

由于期望重建误差是通过蒙特卡洛方法进行经验估计的，因此会引入少量随机性。在所有实验中，我们固定蒙特卡洛噪声样本，使净化过程完全确定。我们的确定性白盒评估假设攻击者完全了解这一固定随机性。

不同于以往仅关注流形一致性或仅关注密度最大化的方法，PurSAMERE 采用了一种混合方法，将这两个目标结合起来：它推动净化样本趋向数据密度函数的局部极大值点，同时确保净化样本尽可能接近由期望重建误差隐式定义的流形。我们的方法具有三项主要优势：

- 首先，该方法在设计上具有确定性；因此，在确定性白盒攻击下，其测试时鲁棒性准确率不会下降，因而更加可靠。此外，在评估确定性净化方法的鲁棒性时无需使用期望变换，从而降低了鲁棒性评估的计算成本。

- 其次，通过最小化期望重建误差，我们的方法将净化样本推动至高斯平滑分布的高密度区域；并且在附加的局部结构假设下，可以得到模型势函数局部最大值点的恢复结果。

- 第三，通过在净化过程中采用 SAM，我们的方法有效地减轻了近似误差的影响。学习得到的评分函数中的这些误差可能会在期望重构误差景观中引入伪局部极小值。这增强了净化过程的可靠性，并有助于提高对抗鲁棒性。

2 相关工作

通过条件期望实现生成模型去噪自编码器（DAE）最初被提出作为一种表征学习方法，用于学习从含噪观测中恢复干净数据，从而增强对扰动的鲁棒性 [25]。实际上，DAE 是给定受噪声污染的数据时条件期望的一种近似。此外，后续的理论分析 [1, 24] 表明，在破坏程度较小时，DAE 能够学习底层数据分布的得分，从而建立起与分数匹配之间的联系 [9, 10]。这一观点后来推动并影响了基于得分的生成模型和扩散生成模型的发展 [19, 20, 8], ，后者已在图像生成任务中取得了最先进的结果。

对抗净化 Song 等人 [21]使用具有可处理似然的生成模型 [23,16] ，通过将对抗样本移向数据密度的局部极大值来对其进行净化。Samangouei 等人 [17] 采用生成对抗网络（GANs） [6]来净化对抗样本

通过在 GAN 流形上寻找距离对抗样本最近的点来实现。如 [2] 所报告的那样，在强自适应攻击下，这两种方法的准确率均显著下降。我们的方法与这两种方法动机相似，即旨在找到数据密度的一个局部最大值，而该点同时也是输入空间中到由期望重建误差隐式定义的流形距离最小的样本。然而，由于以下三个方面的结合，我们的方法更加稳健：基于分数的生成模型能够捕捉数据分布的复杂特性；对噪声扰动取期望会驱动净化样本趋向密度函数呈凹的局部极大值；并且在净化过程中使用 SAM 有助于减轻学习得到的评分函数中近似误差的影响。Shi 等人 [18] 提出了多种对抗净化方法，其中包括一种基于最小化自编码器重构误差的方法。我们的方法与之不同，它最小化的是噪声扰动下的期望重建误差，而非干净样本或单个含噪样本的重构误差。这一特定形式使我们能够将净化样本与数据密度函数平滑近似的局部最大值联系起来。此外，不同于 [18]，的基于自编码器的方法难以很好地泛化到 CIFAR-10 等复杂数据集，我们基于分数的方法能够有效捕捉复杂的数据特征。Yoon等人 [26]使用基于分数的生成模型，借助带随机初始化的简化朗之万动力学来净化对抗样本。Nie 等人 [15] 提出了基于扩散的净化方法，通过扩散模型对对抗样本进行去噪。这两种方法在净化过程中都包含大量随机性。相比之下，我们的方法是确定性的，不仅能取得更高的对抗准确率，而且更加可靠。

对抗攻击 针对图像分类器的常见攻击包括快速梯度符号法（FGSM） [7]，、投影梯度下降（PGD） [14]，和 Carlini & Wagner（C&W） [3]攻击。这些攻击通过利用梯度信息优化扰动，同时对输入施加约束，从而构造对抗样本。对于包含净化等预处理步骤的模型，梯度可能会被遮蔽；为此，人们提出了自适应攻击 [2] 。期望变换通常与自适应攻击结合使用，以提高针对随机净化过程的攻击效果。更强的一种攻击方式是确定性自适应攻击：攻击者完全了解分类器、净化过程以及随机数生成器。此外，净化过程中涉及的随机性会被固定，使净化过程变为确定性。尽管这种确定性设定可能不太实用，但它无需使用期望变换，可降低测试时开销，并能给出对抗准确率的可靠下界。

3 方法

3.1 通过基于分数的生成模型学习数据分布

在这项工作中，我们聚焦于图像分类任务。尽管如此，所提出的净化方法具有通用性，也可扩展到其他数据模态。令 X 表示与输入图像对应的随机变量，其取值属于紧集 $\mathcal{X} = [0,1]^d$，其中 d 是输入图像的维度。我们假设 X 存在概率密度函数，记为 p_X。令 x 为一幅干净图像，并令 x_{adv} 表示相对于分类器 h、由 x 得到的对抗样本。相应的对抗扰动定义为 $\delta_{\text{adv}} := x_{\text{adv}} - x$，并假设其满足 $\|\delta_{\text{adv}}\|_p \le \varrho_{\text{adv}}$，其中 ϱ_{adv} 表示对抗扰动预算。 p 的典型取值为 1，

2 和 ∞。净化的目标是通过净化映射 Pur，从对抗版本 x_{adv} 中恢复干净图像 x，使得 Pur(x_{adv}) $\approx x$。

一类对抗净化方法基于条件期望模型学习数据特征，随后利用该模型净化对抗样本。令 Y_σ 表示 X的噪声污染版本，定义如下：

$$Y_\sigma = X + \sigma \Xi, \tag{1}$$

其中，$\Xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ 是一个与 X独立的高斯随机向量。Y_σ 的概率密度函数由 p_X 与高斯核的卷积给出：

$$p_{Y_\sigma}(y) = \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|y-x\|_2^2}{2\sigma^2}\right) p_X(x)\,\mathrm{d}x. \tag{2}$$

当 σ足够小时，p_{Y_σ} 是数据密度 p_X的平滑近似。

条件期望 $\mathbb{E}[X \mid Y_\sigma]$ 是给定 Y_σ时 X 的最小均方误差（MMSE）估计量，即：

$$\mathbb{E}[X \mid Y_\sigma] = \arg\min_{g \in L_2(\mathbb{R}^d,\, p_{Y_\sigma})} \mathbb{E}\left[\|X - g(Y_\sigma)\|_2^2\right], \tag{3}$$

其中，$L_2(\mathbb{R}^d,\ p_{Y_\sigma})$ 表示可测函数 $g\colon \mathbb{R}^d \to \mathbb{R}^d$ 构成的空间，这些函数关于 p_{Y_σ}平方可积。根据 Tweedie公式，条件期望可表示为：

$$\mathbb{E}[X \mid Y_\sigma = y] = y + \sigma^2 s(y;\sigma), \tag{4}$$

其中，$s(\cdot;\sigma)$ 表示得分函数，定义为：

$$s(y;\sigma) = \nabla_y \log p_{Y_\sigma}(y). \tag{5}$$

我们使用深度神经网络 s_θ 来近似得分函数 s。基于式 (3) 中条件期望的 MMSE 表征和 Tweedie公式 (4)，可以通过最小化去噪分数匹配目标 [24],来学习得分函数

$$\theta = \arg\min_{\theta'} \mathbb{E}_{X,\Xi}\left[\|\Xi + \sigma\, s_{\theta'}(X + \sigma\Xi;\sigma)\|_2^2\right], \tag{6}$$

其中，期望通过对训练数据和噪声样本求经验平均来近似。实际上，求解式 (6) 等价于最小化真实得分函数 s 与分数模型 s_θ之间的期望平方误差，因为

$$\mathbb{E}_{X,\Xi}\left[\|\Xi + \sigma\, s_\theta(X + \sigma\Xi;\sigma)\|_2^2\right] = \sigma^2 \mathbb{E}_{X,\Xi}\left[\|s(X+\sigma\Xi;\sigma) - s_\theta(X+\sigma\Xi;\sigma)\|_2^2\right] + \mathrm{Const}, \tag{7}$$

