arXiv:2302.02936v2 [cs.LG] 5 Oct 2023

重新审视隐私保护 GAN*

AlexBie† Gautam Kamath‡ Guojun Zhang
yabie@uwaterloo.ca g@csail.mit.edu guojun.zhang@huawei.com滑铁卢大学 滑铁卢大学 华为 诺亚方舟实验室

摘要

我们表明，对差分隐私 GAN 进行训练的经典方法——使用差分隐私随机梯度下降（差分隐私随机梯度下降）更新判别器——经过训练方式上的调整后，可以显著改善结果。具体而言，我们认为，现有的该方法实现忽视了添加噪声 *only to discriminator updates* 会抑制判别器训练这一点，从而打破生成器与判别器之间的平衡，而这种平衡是成功进行生成对抗网络训练所必需的。我们表明，一个简单的修正方法——在生成器步骤之间增加判别器步数——能够恢复生成器与判别器之间的平衡，并改善结果。

此外，为了恢复平衡，我们尝试了其他调整方法——即采用大批量和自适应判别器更新频率——以改善判别器训练，并进一步提升生成质量。我们的结果表明，在标准图像合成基准上，差分隐私随机梯度下降优于所有替代的 GAN 隐私保护方案。代码：https://github.com/alexbie98/dpgan-revisit。

1 引言

差分隐私（DP）（Dwork 等人，2006年b）已成为一种极具吸引力的方法，可用于在敏感数据上训练机器学习模型。然而，引入差分隐私需要改变训练流程。值得注意的是，它使建模者无法直接使用敏感数据，从而令调试和探索更加复杂。此外，一旦耗尽分配的隐私预算，建模者便不能再与敏感数据交互。缓解这些问题的一种方法是生成差分隐私合成数据，这种数据可以直接用于现有的机器学习流水线，无需再担心隐私问题。

为了生成高维、复杂的数据（例如图像），已有一系列研究探讨如何对生成对抗网络（GAN）进行隐私保护（Goodfellow 等人，2014年），以生成差分隐私合成数据。早期研究提出，使用差分隐私随机梯度下降（DPSGD）（Abadi 等人，2016年）替代随机梯度下降来更新生成对抗网络判别器，这种方法称为 *DPGAN*（Xie 等人，2018年；Beaulieu-Jones 等人，2019年；Torkzadehmahani 等人，2019年）。

然而，后续研究（Jordon 等人，2019年；Long 等人，2021年；Chen 等人，2020年；Wang 等人，2021年）不再采用这种方法：他们为生成对抗网络提出了其他隐私保护方案，并报告称其效果显著优于 DPGAN 基准。其他差分隐私合成数据生成方法则摒弃了基于 GAN 的架构，在大多数情况下提升了效用（表 2）。这引出了一个问题：生成对抗网络是否适用于差分隐私训练，还是说差分隐私数据生成需要专门设计的架构?

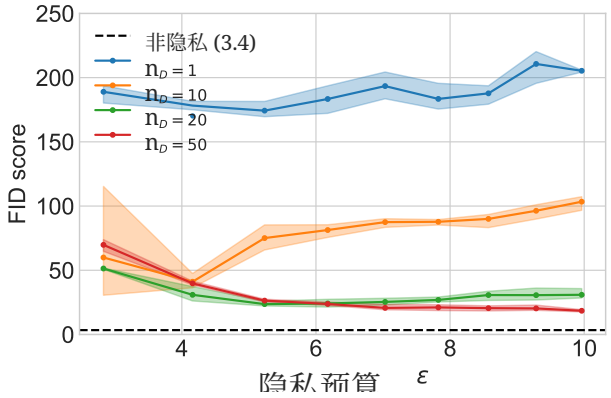
我们的贡献。 我们证明，差分隐私生成对抗网络的效用远超以往研究所展示的水平，在差分隐私图像合成方面，其表现可与几乎所有其他方法相媲美，甚至更胜一筹。[1]

因此，我们得出结论：先前发现的差分隐私生成对抗网络的缺陷，不应归因于框架本身的固有局限，而应归因于训练问题。具体而言，我们认为，

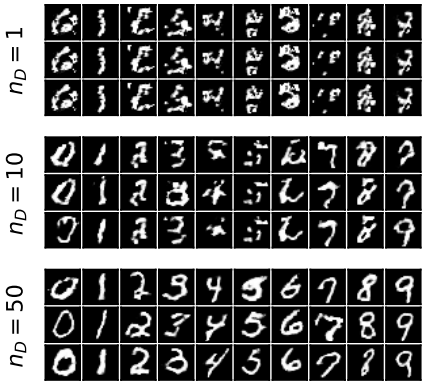
*作者 GK 和 GZ 按字母顺序排列。

†部分工作是在华为实习期间完成的。此外，本研究还获得了 NSERC 探索基金、David R. Cheriton 研究生奖学金和安大略省研究生奖学金的支持。

‡由 NSERC 探索基金、谷歌的无附加条件赠款以及滑铁卢大学的初创研究基金资助。[1]一个值得注意的例外是扩散模型，详见第2节。



(a) 一次训练过程中 MNIST 的 FID



(b) 在 (10，10^{-5})-差分隐私下对应的图像

图 1： 在 (10，10^{-5})-差分隐私下 DPGAN 在 MNIST 上的合成结果。(a) 我们使用 3 个随机种子运行实验，并绘制各次运行中 FID 的均值、最小值和最大值。我们发现，增加 $n_{\mathcal{D}}$（即生成器步骤之间执行的判别器步数）可显著改善图像合成效果。增加 $n_{\mathcal{D}} = 1 \to n_{\mathcal{D}} = 50$ 可将 FID 从 $205.3 \pm 0.9 \to 18.5 \pm 0.9$ 提高。(b) 对应的合成图像（每张图像均使用相同的隐私预算进行训练）。我们观察到，较大的 $n_{\mathcal{D}}$ 可改善视觉质量，而较小的 $n_{\mathcal{D}}$ 会导致模式崩溃。