其中，常数项与 θ无关。这一等价性的证明见附录 A.1。实践中，会使用多个 σ 值来学习不同噪声水平下的得分函数，这有助于更准确地捕获数据分布 [19, 20]。

一旦学得分数函数，就可以用它来净化对抗性输入。例如，在 [26], Yoon等人使用简化版的朗之万动力学从分布中采样p_{Y_σ}。该方法将分数函数用作对数密度的梯度，迭代地更新对抗性输入，使其趋向概率密度较高的区域。这种方法的依据是近似式 $p_X \approx p_{Y_\sigma}$，其中 σ 的值足够小。然而，该算法仅仅是在寻找

分数函数为零的点，这些点可能对应于密度的鞍点p_{Y_σ}，甚至可能是学习得到的分数函数中由分数模型的近似误差 s_θ 导致的伪零点。

在本文中，我们提出了一种基于期望重建误差局部最小化的新型净化方法。我们的方法将净化样本推向密度的局部极大值点 p_{Y_σ}，同时减轻得分函数近似误差的影响。

3.2 通过期望重建误差的尖锐度感知最小化进行净化（PurSAMERE）

我们将给定 x 的期望重建误差定义为

$$R(x;\sigma) = \mathbb{E}_{\Xi}\left[\|\Xi + \sigma s(x+\sigma\Xi;\sigma)\|_2^2\right], \quad \Xi \sim \mathcal{N}(0, \mathbf{I}_d). \tag{8}$$

根据 Tweedie公式， $\mathbb{E}[X \mid Y = x+\sigma\Xi] = x + \sigma\Xi + \sigma^2 s(x+\sigma\Xi;\sigma)$，函数 R 可表示为

$$R(x;\sigma) \equiv \frac{1}{\sigma^2}\mathbb{E}_{\Xi}\left[\|x - \mathbb{E}[X \mid Y_\sigma = x+\sigma\Xi]\|_2^2\right], \tag{9}$$

该误差衡量了在加性噪声扰动 $\mathbb{E}[X \mid Y_\sigma = x+\sigma\Xi]$下， x与其条件期望重建之间的平方距离的期望。

从（9）可以看出，我们对期望重建误差的表述经过了 σ^2归一化。这种归一化有助于稳定 PurSAMERE 中使用的数值优化过程；该过程采用多个噪声尺度，并依次将其减小至零。

关于 X 的 $R(X;\sigma)$ 的期望为

$$\mathbb{E}_X[R(X;\sigma)] \;=\; \frac{1}{\sigma^2}\,\mathbb{E}\big[\|X - \mathbb{E}[X \mid Y_\sigma]\|_2^2\big]\,. \tag{10}$$

由于 $\mathbb{E}[X \mid Y_\sigma]$ 是给定 Y_σ 时 X 的 MMSE估计器，我们可以将其与平凡估计器 $g(Y_\sigma) = Y_\sigma$ 进行比较，从而得到

$$\mathbb{E}\big[\|X - \mathbb{E}[X \mid Y_\sigma]\|_2^2\big] \;\le\; \mathbb{E}\big[\|X - Y_\sigma\|_2^2\big] \;=\; \sigma^2\mathbb{E}\big[\|\Xi\|_2^2\big] \;=\; \sigma^2 d, \tag{11}$$

因此， $\mathbb{E}_X[R(X;\,\sigma)] \le d$。由于经过 σ^2 归一化， $R(x;\,\sigma)$ 的基线取决于维度。因此，在净化过程中，我们依赖其相对于 x 的变化。经验上，在局部邻域内，对于分布内输入， $R(x;\sigma)$ 中依赖于 x 的部分通常较小；而对于分布外输入或对抗性输入，该部分通常较大。

受这一观察启发，我们提出通过在 x_{adv} 的一个小邻域内最小化期望重建误差 $R(x;\sigma)$，来净化给定的、可能带有对抗性的样本 x_{adv}。此外，我们采用尖锐度感知最小化（SAM） [4]，使净化样本趋向于期望重建误差的平坦区域，从而增强净化的鲁棒性，并避免伪极小值。该净化过程表述为以下约束优化问题：

$$x^\star = \arg\min_{x\in[0,1]^d} \max_{\|\varepsilon\|_q \le \varrho_{\mathrm{sam}}} R(x+\varepsilon;\sigma), \quad \text{s.t.} \quad \|x - x_{\mathrm{adv}}\|_q \le \varrho_{\mathrm{pur}}, \tag{12}$$

其中 q 的典型取值为 1、2 和 ∞。在本文中，我们设定 $q = 2$。这里， $\varrho_{\mathrm{pur}} > 0$ 是净化半径，用于控制搜索邻域 $B_2(x_{\mathrm{adv}},\ \varrho_{\mathrm{pur}})$ 的大小； $\varrho_{\mathrm{sam}} > 0$是尖锐度半径，用于控制 SAM 扰动邻域 $B_2(x,\ \varrho_{\mathrm{sam}})$ 的大小。

不同于在训练期间对模型参数进行操作的传统 SAM，我们在净化过程中将类似 SAM 的最坏情况目标应用于输入空间。

这一净化时SAM 步骤在 x_{adv} 的邻域内搜索一个点 $x^{\star}$，即使在最坏情况输入扰动下也能最小化 R_θ，从而确保该极小化点位于 R_θ 的相对平坦盆地中。给定通过 (6) 学习得到的分数模型 s_θ，我们将$R(x;\sigma)$ 近似为

$$R_\theta(x;\sigma) = \mathbb{E}_{\Xi}\left[\|\Xi + \sigma s_\theta(x+\sigma\Xi;\sigma)\|_2^2\right]. \tag{13}$$

该净化准则有两个主要优点。

- 首先，在 x_{adv} 的邻域内最小化 $R(x;\sigma)$，会促使净化器在条件期望模型下找到一个重构误差较小的邻近点 $x^{\star}$。这有效地将 $x^{\star}$ 导向平滑密度 p_{Y_σ} 的局部最大值点，前提是该点存在于 ℓ_2-球 $B_2(x_{\mathrm{adv}}\ \varrho_{\mathrm{pur}})$ 内，如第 3.3 节所述。

- 其次，基于 R_θ 的净化对学习得到的评分函数 s_θ 中的近似误差较为敏感，这些误差可能引入 $R_\theta(x;\sigma)$ 的伪局部极小值，而这些极小值并不对应于真正的密度极大值。为缓解这一问题，我们采用 SAM 将净化样本引导至 $R_\theta(x;\sigma)$ 的平坦区域，从而有助于抑制由近似误差引起的尖锐极小值。

3.3 净化样本与数据密度局部极大值的关系

在本节中，我们分析式 (8) 中定义的期望重建误差$R(x;\sigma)$的极小化点的行为。具体而言，我们证明，最小化 $R(x;\ \sigma)$ 会使 x 趋向于对数密度 $\log p_X$ 的局部极大值点，如图 1 所示。

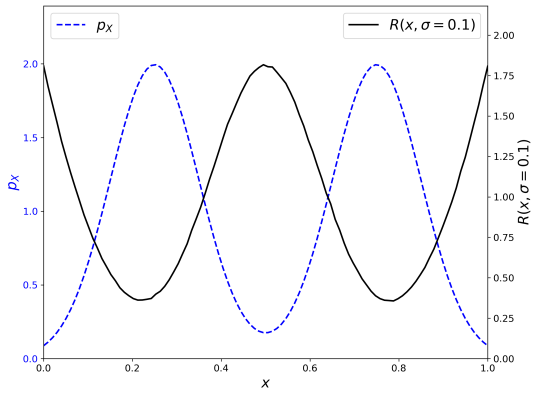

图1：一维截断高斯混合模型的期望重建误差 $R(x;\sigma)$，$p_X(x) \propto 0.5\ \mathcal{N}(x;\ 0.25, 0.1) + 0.5\ \mathcal{N}(x;\ 0.75, 0.1)$。带$\sigma = 0.1$ 的期望重建误差 $R(x;\ \sigma)$（右）近似在密度$\log p_X$ （左）的局部极大值处取得局部极小值点。

命题 1 (期望重构误差展开)。 设 $s:\ \mathbb{R}^d \times [0,\sigma_{\max}] \to \mathbb{R}^d$，且$\Xi \sim \mathcal{N}(0,\ I_d)$。假设对于每个 $x \in \mathbb{R}^d$和 $\xi \in \mathbb{R}^d$，映射 $\sigma \mapsto s(x+\sigma\xi,\sigma)$ 在 $[0,\sigma_{\max}]$上二阶连续可微。定义

$$\phi(x,\sigma,\xi) := \frac{d^2}{d\sigma^2} s(x+\sigma\xi,\sigma).$$

固定一个紧集 $K \subset \mathbb{R}^d$。假设存在常数 $C>0$和 $m \geq 0$，使得对于所有$x \in K$、所有 $\sigma \in [0,\sigma_{\max}]$以及所有 $\xi \in \mathbb{R}^d$，

$$\max_{1\leq i\leq d} |\phi_i(x,\sigma,\xi)| \ \leq\ C\big(1+\|\xi\|_2^m\big).$$

那么，当 $\sigma \to 0$时，

$$R(x;\sigma) = d + \sigma^2\|s(x,0)\|_2^2 + 2\sigma^2\ \mathrm{tr}\big(\nabla_x s(x,0)\big) + \mathcal{O}(\sigma^3), \quad x \in K, \tag{14}$$

其中 $\nabla_x s(x;0)$表示 $s(x;0)$关于 x的雅可比矩阵。

命题 1 的证明见附录 A.2。为了使 $s(x,\sigma)$在 $\sigma=0$ 处对所有 $x \in \mathbb{R}^d$二阶连续可微， p_X必须在 $\mathbb{R}^d$上严格为正。对于紧支撑密度 p_X，得分函数 $s(x,0)$在其支撑集之外未定义。在这种情况下，我们将 $s(x,0)$视为具有固定小噪声尺度 $\sigma_{\min} \ll 1$的高斯平滑密度的得分函数，因此 $s(x,0)$对所有 $x \in \mathbb{R}^d$都有定义。

假设命题 1 的条件成立，且 σ 足够小，让我们分析命题 1 中式 (14) 右侧的各项。第一项 d与 x无关，因此不会影响净化过程。第二项 $\sigma^2\|s(x;0)\|_2^2$ 会惩罚得分幅度较大的点。最小化这一项会使 $s(x;0)\ \to 0$，这对应于对数密度 $\log\ p_X(x)$ 的驻点。由于驻点可能对应极大值点、极小值点或鞍点，因此仅凭这一项不足以刻画典型的数据样本。

第三项提供了额外的几何判别。由于 $s(x,0)\ =\ \nabla_x \log\ p_X(x)$，第三项 $2\sigma^2\,\mathrm{tr}(\nabla_x s(x;0))$ 等于 $2\sigma^2\,\mathrm{tr}(\nabla_x^2 \log\ p_X(x))$，与 $\log\ p_X(x)$ 的海森矩阵的迹成正比。尽管这一项无法保证 $\log\ p_X(x^\star)$的海森矩阵是负定的，但将其最小化会使 x 倾向于 $\log\ p_X(x)$ 局部凹的区域。因此，最小化 $R(x;\sigma)$会使 x 趋向对数密度 $\log\ p_X(x)$ 的局部极大值，从而趋向 $p_X(x)$ 的峰值，前提是搜索邻域内存在此类峰值。因此，我们为净化选择的目标函数 $R(x;\sigma)$比单纯像 [26] 中那样强制 $s(x;\sigma)\ =0$更具判别性。

为从理论上理解 $R_\theta(x;\sigma)$（13）的极小化点的行为，我们考察 s_θ 关于 x分段线性的情形。该假设涵盖某些采用 ReLU 激活函数且不含层归一化的神经网络。在某些假设下，定理 1 表明，当 $\sigma \to 0$时， $R_\theta(x;\sigma)$的极小化点以指数速率收敛到对数密度 f_θ的局部最大值点，其中 $s_\theta = \nabla_x f_\theta$。定理 1 的详细陈述与证明见附录 A.3。

定理 1（密度最大值点的局部恢复（非正式表述））。 考虑一个分数模型 $s_\theta(x,\ \sigma)$，它是分段仿射的，源自势函数 f_θ （因此 $s_\theta\ =\ \nabla_x f_\theta$），并且在区域 $D \in [0,1]^d$内部局部 μ-强凹。假设 $f_\theta(\cdot,\ \sigma)$的最大值点 $x_\sigma^\star$严格位于

D内部，且至少距离其边界 $\sqrt{d}\varrho$ 个 ℓ_2。令 x'_σ 为 R_θ（$\cdot$，σ）在$D_{-\sqrt{d}\,\varrho}$上的任意极小化点，其中

$$D_{-\sqrt{d}\,\varrho} := \{x \in D : \inf_{y \notin D} \|x - y\|_2 \geq \sqrt{d}\,\varrho\}$$

假设其非空。若噪声水平满足 $\sigma < \rho$，则

$$\|x'_\sigma - x^\star_\sigma\|_2 \;\leq\; \frac{\sqrt{C}}{\sigma\mu} \exp\left(-\frac{d}{8}\left(\frac{\varrho}{\sigma} - 1\right)^2\right),$$

其中 $C > 0$ 是一个与 σ 无关的常数。特别地，当 $\sigma \to 0$ 时， x'_σ 以指数速度收敛到 $x^\star_\sigma$ 。

在定理 1 的假设下，关于 x 最小化期望重建误差 $R_\theta(x;\sigma)$，可恢复 $f_\theta(\cdot,\sigma)$ 的局部最大值点，当 $\sigma \to 0$ 时成立。在此极限下， $f_\theta(\cdot,\sigma)$ 收敛到对数数据密度 $\log p_X$，这意味着 $R_\theta(x;\sigma)$ 的极小化点趋近于真实密度最大值点。实践中，我们使用 $\sigma \geq \sigma_{\min} > 0$，因此定理的条件是理想化的；尽管如此，该定理支持这样一种直觉：我们的净化器会将输入推向 p_X 的高密度区域。这些区域对应于典型数据，因此预期具有更强的抵御对抗扰动的鲁棒性。

4 算法

本文采用 [20] 中开发的分数模型作为净化方法。这些分数模型通过对一系列噪声水平 $\sigma_1 > \sigma_2 > \cdots > \sigma_L$ 最小化目标函数 (6) 进行训练，其中 $\{\sigma_i\}_{i=1}^L$ 是一个几何序列。我们通过在逐渐降低的噪声水平 L 上依次细化 x ，求解优化问题 (12)。从 σ_1 开始，并以 $x_0 = x_{\text{adv}}$ 初始化，我们按照以下方式在 L 次迭代中更新 x ，直至 σ_L

$$\begin{aligned}
x_0 &= x_{\text{adv}}, \\
x_k^+ &= x_{k-1} + \varrho_{\text{sam}} \frac{\nabla_x \widehat{R}_\theta(x_{k-1};\sigma_k)}{\|\nabla_x \widehat{R}_\theta(x_{k-1};\sigma_k)\|_2}, \\
x_k^{\max} &= \text{Projection}_{[0,1]^d}(x_k^+), \\
x'_k &= x_{k-1} - \eta_k \nabla_x \widehat{R}_\theta(x_k^{\max};\sigma_k), \\
x_k &= \text{Projection}_{[0,1]^d \cap B_2(x_{\text{adv}},\varrho_{\text{pur}})}(x'_k),
\end{aligned} \tag{15}$$

对于 $k = 1,\ldots,L$。其中， η_k 是第 k 次迭代的步长， $B_2(x_{\text{adv}},\varrho_{\text{pur}})$ 是以 x_{adv} 为中心、半径为 ϱ_{pur} 的 ℓ_2-球， R_θ 是式 (13) 中定义的期望重建误差R_θ 的经验近似，

$$\widehat{R}_\theta(x;\sigma_k) = \frac{1}{m}\sum_{i=1}^m \|\xi_{ki} + \sigma_k s_\theta(x + \sigma_k \xi_{ki};\sigma_k)\|_2^2, \tag{16}$$

其中，$\{\xi_{ki}\}_{k=1,\ldots,L;\, i=1,\ldots,m}$ 是 Ξ 的独立同分布样本。

为了更新净化样本，我们使用 Adam [11] 作为优化器。完整算法总结于算法 1。

算法 1 PurSAMERE

要求：对抗样本 x_{adv}，分数模型 s_θ 针对包括 $\{\sigma_i\}_{i=1}^{L}$ 在内的噪声水平进行训练，净-
化半径 ϱ_{pur}，SAM 半径 ϱ_{sam}，蒙特卡洛样本数量 m，Adam优化器
$\mathcal{A}$。学习率最大值与最小值： $\eta_{\max} = 0.1$， $\eta_{\min} = 0.001$。可选：生成器 G 和随机种子
$\{\text{sd}_i\}_{i=1}^{L}$。