隐私 ε	方法	MNIST FID	MNIST 准确率（%）	FashionMNIST FID	FashionMNIST 准确率（%）	CelebA-Gender FID	CelebA-Gender 准确率（%）
$\varepsilon = \infty$	真实数据	1.0	99.2	1.5	92.5	1.1	96.6
	GAN	3.4 ± 0.1	97.0 ± 0.1	16.5 ± 1.7	79.5 ± 0.8	30.0 ± 1.6	92.0 ± 0.4
$\varepsilon = 10$	最佳隐私生成对抗网络	61.34	80.92	131.34	70.61	-	70.72
	DPGAN	179.16	80.11	243.80	60.98	-	54.09
$\varepsilon = 9.32$	我们的 DPGAN	12.8 ± 0.3	95.1 ± 0.1	62.3 ± 8.7	74.7 ± 0.4	170.8 ± 20.3	82.4 ± 4.4

表1： 与先前关于私有生成对抗网络的研究结果相比，我们的结果汇总如下。对于我们的结果，我们运行了3个随机种子，并报告均值 $\pm$ 标准差。 *Acc.(%)* 指使用生成数据训练的卷积神经网络模型的下游分类准确率。中间两行综合了文献中报告的 DPGAN 和其他生成对抗网络隐私保护方案的最佳结果（*"Best Private GAN"*）；对应关系见表 2 和 3 。这里，我们使用 Gopi 等人 (2021年) 的隐私核算来计算我们的结果。我们发现，与所有先前基于生成对抗网络的差分隐私图像合成方法相比，我们的结果有显著提升。

在差分隐私生成对抗网络中采用非对称噪声添加（仅对判别器更新添加噪声）会抑制判别器训练，却不影响生成器训练，从而破坏成功进行 GAN 训练所必需的平衡。事实上，Goodfellow 等人（2014年）的开创性研究指出，GAN 训练面临着"使判别器与生成器同步"的挑战，并提出"$\mathcal{G}$ 在更新 $\mathcal{D}$之前，不能对其进行过多训练，否则会导致'Helvetica 情景' [模式崩溃]"。以往文献中的 DPGAN实现，在移植非隐私生成对抗网络训练方案时，并未考虑这一点。

我们提出，在生成器更新之间增加判别器步数，可以解决噪声引入的不平衡问题。采用这一改动后，差分隐私生成对抗网络的效果显著提升（见图 1 和表 1）。此外，我们还表明，这一关于 DPGAN 训练的视角（"恢复受到差分隐私噪声削弱的判别器的平衡"）可用于改进训练。我们还对判别器训练作出其他调整——采用更大的批量大小和自适应判别器更新频率——以改进判别器训练，并进一步提升上述结果。总之，我们作出以下贡献：

- 我们发现，在生成器步骤之间增加判别器步数，能显著提升原始DPGAN的表现。与以往文献中的结果相反，DPGAN的表现优于其他GAN隐私保护方案。

- 我们给出了实证结果，以探究为什么更频繁地进行判别器步数更新会有所帮助。我们提出了一种基于非对称噪声添加 的解释，说明原始DPGAN表现不佳的原因，以及增加判别器步数为何能带来改善。

- 我们将这一解释作为设计更优隐私GAN训练方案的原则——采用更大的批量大小和自适应判别器更新频率——并且确实能够在上述结果的基础上取得改进。

2 相关工作

隐私GAN。采用经差分隐私随机梯度下降训练的判别器的基准DPGAN由Xie等人（2018）提出，随后由Torkzadehmahani等人（2019）和Beaulieu-Jones等人（2019）进行了研究。尽管该方法受到了广泛关注（截至撰写本文时已有≈400 篇引用），我们未能找到探索我们所进行的修改，或发现改进DPGAN训练的类似原则的研究。我们注意到，每个生成器步骤所执行的判别器步数 $n_{\mathcal{D}}$，在Goodfellow等人（2014）的开创性研究所概述的框架中，以及WGAN（Arjovsky 等人，2017）等后续工作中，都作为一个超参数出现。Xie等人（2018）将WGAN隐私化，采用了其 $n_{\mathcal{D}} = 5$的不平衡步进策略，但没有提及该参数的重要性（Torkzadehmahani等人（2019）同样如此，其使用了 $n_{\mathcal{D}} = 1$）。正如我们在图1a中所示，确保 $n_{\mathcal{D}}$ 处于关键范围内（由DPSGD超参数决定），是将GAN训练方案调整为差分隐私版本的关键； $n_{\mathcal{D}}$ 的选择决定了能否达到具有最先进水平竞争力的性能，还是会完全无法运行。[2]

因此，后续研究逐渐转向脱离差分隐私生成对抗网络，探索针对生成对抗网络的其他隐私化方案（Jordon 等人，2019年；Long 等人，2021年；Chen 等人，2020年；Wang 等人，2021年）。总体而言，这些方法通过 PATE 方法（Nissim 等人，2007年）采用子采样与聚合（Papernot 等人，2017年），将数据划分为 $\geq$ 1K 个互不相交的分区，并分别在每个分区上训练教师判别器。我们的工作表明，这些隐私化方案的表现不及差分隐私随机梯度下降。

差分隐私生成模型。 其他生成建模框架也已被用于生成差分隐私合成数据：变分自编码器（Chen 等人，2018年）、最大均值差异（Harder 等人，2021年；Vinaroz 等人，2022年；Harder 等人，2022年）、Sinkhorn散度（Cao 等人，2021年）、归一化流（Waites 和 Cummings，2021年）以及扩散模型（Dockhorn 等人，2022年）。另一方面，Chen 等人 （2022年）不采用生成模型，而是生成一个示例核心集（≈ 20 每个类别一个），用于训练分类器。这些方法可分为两类：将差分隐私随机梯度下降应用于现有的高性能生成模型；或专门为隐私设计的定制方法，而在以其非隐私上限进行评估时，这些方法的表现不及生成对抗网络（$\varepsilon \to \infty$）。

关于差分隐私扩散模型的同期研究。Dockhorn 等人（2022年）首次研究了扩散模型的差分隐私训练。他们在多种设置下取得了令人瞩目的最先进差分隐私图像合成结果，尤其是优于本文报告的差分隐私生成对抗网络结果。我们认为，我们的结果对学界仍具有重要意义，因为我们质疑了关于差分隐私生成对抗网络存在缺陷的传统观点，并表明它们的效用远高于以往的认知。事实上，生成对抗网络仍是最受欢迎、研究最充分的生成模型之一，因此在许多情况下，人们会更倾向于选择生成对抗网络，而非其他方法。通过重新审视差分隐私生成对抗网络中的若干设计选择，我们为如何将差分隐私无缝引入此类流程提供了指导。此外，我们的工作与Dockhorn 等人（2022年）的工作都支持一个更广泛的观点：使用差分隐私随机梯度下降训练传统机器学习架构，往往能在差分隐私条件下取得最先进的结果。事实上，我们双方的结果都优于几乎所有专为差分隐私图像合成设计的定制方法。这进一步印证了近期在其他隐私保护机器学习场景中得到验证的类似观点，包括图像分类（De 等人，2022年）和自然语言处理（Li 等人，2022年；Yu 等人，2022年）。

[2]关于超参数在差分隐私机器学习中作用的进一步讨论，请参阅附录 F。

3 预备知识

我们的目标是在敏感数据上训练一个可安全发布的生成模型，也就是说，该模型不会泄露训练数据集中个人的秘密。为此，我们确保训练算法 $\mathcal{A}$ ——该算法以敏感数据集 D 为输入，并输出训练完成的（生成）模型 θ 的参数——满足差分隐私。

定义 1（差分隐私，Dwork 等人 2006年b）。 随机算法 $\mathcal{A}: \mathcal{U} \to \Theta$ 是（ε，δ）-差分隐私的，当且仅当对于任意一对相邻数据集 D，$D' \in \mathcal{U}$，都有

$$\mathbb{P}\{\mathcal{A}(D) \in S\} \leq \exp(\varepsilon) \cdot \mathbb{P}\{\mathcal{A}(D') \in S\} + \delta \qquad \text{for all (measurable) } S \subseteq \Theta.$$

本文工作采用差分隐私的添加/删除定义，并将两个数据集 D 和 D' 视为相邻数据集，当且仅当它们至多有一项不同，也就是说，$D = D' \cup \{x\}$ 或 $D' = D \cup \{x\}$。

差分隐私的一个实用性质是后处理闭包，它指出，从差分隐私算法的输出进一步计算得到的结果（不以任何其他方式访问私有数据）可以安全发布，并满足与原始输出相同的差分隐私保证。在本文的情形中，这意味着与隐私化模型交互（例如，使用它在非敏感数据上计算梯度、生成样本）不会导致进一步的隐私侵犯。

差分隐私随机梯度下降。 基于梯度的学习算法可以通过采用差分隐私随机梯度下降（差分隐私随机梯度下降）（Song 等人，2013年；Bassily 等人，2014年；Abadi 等人，2016年）来替代随机梯度下降，而无需更改其他部分。差分隐私随机梯度下降通过裁剪逐样本梯度并向其总和添加高斯噪声，从而有效地限制并掩盖任意单个数据点对最终模型参数的贡献。差分隐私随机梯度下降的隐私分析基于差分隐私工具箱中的若干经典工具：高斯机制、子采样隐私放大和组合（Dwork 等人，2006年a；Dwork 和 Roth，2014年；Abadi 等人，2016年；Wang 等人，2019年）。

在我们的工作中，我们针对差分隐私随机梯度下降使用了两种不同的隐私核算方法：(a) Mironov 等人（2019年）提出的经典方法，由 Opacus（Yousefpour 等人，2021年）实现；(b) Gopi 等人（2021年）提出的最新精确隐私核算方法。默认情况下，为了与既有研究进行更直接的比较，我们采用前一种方法（不过需要指出的是，一些既有研究采用的核算方法甚至更宽松）。然而，后一种方法对实际隐私损失给出了更紧的界限；在所有实际应用中，它都是首选的隐私核算方法。我们仅在表 Gopi 等人（2021年）的核算方法1、2和3中明确标注的地方使用该方法。