$x_0 \leftarrow x_{\text{adv}}$
设置 Adam 优化器 $\mathcal{A}$ ，初始学习率为 $\eta_{\max}$
</style>对于 至 L ，**执行**
使用生成器 G，以种子 sd_k 生成 m独立同分布样本 $\{\xi_{ki}\}_{i=}$ $_1$ 来自 $\mathcal{N}(0, \mathbf{I}_d)$
计算近似梯度 ∇_x $\widehat{R}_\theta(x_{k-1};\sigma_k)$，基于式 (16) 使用反向传播计算

$$x_k^+ \leftarrow x_{k-1} + \varrho_{\text{sam}} \frac{\nabla_x \widehat{R}_\theta(x_{k-1};\ \sigma_k)}{\|\nabla_x \widehat{R}_\theta\ (x_{k-1};\ \sigma_k)\ \|_2}$$

$x_k^{\max} \leftarrow$ 将 x_k^+ 投影到 $[0,1]^d$
计算近似梯度 ∇_x $\widehat{R}_\theta$基于式 (16) 通过反向传播计算 $(x_k^{\max};\sigma_k)$
$x'_k \leftarrow$使用 ∇_x更新 Adam优化器 $\mathcal{A}$ 的状态 $\widehat{R}(x_k^{\max};\sigma_k)$
$x_k \leftarrow$投影 x' $_k$ 投影到 $[0,1]^d \cap B_2(x_{\text{adv}},\ \varrho_{\text{pur}})$
$lr = \eta_{\min} + (\eta_{\max} - \eta_{\min}) \times (k-1)/(L-1)$
将 Adam优化器的学习率更新 $\mathcal{A}$为 lr
结束循环
返回净化后的样本 x_L

在算法 1 中，学习率从最大噪声尺度 σ_1 下的 0.1 线性降低至最小噪声尺度 σ_L下的 0.001。此过程对于确保在细噪声尺度下进行稳定而精确的更新至关重要，同时也能在粗噪声尺度下进行足够大的探索步进。

尽管我们的净化方法在设计上是确定性的，但对期望重建误差（16）进行蒙特卡洛近似会引入少量随机性。可以通过结合以下技术来控制这种随机性：增加蒙特卡洛样本数量 m、降低学习率，以及使用能够抵御噪声梯度的自适应优化器。实验表明，蒙特卡洛近似引入的随机性不会影响我们算法的净化性能。尽管如此，我们仍可选择使用带固定种子的随机数生成器来生成蒙特卡洛近似所需的噪声样本。本文报告的所有使用我们方法进行的实验均启用了这一固定种子选项，因此在给定输入的条件下，净化映射是确定性的。在确定性白盒威胁模型中，假定攻击者知道生成器和随机种子（因此无需使用期望变换）。

用于训练分类器的数据增强 为进一步提升对抗鲁棒性，我们使用净化样本增强分类器的训练数据集。给定一个干净训练样本 x，我们以 x 作为输入，应用算法 1 中所述的净化方法，生成其净化版本 x_{pur} 。具体而言，在净化过程中停用 SAM 选项，以促使分类器适应分数模型 s_θ的原始景观。随后，使用净化样本 $\{(x_{\text{pur}},\ y)\}$增强训练数据集，其中 y 是干净样本 x的标签。分类器在原始训练样本和净化样本上共同进行训练。

用于训练分类器的损失函数如下所示：

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\frac{1}{2}\ell(h(x),y) + \frac{1}{2}\ell(h(x_{\text{pur}}),y)\right], \tag{17}$$

其中， $\mathcal{D}$ 是训练数据集， h 是分类器， ℓ 是交叉熵损失函数。这种数据增强策略鼓励分类器对净化样本具备鲁棒性。

5 实验

我们评估 净化方法对以下攻击的对抗鲁棒性：

- gPGD20：一种使用 20 个 PGD 步骤的灰盒攻击。
- BPDAs-det：一种确定性的白盒自适应攻击，使用 s个步骤和反向传递可微近似（BPDA）；净化和攻击过程使用固定的随机数生成器和种子。

对于 gPGD20 攻击，攻击者可以访问分类器，但无法获知净化过程；对抗样本通过攻击分类器生成。对于确定性的白盒自适应攻击 BPDAs-det，防御是确定性的，攻击者完全了解分类器和净化过程，包括随机数生成器。

在白盒攻击中，我们采用 BPDA 技术来估计损失函数相对于输入的梯度，因为净化映射由于投影/裁剪和迭代优化而无法直接求导。在我们的 BPDA 实现中，我们使用 Athalye等人 [2, 第 4.1 节1]提出的直通式（恒等）反向近似。在前向传播中，我们计算 $x_{\text{pur}} = \text{Pur}(x)$；而在反向传播中，我们将 $\nabla_x \text{Pur}(x)$替换为恒等映射。等价地，

$$\nabla_x\, \ell(h(\text{Pur}(x)),y) \;\approx\; \nabla_x\, \ell(h(x),y)\,\big|_{x=\text{Pur}(x)}.$$

这是一个近似梯度；因此，我们使用足够多的 BPDA 步数（例如， $s = 20$和 $s = 200$）来获得强大的自适应攻击。

使用这些近似梯度，迭代生成对抗样本。可以采用各种基于优化的更新规则，例如快速梯度符号法（FGSM） [7],PGD [14], 或 Carlini & Wagner（C&W） [3] 攻击。本文采用 PGD，因为它比 FGSM 有效得多，且性能可与其他强大的基于迭代优化的攻击相媲美。

对于涉及显著随机性的净化方法，需要使用期望变换技术来正确估计梯度。我们将这种攻击记为 BPDAs+期望变换k，它运行 k 个统计独立的净化过程，以估计每个 sBPDA 步数所需的梯度。这是标准的期望变换估计器：在每个 BPDA 步数使用 k 个蒙特卡洛样本来估计 $\nabla_x\, \mathbb{E}_\omega[\ell(h(\text{Pur}_\omega(x)), y)]$, ；参见 Athalye等人 [2, 第 4 节2]。

CIFAR-10 数据集上的数值结果 我们对所有测试样本执行 gPGD20 和 BPDA20-det 攻击；受计算资源限制，BPDA200-det 和 BPDA200+期望变换20 攻击仅在 1,000 个测试样本的子集上进行。此外，BPDA200-det 攻击针对 ℓ_∞、 ℓ_2 和 ℓ_1 扰动进行评估，而 gPGD20 和

BPDA20-det 攻击仅针对 ℓ_∞ 对抗扰动进行测试。 ℓ_∞、 ℓ_2 和 ℓ_1 攻击的对抗扰动预算分别设为 $8/255$、1 和 12。在ℓ_∞ 扰动预算下，gPGD20 和 BPDA20-det 攻击中每个 PGD 步骤的步长设为 $2/255$ 。对于 BPDA200-det 攻击，步长计算为 $2.5\varrho_{\mathrm{adv}}/200$，对应于 ℓ_∞、 ℓ_2 和 ℓ_1 攻击时分别为 0.00039、0.0125 和 0.15。

我们采用来自 [20] 的已训练基于分数的模型，且不更改其设置。该模型使用 $L = 232$ 个噪声水平进行训练，这些噪声水平从 $\sigma_1 = 50$ 到 $\sigma_L = 0.01$ 按几何序列排列。我们设置 $m = 4$ ，以近似期望重建误差。

我们将净化预算 ϱ_{pur} 在范围 $[2, 3]$ 内进行变化——也就是说，取 ℓ_2 对抗扰动预算的 2 到 3 倍——以研究干净数据与对抗数据上的准确率之间的权衡。尖锐度半径必须小于 ϱ_{pur} ，但与其处于同一数量级，以确保净化样本位于期望重建误差 R_θ 的平坦区域中。因此，我们设置 $\varrho_{\mathrm{sam}} = 1.5$。我们采用 WideResNet-28-10 [27] 架构作为分类器，并使用通过 $\varrho_{\mathrm{pur}} = 3$ 生成的原始训练样本和净化训练样本对其进行训练。

我们的方法在 gPGD20 攻击下的对抗准确率汇总于表 1。我们的方法能够有效地在干净数据和对抗数据上均保持较高的准确率。

表 1：在扰动上界为 $\varrho_{\mathrm{adv}} = 8/255$ 的情况下，我们的方法在 CIFAR-10 上遭受 gPGD20 攻击时的对抗准确率（%）。

	干净	gPGD20
我们的方法		
$\varrho_{\mathrm{pur}} = 2$	**91.18**	**87.88**
$\varrho_{\mathrm{pur}} = 2.5$	88.45	86.73
$\varrho_{\mathrm{pur}} = 3$	85.68	84.17

表 2 和表 3 分别汇总了在 BPDA20-det 和 BPDA20+期望变换10 攻击下，我们的净化方法与几种具有代表性的净化方法的对抗准确率。与现有净化技术相比，所提出的方法始终展现出更强的鲁棒性。特别是在强大的 BPDA20-det 攻击下，该方法的对抗准确率达到 69.08%，显著高于此前报告的 39.16% 的最先进准确率 [13]。此外，表 3 展示了在另一种强攻击配置（BPDA200+期望变换20）下获得的结果。在 ℓ_∞、 ℓ_1 和 ℓ_2 攻击下，我们的方法平均对抗准确率达到 66.23%，优于其他净化方法。

表 2：在扰动上界为 $\varrho_{\text{adv}} = 8/255$ 的 ℓ_∞ BPDA20-det 攻击下，CIFAR-10 上各种净化方法的对抗准确率（%）。

	干净	BPDA20-det
我们的方法		
$\varrho_{\text{pur}} = 2$	91.18	48.17
$\varrho_{\text{pur}} = 2.5$	88.45	63.35
$\varrho_{\text{pur}} = 3$	85.68	69.08
Liu 等人 [13]		
DP_{DDPM}	85.94	16.8
DP_{DDIM}	88.38	4.98
DP_{DDPM}，使用 ADDT	85.64	17.09
DP_{DDIM}，使用 ADDT	88.77	39.16

表 3：在 BPDA200-det 攻击和 BPDA200+期望变换20 攻击下，各种净化方法在 CIFAR-10 上的对抗准确率（%）。后者使用 200 个 BPDA 步数，每步使用 20 个期望变换样本。标记为 * 的方法的数值结果引自 [12]。

	干净	ℓ_∞	ℓ_1	ℓ_2	平均值
BPDA200-det 我们的方法 ($\varrho_{\text{pur}} = 3$)	84.6	**67.9**	**66.9**	**63.9**	**66.23**
BPDA200+期望变换20					
Li等人 [12] *	91.9	47.7	53.5	63.3	53.5
Yoon等人 [26] *	87.93	37.65	36.87	57.81	44.11
Nie 等人 [15] *	**92.5**	42.2	44.3	60.8	49.1

6 结论

本文提出了一种新颖的确定性净化方法，用于增强神经网络分类器的对抗鲁棒性。我们的方法基于对噪声扰动下期望重建误差进行基于SAM的优化。我们证明，PurSAMERE 会将净化样本推向平滑密度 p_{Y_σ} 的局部极大值；在理想化的局部区域内（定理 1），该方法可恢复模型势函数 $\sigma \to 0$ 的一个局部最大值点。我们在 CIFAR-10 数据集上的实验结果表明，与最先进的净化方法相比，我们的方法显著提升了对抗鲁棒性，尤其是在攻击者完全了解净化过程的确定性白盒攻击下。作为一种确定性净化方法，我们的方法在此类攻击下不会出现有效鲁棒性下降，而其他随机净化方法中通常会观察到这一现象。因此，我们的方法能够提供更可靠的对抗准确率。

7 致谢

V.H.、S.K.、H.R. 和 R.T. 感谢德国研究联合会（DFG）通过协同研究中心“稀疏性与奇异结构”（SFB 1481）提供的资助，项目编号为 442047500。V.H. 感谢德国联邦研究、技术与航天部（BMFTR）以及北莱茵-威斯特法伦州经济事务、工业、气候行动与能源部通过 HC-H2 项目提供的资助。

参考文献

[1] Guillaume Alain 和 Yoshua Bengio。正则化自编码器从数据生成分布中学到了什么。《机器学习研究杂志》，15(1):3563–3593，2014年1月。ISSN 1532- 4435。[2] Anish Athalye、Nicholas Carlini 和 David Wagner。混淆梯度带来虚假的安全感：绕过针对对抗样本的防御。在 Jennifer Dy 和 Andreas Krause（编）主编的第*35*届国际机器学习会议论文集中，第80卷，机器学习研究论文集，第274–283页。PMLR，2018年7月10日至15日。URL https://proceedings.mlr.press/v80/athalye18a.html。[3] Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。在*2017*年 *IEEE*安全与隐私研讨会（*SP*）中，第39–57页，2017年。doi: 10.1109/SP. 2017.49。[4] Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。通过锐度感知最小化高效提升泛化能力。在国际学习表征会议中，2021年。URL https://openreview.net/forum?id=6Tm1mposlrM。[5] Simon Foucart 和 Holger Rauhut。压缩感知数学导论。Birkh¨auser 纽约，纽约州，2013年。doi: 10.1007/978-0-8176-4948-7。[6] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。《*Commun. ACM*》，63(11):139–144，2020年10月。ISSN 0001-0782。doi: 10.1145/3422622。URL https: //doi.org/10.1145/3422622。[7] Ian J. Goodfellow、J onathon Shlens 和 Christian Szegedy。解释并利用对抗样本。在 Yoshua Bengio 和 Yann LeCun（编）主编的第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集中，2015年。URL http://arxiv.org/abs/1412.6572。[8] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。在第*34*届神经信息处理系统大会论文集中，NIPS ’20，美国纽约州雷德胡克，2020年。Curran Associates Inc. ISBN 9781713829546。[9] Aapo Hyv¨arinen。通过分数匹配估计非归一化统计模型。《机器学习研究杂志》，6(24):695–709，2005年。URL http://jmlr.org/papers/v6/ hyvarinen05a.html。

[10] Aapo Hyvarinen。连续值变量的分数匹配、对比散度与伪似然之间的联系。*IEEE* 神经网络汇刊，18(5): 1529–1531，2007年。[11] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。在 Yoshua Bengio 和 Yann LeCun（编）主编的第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集中，2015年。URL http://arxiv.org/abs/1412.6980。[12]Xiao Li、Wenxuan Sun、Huanran Chen、Qiongxiu Li、Yingzhe He、Jie Shi 和 Xiaolin Hu。ADBM：用于可靠对抗净化的对抗扩散桥模型。在第十三届国际学习表征会议中，2025年。URL https: //openreview.net/forum?id=g0rnZeBguq。[13] Yiming Liu、Kezhao Liu、Yao Xiao、ZiYi Dong、Xiaogang Xu、Pengxu Wei 和 Liang Lin。迈向理解基于扩散的净化的鲁棒性：一种随机视角。在第十三届国际学习表征会议中，2025年。URL https: //openreview.net/forum?id=shqjOIK3SA。[14] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。在国际学习表征会议中，2018年。URL https://openreview.net/forum?id= rJzIBfZAb。[15]Weili Nie、Brandon Guo、Yujia Huang、Chaowei Xiao、Arash Vahdat 和 Animashree Anandkumar。用于对抗净化的扩散模型。在 Kamalika Chaudhuri、Ste- fanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编）主编的第*39*届国际机器学习会议论文集中，第162卷，机器学习研究论文集，第16805–16827页。PMLR，2022年7月17日至23日。URL https://proceedings.mlr.press/v162/nie22a.html。[16] Tim Salimans、Andrej Karpathy、Xi Chen 和 Diederik P. Kingma。PixelCNN++：通过离散化逻辑斯蒂混合似然及其他改进提升 pixelCNN。在国际学习表征会议中，2017年。URL https://openreview.net/forum? id=BJrFC6ceg。[17] Pouya Samangouei、Maya Kabkab 和 Rama Chellappa。Defense-GAN：利用生成模型保护分类器免受对抗攻击。在国际学习表征会议中，2018年。URL https://openreview.net/forum?id=BkJ3ibb0-。[18] Changhao Shi、Chester Holtz 和 Gal Mishne。基于自监督学习的在线对抗净化。在国际学习表征会议中，2021年。URL https://openreview.net/forum?id=_i3ASPp12WS。[19] Yang Song 和 Stefano Ermon。通过估计数据分布的梯度进行生成建模。在 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett（编）主编的神经信息处理系统进展中，第32卷。Curran Associates, Inc.，2019年。URL https://proceedings.neurips.cc/paper_ files/paper/2019/ file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf。