差分隐私生成对抗网络。算法1 详细说明了差分隐私生成对抗网络的训练算法，它实际上是差分隐私随机梯度下降的一种具体实现。请注意，只有判别器的梯度 $\mathcal{D}$ 必须经过隐私保护处理（通过裁剪和添加噪声），而生成器的梯度 $\mathcal{G}$则不需要。这是后处理闭包的结果——生成器仅通过判别器参数间接接触敏感数据集，因此无需进一步进行隐私保护处理。

4 频繁增加判别器步数可改善私有生成对抗网络 S

本节讨论我们的主要发现： $n_{\mathcal{D}}$，每个生成器步骤之间执行的判别器步数（见算法 1）对 DPGAN 训练的成效有显著影响。

固定一组 DPSGD 超参数后， $n_{\mathcal{D}}$ 存在一个最优取值范围，可使生成质量最大化；生成质量既体现在视觉质量上，也体现在下游分类器训练的效用上。这个值可能相当大（某些情况下为 $n_{\mathcal{D}} \approx 100$ ）。

4.1 实验细节

设置。我们重点研究 MNIST（LeCun 等人，1998 年）和 FashionMNIST（Xiao 等人，2017年）的带标签生成。这两个数据集都包含 6 万张 28×28 灰度图像，分为 10 个类别。为了构建一个

算法 1 TrainDPGAN(D; ϕ_0, θ_0, OptD, OptG $n_{\mathcal{D}}$, T, B, C, σ, δ)

输入： 带标签数据集 $D = \{(x_j, y_j)\}_{j=1}^{n}$。判别器 $\mathcal{D}$ 和生成器 $\mathcal{G}$ 的初始化值 ϕ_0 和 θ_0。优化器 Optimizers OptD、OptG。超参数： $n_{\mathcal{D}}$ （每个 $\mathcal{G}$ 步的步数）， T （ $\mathcal{D}$ 步的总数）， B （预期批量大小）， C （裁剪范数），以及 σ （噪声水平）。隐私参数 δ。

$q \leftarrow B/|D|$ 和 t， $k \leftarrow 0$ ▷ 计算采样率 q，初始化计数器。

当 $t < T$ **时执行** ▷使用差分隐私随机梯度下降更新 $\mathcal{D}$ 。

  $S_t \sim$ 泊松采样(D， q) ▷ 通过以概率 q 纳入每个 $(x, y) \in D$ ，采样一个真实批次 S_t 。

  $\tilde{S}_t \sim \mathcal{G}(\cdot;\theta_k)^B$ ▷采样一个虚假批次 $\tilde{S}_t$。

  $g_{\phi_t} \leftarrow \sum_{(x,y)\in S_t}$ 裁剪（∇_{ϕ_t}（$-\log$（$\mathcal{D}$（x, y； ϕ_t））； C）
  $+\sum_{(\tilde{x},\tilde{y})\in\tilde{S}_t}$ 裁剪（∇_{ϕ_t}（$-\log$（$1-\mathcal{D}$（$\tilde{x}, \tilde{y}$； ϕ_t））； C） ▷裁剪逐样本梯度。

  $\widehat{g}_{\phi_t} \leftarrow \frac{1}{2B}$（$g_{\phi_t} + z_t$），其中 $z_t \sim \mathcal{N}(0, C^2\sigma^2 I)$ ▷添加高斯噪声。

  $\phi_{t+1} \leftarrow$ OptD(ϕ_t， $\widehat{g}_{\theta_t}$)

  $t \leftarrow t + 1$

  若 $n_{\mathcal{D}}$ 整除 t **，则** ▷每隔 $n_{\mathcal{D}}$ 步数执行一次 $\mathcal{G}$ 更新。

    $\tilde{S}'_t \sim \mathcal{G}(\cdot;\theta_k)^B$

    $g_{\theta_k} \leftarrow \frac{1}{B}\sum_{(\tilde{x},\tilde{y})\in\tilde{S}'_t} \nabla_{\theta_k}(-\log(\mathcal{D}(\tilde{x}, \tilde{y}; \phi_t)))$

    $\theta_{k+1} \leftarrow$ OptG(θ_k， g_{θ_k})

    $k \leftarrow k + 1$

  结束条件判断

结束循环

$\varepsilon \leftarrow$ 隐私会计器(T， σ， q， δ) ▷ 计算已消耗的隐私预算。

输出： 最终 $\mathcal{G}$ 参数 θ_k 以及(ε， δ)-差分隐私保证。

强基准，我们从一个开源的 PyTorch（Paszke 等人，2019年）深度卷积生成对抗网络实现开始[3] （Radford 等人，2016年），该实现以非隐私方式表现良好，因此我们沿用其训练方案。随后，我们针对自身需求调整其架构：移除与差分隐私随机梯度下降不兼容的批归一化层，并添加标签嵌入层以实现带标签生成。以非隐私方式训练此配置后，可实现带标签生成，在 MNIST 上取得 3.4 ± 0.1 FID 分数，在 FashionMNIST 上取得 16.5 ± 1.7 FID 分数。 $\mathcal{D}$和$\mathcal{G}$ 的可训练参数分别为 1.72M 和 2.27M。更多详情请参阅附录 B.1。