[20]Yang Song和Stefano Ermon。训练基于分数的生成模型的改进技术。载于H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan和H. Lin编，神经信息处理系统大会进展，第33卷，第12438–12448页。Curran Associates, Inc.，2020年。网址 https://proceedings.neurips.cc/paper_files/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.[21] Yang Song、Taesup Kim、Sebastian Nowozin、Stefano Ermon和Nate Kushman。PixelDefend：利用生成模型理解并防御对抗样本。载于国际学习表征会议，2018年。网址 https://openreview. net/forum?id=rJUYGxbCW.[22] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian J. Goodfellow和Rob Fergus。神经网络的有趣性质。载于Yoshua Bengio和Yann LeCun编，第二届国际学习表征会议，*ICLR 2014*，加拿大班夫，*AB*，*2014*年*4*月*14-16*日，会议轨道论文集，2014年。网址 http: //arxiv.org/abs/1312.6199.[23] A¨aron Van Den Oord、Nal Kalchbrenner和Koray Kavukcuoglu。像素循环神经网络。载于第*33*届国际机器学习国际会议论文集 - 第*48*卷，ICML'16，第1747–1756页。JMLR.org，2016年。[24] Pascal Vincent。分数匹配与去噪自编码器之间的联系。神经计算，23(7):1661–1674，2011年7月。ISSN 0899-7667。doi: 10.1162/NECO a 00142。网址 https://doi.org/10.1162/NECO_a_00142。[25] Pascal Vincent、Hugo Larochelle、Yoshua Bengio和Pierre-Antoine Manzagol。利用去噪自编码器提取并组合鲁棒特征。载于第*25*届国际机器学习会议论文集，ICML '08，第1096–1103页，美国纽约州纽约，2008年。计算机协会。ISBN 9781605582054。doi: ¯ ¯ 10.1145/1390156.1390294。网址 https://doi.org/10.1145/1390156.1390294。[26] Jongmin Yoon、Sung Ju Hwang和Juho Lee。使用基于分数的生成模型进行对抗净化。载于Marina Meila和Tong Zhang编，第*38*届国际机器学习会议论文集，机器学习研究论文集第139卷，第12062–12072页。PMLR，2021年。网址 https://proceedings.mlr.press/v139/yoon21a. html。[27] Sergey Zagoruyko和Nikos Komodakis。宽残差网络。载于Richard C. Wilson、Edwin R. Hancock和William A. P. Smith编，*2016*年英国机器视觉会议论文集，*BMVC 2016*，英国约克，*2016*年*9*月*19-22*日。BMVA Press，2016年。网址 https://bmva-archive.org.uk/bmvc/2016/papers/paper087/index.html。

理论分析

A.1 去噪得分匹配目标函数

引理 1。　对于所有 $s_\theta \in L_2(\mathbb{R}^d, p_{Y_\sigma})$，去噪得分匹配目标函数 (6) 满足

$$\mathbb{E}_{X,\ \Xi}[\|\Xi + \sigma s_\theta(X + \sigma\Xi;\ \sigma)\|_2^2] = \sigma^2 \mathbb{E}_{X,\ \Xi}[\|s(X + \sigma\Xi;\ \sigma) - s_\theta(X + \sigma\Xi;\ \sigma)\|_2^2] + \tag{18}$$

常数项，其中常数项与 s_θ 无关。

证明。根据 (3)，对于所有函数 $g \in L_2(\mathbb{R}^d,\ p_{Y_\sigma})$，

$$\mathbb{E}_{X,\Xi}\left[(X - \mathbb{E}[X \mid Y_\sigma = X + \sigma\Xi])^\top g(X + \sigma\Xi)\right] = 0.$$

因此，我们有

$$\begin{aligned}\mathbb{E}_{X,\Xi}\left[\|\Xi + \sigma s_\theta(X + \sigma\Xi;\sigma)\|_2^2\right] &= \mathbb{E}_{X,\Xi}\left[\|\Xi + \sigma s(X + \sigma\Xi;\sigma)\|_2^2\right] \\ &+ \sigma^2 \mathbb{E}_{X,\Xi}\left[\|s(X + \sigma\Xi;\sigma) - s_\theta(X + \sigma\Xi;\sigma)\|_2^2\right] \\ &+ 2\sigma\mathbb{E}_{X,\Xi}\left[(\Xi + \sigma s(X + \sigma\Xi;\sigma))^\top (s(X + \sigma\Xi;\sigma) - s_\theta(X + \sigma\Xi;\sigma))\right]\end{aligned}$$

其中，右侧第一项是与 s_θ 无关的常数，右侧第三项可表示为

$$2\sigma\mathbb{E}_{X,\Xi}\left[(X - \mathbb{E}[X \mid Y_\sigma = X + \sigma\Xi])^\top (s(X + \sigma\Xi;\sigma) - s_\theta(X + \sigma\Xi;\sigma))\right],$$

等于零。 □

A.2 命题 1 的证明

命题 *1* 的证明。一维映射 $\sigma \mapsto s(x+\sigma\xi, \sigma)$ 在 $\sigma = 0$ 处的泰勒展开为

$$s(x + \sigma\xi, \sigma) = s(x, 0) + \sigma\nabla_x s(x, 0)\xi + \sigma\partial_\sigma s(x, 0) + \frac{1}{2}\sigma^2\,\phi(x, \sigma_t, \xi),$$

其中 $\sigma_t \in (0, \sigma)$ 取决于 x 和 ξ，

$$\phi(x, \sigma, \xi) = \xi^\top \nabla_x^2 s(x + \sigma\xi, \sigma)\xi + 2\partial_\sigma \nabla_x s(x + \sigma\xi, \sigma)\xi + \partial_{\sigma\sigma} s(x + \sigma\xi, \sigma).$$

Let

$$A(x, \xi) := \xi + \sigma s(x, 0) + \sigma^2 \nabla_x s(x, 0)\xi + \sigma^2 \partial_\sigma s(x, 0).$$

由于 $\Xi \sim \mathcal{N}(0, \mathbf{I}_d)$，我们有

$$\begin{aligned}\mathbb{E}_\Xi\left[\|A(x, \Xi)\|^2\right] &= \mathbb{E}_\Xi\left[\|\Xi + \sigma\, s(x, 0) + \sigma^2 \nabla_x s(x, 0)\Xi + \sigma^2 \partial_\sigma s(x, 0)\|_2^2\right] \\ &= d + \sigma^2 \|s(x, 0)\|_2^2 + 2\sigma^2 \operatorname{tr}(\nabla_x s(x, 0)) + \mathcal{O}(\sigma^3).\end{aligned}$$

此外，存在常数 $C > 0$ 和 $m \geq 0$，使得对于所有 $x \in K$、所有 $\sigma \in [0, \sigma_{\max}]$ 以及所有 $\xi \in \mathbb{R}^d$，

$$\max_{1 \leq i \leq d} |\phi_i(x, \sigma, \xi)| \ \leq\ C\big(1 + \|\xi\|_2^m\big),$$

因此，

$$\mathbb{E}_\Xi\left[\|A(x, \Xi) + \frac{1}{2}\sigma^3 \phi(x, \sigma_t, \Xi)\|_2^2\right] = \mathbb{E}_\Xi\left[\|A(x, \Xi)\|^2\right] + \mathcal{O}(\sigma^3).$$

最后，我们得到：

$$\mathbb{E}_\Xi\left[\|\Xi + \sigma\, s(x + \sigma\Xi, \sigma)\|_2^2\right] = d + \sigma^2 \|s(x, 0)\|_2^2 + 2\sigma^2 \operatorname{tr}(\nabla_x s(x, 0)) + \mathcal{O}(\sigma^3).$$

□

A.3 定理：密度最大值点的局部恢复

我们以正式形式重述定理 1。

定理 2（密度最大值点的局部恢复（形式化表述））。设 $s_\theta\colon \mathbb{R}^d \times (0, \sigma_{max}] \to \mathbb{R}^d$ 是一个分数模型，用于逼近得分函数 $s(x;\sigma) = \nabla_x \log p_{Y_\sigma}(x)$ 对于 $\sigma \in (0, \sigma_{max}]$。