隐私实现。 为使训练具有隐私保护，我们使用 Opacus（Yousefpour 等人，2021年），它实现了逐样本梯度计算。如前所述，我们使用 R´enyi 差分隐私（RDP）会计方法，来自Mironov 等人（2019年）（少数特别注明的情况除外，在这些情况下我们改用更紧的Gopi 等人（2021年）会计方法）。对于基准设置，我们采用以下差分隐私随机梯度下降超参数：保持非隐私（期望）批量大小为 $B = 128$，并使用噪声水平 $\sigma = 1$ 和裁剪范数 $C = 1$。在这些设置下，当目标为(10， 10^{-5})-差分隐私时，我们的隐私预算足以支持 $T = 450$K 个判别器步数。

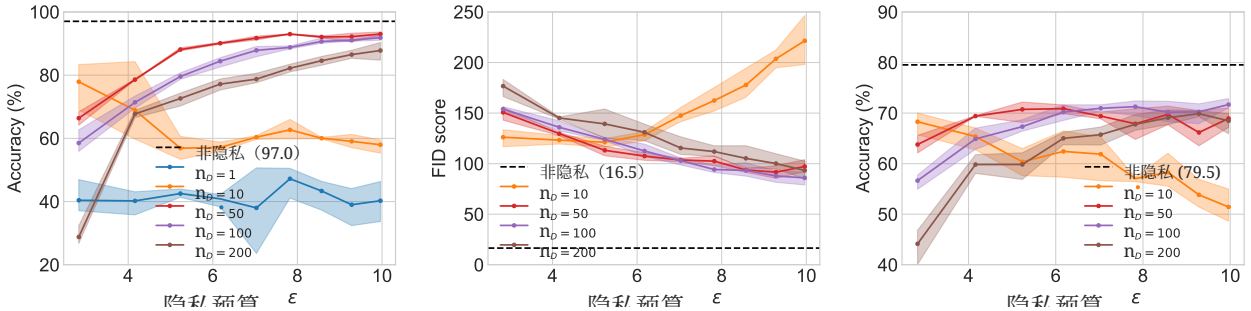
评估。 我们通过考察生成图像的视觉质量 和 对下游任务的效用 来评估我们的生成模型。遵循先前的工作，我们通过计算 Fr´echet Inception 距离（FID）（Heusel 等人，2017年）来衡量视觉质量，该距离计算基于 6 万张生成图像与完整测试集之间的差异。[4] 为了衡量下游任务的效用，我们再次遵循先前的工作，在 6 万个生成的图像—标签对上训练卷积神经网络分类器，并报告其在真实测试集上的准确率。

4.2 结果

更频繁的判别器步数能够改善生成效果。 我们在图 1a 和2 中绘制了在不同判别器更新频率下，MNIST 和 FashionMNIST 上 DPGAN 训练期间 FID 与准确率的变化 $n_{\mathcal{D}}$。该参数对最终结果的影响尤为显著。对于 MNIST， $n_{\mathcal{D}} = 50$ 能够取得最佳结果；对于 FashionMNIST， $n_{\mathcal{D}} = 100$ 效果最佳。

我们强调，相对于生成器步数，增加判别器步数的频率不会影响算法1的隐私成本。对于 $n_{\mathcal{D}}$的任何设置，我们执行相同数量的加噪梯度

[3]感谢 Hyeonwoo Kang（https://github.com/znxlwm）。代码可通过此链接获取。 [4]我们使用开源的 PyTorch 实现来计算 FID： https://github.com/mseitzer/pytorch-fid。



(a) MNIST 准确率　(b) FashionMNIST FID　(c) FashionMNIST 准确率

图 2： 采用不同判别器更新频率的 DPGAN 在各训练运行中的结果 $n_{\mathcal{D}}$，目标为 $(10,\ 10^{-5})$- 差分隐私。每条绘制的曲线表示 3 次采用不同随机种子的训练运行的效用均值、最小值和最大值，随着隐私预算的消耗而变化。 **(a)** 我们绘制了仅使用生成数据训练的卷积神经网络在测试集上的准确率。准确率与图 1a中的 FID 分数走势一致。从 $n_{\mathcal{D}} = 1$ 提高到 $n_{\mathcal{D}} = 50$，准确率从 $40.3 \pm 6.3\% \rightarrow 93.0 \pm 0.6\%$。进一步 $n_{\mathcal{D}}$ 提高会损害准确率。 **(b) 和 (c)** FashionMNIST 上也得到类似结果。请注意，最优 $n_{\mathcal{D}}$ 更高（约为 $n_{\mathcal{D}} \approx 100$）。在 $n_{\mathcal{D}} = 100$时，我们得到的 FID 为 85.9 ± 6.4，准确率为 $71.7 \pm 1.0\%$。

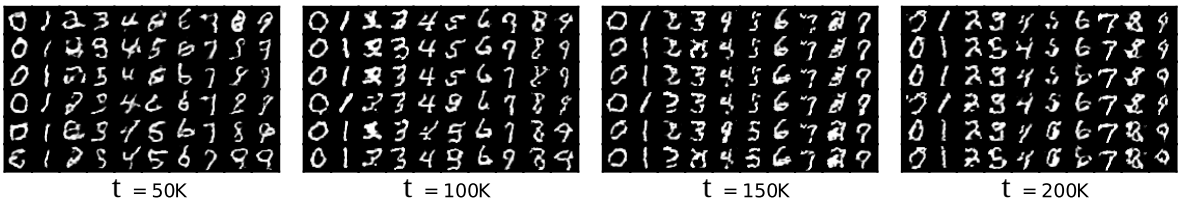


图 3： 训练期间抽取的样本演变过程，其中 $n_{\mathcal{D}} = 10$，目标为 $(10, 10^{-5})$-差分隐私。该设置在 $t = 50$K 次迭代（$\varepsilon \approx 2.85$）时取得最佳 FID 和下游准确率。随着训练继续超过这一点，我们观察到若干类别出现模式崩溃（例如数字 6 和 7，尤其是在 $t = 150$K 时），同时评估指标也随之恶化（这些样本对应于图 1a和 2a中 $n_{\mathcal{D}} = 10$ 线的前 4 个数据点）。

查询真实数据——变化之处在于，训练过程中执行的生成器步骤总数减少了 $n_{\mathcal{D}}$倍。

隐私GAN 正走向模式崩溃。 对于我们的 MNIST 结果，我们观察到，在较低的判别器更新频率（$n_{\mathcal{D}} = 10$）下，最佳 FID 和准确率分数在训练早期出现，远早于我们设定的隐私预算耗尽之前。[5] 查看图 1a 和2a 中 50K 个判别器步数处的数据点（图表中最左侧的点；$\varepsilon \approx 2.85$），$n_{\mathcal{D}} = 10$ 运行（橙色）的 FID 和准确率优于两者：(a) $n_{\mathcal{D}} = 10$ 运行在训练更久并消耗 更多隐私预算后的较晚检查点；以及 (b) 训练处于该阶段时的 $n_{\mathcal{D}}$ 其他设置。

我们将训练时间越长生成器性能越差归因于模式崩溃：这是生成对抗网络的一种已知失效模式，此时生成器会转而生成少量样本，而不是展现底层数据分布中的全部变化。在图3中，我们绘制了一个$n_{\mathcal{D}} = 10$ 运行中生成图像随训练进程的变化，并观察到模式崩溃的定性证据：在 50K 步时，所有生成图像各不相同；而在 150K 步时，许多列（尤其是数字 6 和 7）中的图像只是同一图像的细微变体。相比之下，成功训练的生成对抗网络不会出现这种现象（见 $n_{\mathcal{D}} = 50$ 图1b中的图像）。模式崩溃与 $n_{\mathcal{D}} = 10$ 运行（图橙色）前 4 个数据点中观察到的 FID 和准确率下降同时出现，见图1a 和2a。

[5]Neunhoeffer 等人 Neunhoeffer 等人 （2021年），曾报告过这一观察，这也促使他们采取补救措施：从训练过程中出现的中间模型中取样并进行混合。据我们所知，报告带标签图像合成 DPGAN 基准的论文中，并未提及 DPGAN 训练的这一方面。

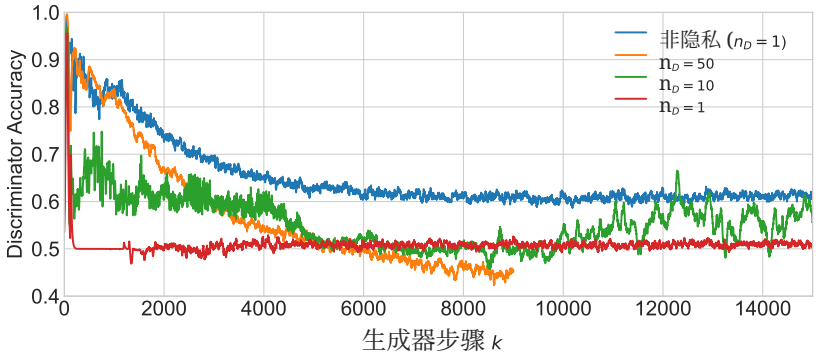


图 4：在每个生成器步骤之前，GAN 判别器在小批次上的准确率的指数移动平均值（$\beta = 0.95$）。非隐私情况下，判别器的准确率维持在 60%；而当 $n_{\mathcal{D}} = 1$ 时，隐私判别器实际上只能随机猜测。增加判别器步数，可以让判别器在训练初期重新获得优势，从而改进生成器。随着生成器不断改进，判别器的任务会变得更加困难，准确率也随之下降。

最优判别器更新频率。这些结果表明，在固定其他差分隐私随机梯度下降超参数的情况下，判别器步频存在一个最优设置 $n_{\mathcal{D}}$，能够在以下两种情况之间取得平衡：（a）步频过低，导致生成质量在训练早期达到峰值，随后发生模式崩溃；之后的训练只会消耗额外的隐私预算，却无法提升模型；（b）步频过高，使生成器在隐私预算耗尽前无法采取足够多的步数以实现收敛（图 $n_{\mathcal{D}} = 200$ 2a 中的运行就是一个例子）。在这两者之间取得平衡，可以最有效地利用隐私预算来提升生成器。

5 为什么增加步数会有帮助?