假设 ***1***：函数 $x \mapsto s_\theta(x;\sigma)$ 是分段仿射的，且在 $\mathbb{R}^d$ 上全局满足 L-利普希茨条件。更确切地说，存在与 σ 无关的 $\mathbb{R}^d$ 的闭集 $\{D_i\}_{i=1}^N$，使得

- *(i)* $\cup_{i=1}^N D_i = \mathbb{R}^d$,
- （*ii*）$\mathrm{int}(D_i) \cap \mathrm{int}(D_j) = \emptyset$ 对于 $i \neq j$,
- *(iii)* 对于每个 i，D_i 的内部具有正的勒贝格测度，且边界 ∂D_i 的勒贝格测度为零。

对于每个 i 和 $\sigma \in (0,\ \sigma_{\max}]$，$s_\theta(\cdot;\ \sigma)$ 在 $\mathrm{int}(D_i)$ 上是仿射的，并且连续延拓至 D_i。此外，存在一个常数 $L > 0$，使得

$$\|s_\theta(x;\sigma) - s_\theta(y;\sigma)\|_2 \le L\|x-y\|_2, \quad \forall x, y \in \mathbb{R}^d,\ \sigma \in (0, \sigma_{\max}].$$

假设 ***2***：存在一个函数 $f_\theta\colon \mathbb{R}^d \times (0,\ \sigma_{max}] \to \mathbb{R}$，使得 $s_\theta(x;\ \sigma) = \nabla_x f_\theta(x;\ \sigma)$。此外，对于每个 $\sigma \in (0,\ \sigma_{max}]$，函数 $x \mapsto \exp(f_\theta(x;\ \sigma))$ 在 $\mathbb{R}^d$上定义了一个密度函数。也就是说，$\int_{\mathbb{R}^d} \exp(f_\theta(x;\ \sigma))\mathrm{d}x = 1$。

假设 ***3***：存在一个紧集 $D \in \{D_1, \cdots, D_N\}$ 以及常数 $\mu > 0$ 和 $0 < \varrho < \sigma_{max}$，使得 $f_\theta(x,\ \sigma)$ 对所有 $\sigma \in (0,\ \varrho]$ 均在 D 上是 μ-强凹的。此外，D 的 $\sqrt{d}\,\varrho$ 偏移集非空，即

$$D_{-\sqrt{d}\,\varrho} = \{x \in D : \inf_{y \notin D} \|x-y\|_2^2 \ge d\varrho^2\} \neq \emptyset,$$

对每个 $\sigma \le \varrho$，且 $f_\theta(\cdot,\sigma)$ 的（唯一）最大值点 $x_\sigma^\star$ 位于 $D_{-\sqrt{d}\,\varrho}$ 中。

在假设 *1*、*2* 和 *3* 下，对于所有 $0 < \sigma \le \varrho$，一个极小化点 $x'_\sigma \in D_{-\sqrt{d}\,\varrho}$ 满足

$$R_\theta(x;\sigma) := \mathbb{E}_\Xi\left[\|\Xi + \sigma s_\theta(x + \sigma\Xi, \sigma)\|_2^2\right], \quad \Xi \sim \mathcal{N}(0, \mathbf{I}_d)$$

满足

$$\|x'_\sigma - x^\star_\sigma\|_2 \le \frac{\sqrt{C}}{\sigma\mu} \exp\left(-\frac{d}{8}\left(\frac{\varrho}{\sigma} - 1\right)^2\right),$$

其中某个常数 $C > 0$ 与 σ 无关。

假设 1 要求满足两个条件：首先，分数模型是分段线性的；其次，划分 $\{D_i\}_{i=1}^N$ 与 σ 相互独立。第一个条件由使用 ReLU 激活函数且不含层归一化的常见神经网络架构满足。若将分数模型参数化为 $s_\theta(x;\sigma) = \gamma_\theta(x)/\sigma$，其中 γ_θ 是如 [20] 中提出的神经网络，则可以满足第二个条件。假设 2 通常能够满足，只要分数模型训练得足够充分，因为它近似于某个密度函数的得分函数。假设 3 要求 f_θ 存在一个强凹区域，且局部最大值点严格位于该区域内部。这个假设是合理的，因为数据密度函数 p_X 通常有多个局部极大值，而当 σ 较小时，平滑密度 p_{Y_σ} 能很好地近似 p_X。定理 2 的证明依赖于以下两个引理。

引理 2. 设 D 是 $\mathbb{R}^d$ 中的一个闭集，$\Xi \sim \mathcal{N}(0, \mathbf{I}_d)$，且 $g\colon \mathbb{R}^d \to \mathbb{R}$ 是$L^2(p_\Xi)$ 中的一个函数，也就是说，存在一个常数 $C > 0$ 满足

$$\mathbb{E}_\Xi \left[|g(\Xi)|^2\right] \le C^2.$$

对于所有使 D 的 $\sqrt{d}\varrho$-偏移非空的 $\varrho > 0$，即

$$D_{-\sqrt{d}\varrho} := \{x \in D : \inf_{y \notin D} \|x - y\|_2 \ge \sqrt{d}\varrho\} \neq \emptyset,$$

且对于任意 $0 < \sigma < \varrho$，以下不等式成立：

$$\left|\mathbb{E}\left[\mathbb{1}_{D^c}(x + \sigma\Xi)g(\Xi)\right]\right| \le C \exp\left(-\frac{d}{4}\left(\frac{\varrho}{\sigma} - 1\right)^2\right), \quad \forall x \in D_{-\sqrt{d}\varrho},$$

其中 $\mathbb{1}_{D^c}$ 表示集合 $D^c = \mathbb{R}^d \setminus D$ 的指示函数。

引理 3. 设 $f : D \to \mathbb{R}$，其中 D 是 $\mathbb{R}^d$ 中的一个紧集。假设

$$f(x) = (x - x^\star)^\top H(x - x^\star) + g(x), \quad \forall x \in D,$$

对于某个 $x^\star \in D$，其中 H 是对称正定的，且对于某个 $\mu > 0$，有 $H \succeq \mu\mathbf{I}_d$；并且 $g\colon D \to \mathbb{R}$ 连续且满足：对于所有 $x \in D$，存在某个 $\varepsilon > 0$。那么，f 在 D 中的任意极小化点 $x_{\min}$ 都位于 $x^\star$ 的邻域内，即 $\|x_{\min} - x^\star\|_2 \le \sqrt{2\varepsilon/\mu}$。

A.3.1 证明

证明　　定理 2. 由于 s_θ 关于 D 是线性的，其中 D 是满足假设 3 的定义域，我们有　　ave

$$s_\theta(x;\sigma) = H_\theta(\sigma)(x - x^\star_\sigma) + s_\theta(x^\star_\sigma, \sigma), \quad \text{for } x \in D$$

其中

$$H_\theta(\sigma) = \nabla_x s_\theta(x;\sigma), \quad x \in D,$$

是 $f_\theta(x,\sigma)$ 在 D 中的常数海森矩阵。假设 3 导致 $H_\theta(\sigma) \preceq -\mu I_d$ 且 $s_\theta(x^\star_\sigma, \sigma) = 0$。因此，对于 $x \in D$，有 $s_\theta(x;\sigma) = H_\theta(\sigma)(x - x^\star_\sigma)$。

对于 $x \in D_{-\sqrt{d}\varrho}$，期望重建误差 (13) 可表示为

$$\begin{aligned}
R_\theta(x;\sigma) &= \mathbb{E}_\Xi\left[\|\Xi + \sigma s_\theta(x + \sigma\Xi)\|_2^2\right] \\
&= \mathbb{E}_\Xi\left[\|\Xi + \sigma H_\theta(\sigma)(x + \sigma\Xi - x^\star_\sigma)\|_2^2\right] \\
&\quad - \mathbb{E}_\Xi\left[\mathbb{1}_{D^c}(x + \sigma\Xi)\|\Xi + \sigma H_\theta(\sigma)(x + \sigma\Xi - x^\star_\sigma)\|_2^2\right] \\
&\quad + \mathbb{E}_\Xi\left[\mathbb{1}_{D^c}(x + \sigma\Xi)\|\Xi + \sigma s_\theta(x + \sigma\Xi, \sigma)\|_2^2\right].
\end{aligned}$$