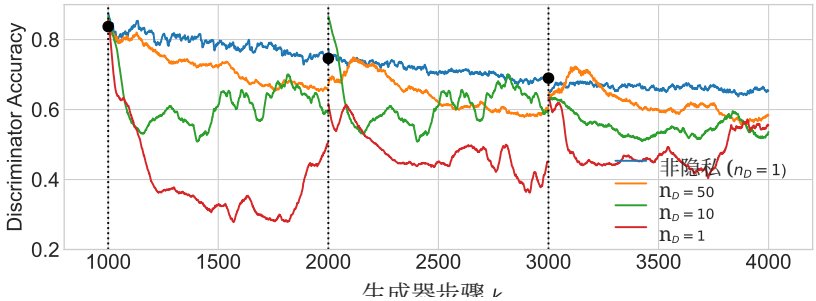
本节将展示实证结果，以帮助理解为什么更频繁地更新判别器能改善 DPGAN 训练。我们提出了一种与实证结果相符的解释。

差分隐私如何影响生成对抗网络训练? 图 4比较了在不同 $n_{\mathcal{D}}$ 设置下，私有训练与非隐私训练中，每次生成器步骤之前判别器对留出的真实和伪造样本的准确率。我们观察到，在非隐私训练中，当 $n_{\mathcal{D}} = 1$ 时，判别器准确率稳定在约 60%。直接引入差分隐私（$n_{\mathcal{D}} = 1$）会导致定性差异：训练一开始，差分隐私就使判别器准确率降至 50%（即与随机猜测的准确率相当），且之后始终无法恢复。[6]

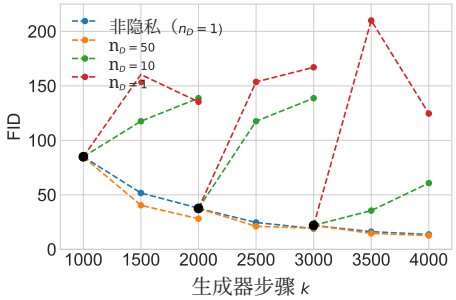
对于 $n_{\mathcal{D}}$ 的其他设置，我们有以下观察：（1）较大的 $n_{\mathcal{D}}$ 对应于训练早期较高的判别器准确率；（2）在一次训练过程中，随着生成器不断改进，判别器准确率会持续下降；（3）判别器准确率低于某个阈值后，生成器的表现会退化或改进有限。[7] 基于这些观察，我们提出以下解释，说明为什么更频繁地更新判别器会有所帮助：

- 当判别器能够有效地区分真实数据和伪造数据时，生成器就会得到改进。
- DP 对判别器引入的非对称噪声添加 使这项任务变得困难，导致生成器改进有限。
- 让判别器在固定的生成器上训练更长时间，可以提高其准确率，从而恢复到非隐私情形，即生成器与判别器达到平衡。

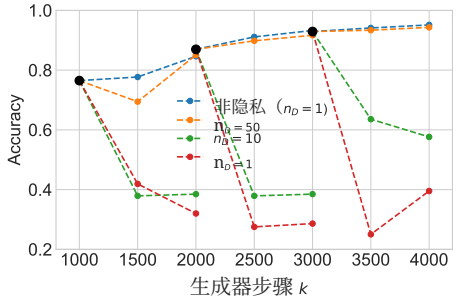
[6]我们的图仅显示前 15K 个生成器步数，但值得注意的是，这种情况会持续到训练结束（450K 步）。[7]对于 $n_{\mathcal{D}} = 10$，准确率在 5K $\mathcal{G}$ 步数（$= 50$K $\mathcal{D}$ 步）后降至 50% 以下，这对应于 $n_{\mathcal{D}} = 10$ 图 1a 和 2a中的实线上的第一个点。对于 $n_{\mathcal{D}} = 50$，准确率在 5K $\mathcal{G}$ 步数（$= 250$K $\mathcal{D}$ 步）后降至 50% 以下，这对应于 $n_{\mathcal{D}} = 50$ 图 1a 和 2a中实线上的第 5 个点。



(a) 指数　　检查点重启后，小批次上判别器准确率的移动平均值（$\beta = 0.95$）　　ts



(b) 检查点重启后的 FID



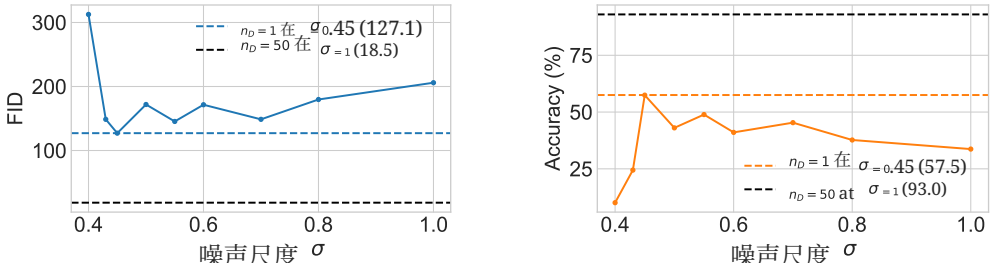
(c) 检查点重启后的准确率

图 5： 我们在非隐私训练进行到生成器步数为 1K、2K 和 3K 时取得的 3 个检查点处，采用不同的隐私和 $n_{\mathcal{D}}$ 设置重新开始训练。我们绘制了判别器准确率、FID 和下游分类准确率的变化过程。黑点表示检查点的初始值。我们观察到，较低的 $n_{\mathcal{D}}$设置无法达到与非隐私训练相当的判别器准确率（a），并会导致效用下降（（b）和（c））。对于 $n_{\mathcal{D}} = 50$，判别器准确率与非隐私训练相当；与非隐私设置一样，我们观察到效用在整个训练过程中持续提升。