Let

$$\begin{aligned}
g_1(\xi) &:= \left\|\xi + \sigma H_\theta(\sigma)(x + \sigma\xi - x^\star_\sigma)\right\|_2^2, \\
g_2(\xi) &:= \left\|\xi + \sigma s_\theta(x + \sigma\xi;\sigma)\right\|_2^2.
\end{aligned}$$

我们的目标是用一个与 σ 无关的常数来界定 $\mathbb{E}_{\Xi}\left[|g_2(\Xi)-g_1(\Xi)|^2\right]$。由于对于所有 σ，$\|s_\theta(x;\sigma)-s_\theta(y;\sigma)\|_2 \le L\|x-y\|_2$, $\|H_\theta(\sigma)\|_2 \le L$。此外，存在一个与 σ 无关的常数 M_D，使得当 D 是有界集时，$\|x-x^\star_\sigma\| \le M_D$。因此，对于 $\sigma \le \varrho$

$$\begin{aligned} g_1(\xi) &\le 2\left\|(I+\sigma^2 H_\theta(\sigma))\xi\right\|_2^2 + 2\left\|\sigma H_\theta(\sigma)(x-x^\star_\sigma)\right\|_2^2 \\ &\le 2\|I+\sigma^2 H_\theta(\sigma)\|_2^2\,\|\xi\|_2^2 + 2\|\sigma H_\theta(\sigma)\|_2^2\|x-x^\star_\sigma\|_2^2, \\ &\le 2(1+L\varrho^2)^2\|\xi\|_2^2 + 2(L\varrho)^2 M_D^2, \end{aligned}$$

and

$$\begin{aligned} g_2(\xi) &\le 2\|\xi\|_2^2 + 2\sigma^2\|s_\theta(x+\sigma\xi;\sigma)\|_2^2 \\ &\le 2\|\xi\|_2^2 + 2\sigma^2\left(\|s_\theta(x^\star_\sigma;\sigma)\|_2 + L\|x+\sigma\xi-x^\star_\sigma\|_2\right)^2 \\ &\le 2\|\xi\|_2^2 + 2\sigma^2\left(L\|x-x^\star_\sigma\|_2 + L\,\sigma\|\xi\|_2\right)^2 \\ &\le 2(1+2L^2\sigma^4)\|\xi\|_2^2 + 4\sigma^2L^2\|x-x^\star_\sigma\|_2^2, \\ &\le 2(1+2L^2\varrho^4)\|\xi\|_2^2 + 4\varrho^2L^2M_D^2. \end{aligned}$$

由于 $\Xi \sim \mathcal{N}(0,\ \mathbf{I}_d)$，$g_1(\Xi)$ 和 $g_2(\Xi)$ 的二阶矩有界。因此，存在一个与 σ 无关的常数 $C>0$，使得 $\mathbb{E}_{\Xi}\left[|g_2(\Xi)-g_1(\Xi)|^2\right] \le C^2$。记 $A = \mathbb{E}_{\Xi}\left[\mathbb{1}_{D^c}(x+\sigma\Xi)\,(g_2(\Xi)-g_1(\Xi))\right]$。利用引理 2，我们可以对 A 作如下界定:

$$|A| \le C\exp\left(-\frac{d}{4}\left(\frac{\varrho}{\sigma}-1\right)^2\right).$$

因此,

$$R_\theta(x;\sigma) \le \mathbb{E}_{\Xi}\left[\|\Xi+\sigma H_\theta(\sigma)(x+\sigma\Xi-x^\star_\sigma)\|_2^2\right] + C\exp\left(-\frac{d}{4}\left(\frac{\varrho}{\sigma}-1\right)^2\right).$$

右侧第一项是 $R_\theta(x;\sigma)$ 的主要贡献，可表示为

$$\begin{aligned} \mathbb{E}_{\Xi}\left[\|\Xi+\sigma H_\theta(\sigma)(x+\sigma\Xi-x^\star_\sigma)\|_2^2\right] =& d + \sigma^2\|H_\theta(\sigma)(x-x^\star_\sigma)\|_2^2 \\ &+ 2\sigma^2\operatorname{tr}(H_\theta(\sigma)) + \sigma^4\operatorname{tr}(H_\theta(\sigma)H_\theta(\sigma)^\top), \end{aligned}$$

这是关于 x 的二次函数，其唯一极小化点为 $x^\star_\sigma$。由于 $H_\theta(\sigma) \preceq -\mu I_d$，我们有 $\sigma^2 H_\theta(\sigma)^\top H_\theta(\sigma) \succeq \sigma^2\mu^2\mathbf{I}_d$。利用引理 3，我们得出

$$\|x'_\sigma - x^\star_\sigma\|_2 \le \frac{\sqrt{2C}}{\sigma\mu}\exp\left(-\frac{d}{8}\left(\frac{\varrho}{\sigma}-1\right)^2\right),$$

由此完成证明。 □

引理 *2* 的证明。根据利普希茨函数的高斯测度集中性质，参见例如 [5, 定理 8.40], 对所有 $t>0$,

$$\mathbb{P}(\|\Xi\|_2 \ge \mathbb{E}\left[\|\Xi\|_2\right] + t) \le \exp\left(-\frac{t^2}{2}\right).$$

詹森不等式意味着 $\mathbb{E}\left[\|\Xi\|_2\right] \le \left(\mathbb{E}\left[\|\Xi\|_2^2\right]\right)^{1/2} = \sqrt{d}$，因此，对于所有 $t>0$,

$$\mathbb{P}(\|\Xi\|_2 \ge \sqrt{d} + t) \le \exp\left(-\frac{t^2}{2}\right),$$

另请参见 [5, 公式 (8.89)]。令 $\sqrt{d}+t=\sqrt{d}\,\frac{\varrho}{\sigma}$，即 $t=\sqrt{d}\left(\frac{\varrho}{\sigma}-1\right)$，可得

$$\mathbb{P}(\|\Xi\|_2 \geq \sqrt{d}\frac{\varrho}{\sigma}) \leq \exp\left(-\frac{d}{2}\left(\frac{\varrho}{\sigma}-1\right)^2\right).$$

由于 $x \in D_{-\sqrt{d}\,\varrho}$，开球 $B(x,\ \sqrt{d}\,\varrho)$ 包含于 D。因此，若 $x+\sigma\Xi \notin D$，则必然有

$$\|x+\sigma\Xi-x\|_2 \geq \sqrt{d}\,\varrho,$$

这意味着

$$\|\Xi\|_2 \geq \sqrt{d}\,\frac{\varrho}{\sigma}.$$

因此，

$$\mathbb{P}(x+\sigma\Xi \notin D) \leq \mathbb{P}(\|\Xi\|_2 \geq \sqrt{d}\frac{\varrho}{\sigma}) \leq \exp\left(-\frac{d}{2}\left(\frac{\varrho}{\sigma}-1\right)^2\right).$$

最后，我们有

$$\begin{aligned}\left|\mathbb{E}_\Xi\left[\mathbb{1}_{D^c}(x+\sigma\Xi)g(\Xi)\right]\right| &\leq \mathbb{E}_\Xi\left[\mathbb{1}_{D^c}(x+\sigma\Xi)\right]^{1/2}\mathbb{E}_\Xi\left[g^2(\Xi)\right]^{1/2}\\ &\leq C\mathbb{E}_\Xi\left[\mathbb{1}_{D^c}(x+\sigma\Xi)\right]^{1/2} = C\mathbb{P}(x+\sigma\Xi \notin D)^{1/2}\\ &\leq C\exp\left(-\frac{d}{4}\left(\frac{\varrho}{\sigma}-1\right)^2\right),\end{aligned}$$

由此完成证明。 □

引理 3 的证明。由于 $H \succeq \mu\mathbf{I}_d$，我们有

$$f(x) \geq \mu\|x-x^\star\|_2^2 + g(x), \quad \forall x \in D.$$

令 $r=\sqrt{2\varepsilon/\mu}$。对于所有满足 $\|x-x^\star\|_2 > r$ 的 x，我们有

$$f(x) > \mu r^2 - \varepsilon = \varepsilon$$

另一方面，我们有 $f(x^\star) = g(x^\star) \leq \varepsilon$。因此，$D$ 中 f 的极小化点 $x_{\min}$ 必须满足 $\|x_{\min}-x^\star\|_2 \leq r = \sqrt{2\varepsilon/\mu}$。 □