检查点重启实验。我们进行了一项检查点重启实验，以便在更可控的设置下检验这一解释。我们对一个非隐私 GAN 训练 3K 个生成器步骤，并在 1K、2K 和 3K 个生成器步骤时保存 $\mathcal{D}$ 和 $\mathcal{G}$ （以及各自的优化器）的检查点。我们从每个检查点重启训练，在不同 $n_{\mathcal{D}}$ 和隐私设置下继续训练 1K 个生成器步骤。我们绘制了判别器准确率、FID 和下游分类准确率的变化情况。结果见图5。总体而言，我们的结果证实了以下观察：判别器准确率会随着 $n_{\mathcal{D}}$ 增大而提高，并随生成器性能改善而降低；当判别器准确率足够高时，生成器也会有所改进。

降低噪声能达到同样的效果吗？ 鉴于上述解释，我们想知道降低噪声水平 σ 能否像增加步数一样带来相同的提升，因为降低 σ 也应该能在生成器步骤之前提高判别器准确率。为验证这一点，我们从第 4 节中的设置出发，固定$n_{\mathcal{D}} = 1$，并以 $\varepsilon = 10$ 为目标数据集 MNIST，在一组噪声水平上 σ （其中最低的$\sigma = 0.4$所对应的预算仅允许进行 $T = 360$ 个判别器步数）进行搜索。结果如图 6 所示。我们在噪声水平 $\sigma = 0.45$ 下，取得了 127.1 的最佳 FID 和 57.5% 的最佳准确率。因此，我们可以得出结论：在这一实验设置中，将判别器更新频率纳入设计空间，能够更有效地利用隐私预算来提升生成质量。

增加判别器步数总是有帮助吗？ 正如我们在第6.1节中更详细讨论的那样，如果我们能通过增加步数以外的其他方式改进判别器，那么调整判别器更新频率可能不会带来提升。为了用一个极端案例加以说明，不妨考虑取消隐私约束。在非隐私 GAN 训练中，众所周知，增加步数并无必要。我们



(a) 仅改变 σ 与 $n_{\mathcal{D}} = 1$下 FID 的对比　　(b) 仅改变 σ 与 $n_{\mathcal{D}} = 1$下准确率的对比

图 6： 在 MNIST 上，我们固定 $n_{\mathcal{D}} = 1$，并报告 DPSGD 噪声水平 σ 在各种设置下的结果，其中迭代次数 T 针对每个 σ 进行选择，以达到 (10，10^{-5})-差分隐私。虚线之间的间隔表示将 $n_{\mathcal{D}}$ 的选择纳入我们的设计空间后，效用前沿的推进。

证实了这一结果：我们使用相同数量的生成器步骤，运行第4 节中的非隐私基线，但选择在每个生成器步骤之间执行 10 个判别器步数，而不是 1 个。FID 从$3.4 \pm 0.1 \to 8.3$恶化至，准确率则从 $97.0 \pm 0.1\% \to 91.3\%$ 恶化至。

6 更好的判别器带来更好的生成器

我们在第5节提出的解释，为改进生成对抗网络训练提供了具体建议：有效利用隐私预算，在判别器准确率足够高时，尽可能增加生成器步骤。我们针对这些目标，对隐私GAN训练方案进行了修改实验，从而提升了生成效果。

6.1 更大的批量大小

近期多项研究表明，对于分类任务，在相应调整噪声水平 σ 后，DPSGD使用更大的批量大小可取得更高的准确率（Tram`er和Boneh，2021年；Anil等人，2022年；De等人，2022年）。在较简单、较不多样化的数据集（如MNIST、CIFAR-10和FFHQ）上，GAN训练通常采用较小的批量大小（例如，深度卷积生成对抗网络使用 $B = 128$ （Radford等人，2016年），我们也采用这一设置；StyleGAN(2/3)使用 $B = 32/64$ （Karras等人，2019年、2020年、2021年））。[8] 因此，探究大批量是否有助于我们所研究的场景颇有意义。我们的实验验证了，在MNIST上，更大的批量大小并不能显著改善第4节中的非隐私基线：当批量大小从 $B = 128$增至 $B = 2048$ 时，FID从 $3.4 \pm 0.1 \to 3.2$ 变为 $97.0 \pm 0.1\% \to 97.5\%$。

结果。 我们扩大批量大小，考察 $B \in \{128, 512, 2048\}$，并搜索最佳噪声水平σ 和 $n_{\mathcal{D}}$ （详情见附录 B.2）。我们的目标是 $\varepsilon = 1$ 和 $\varepsilon = 10$。我们在表 2中报告超参数搜索所得的最佳结果。我们发现，更大的批量大小能带来改进：对于 $\varepsilon = 1$ 和$\varepsilon = 10$，分别在 $B = 512$ 和 $B = 2048$ 时取得最佳结果。我们还注意到，对于大批量，最佳生成器步数可能相当少。对于 $B = 2048$， $\sigma = 4.0$，在$\varepsilon = 10$时以 MNIST 为目标， $n_{\mathcal{D}} = 5$ 是最佳判别器更新频率，并且优于我们采用 $B = 128$ 的最佳 $n_{\mathcal{D}} = 50$设置。完整结果见附录 D.3。

6.2 自适应判别器步频

第4节和5节的观察结果促使我们考虑采用自适应的判别器步数频率。如图4和5a所示，随着生成器不断改进，训练期间判别器准确率会下降。在这种情况下，我们希望增加判别器的更新步数，以改进判别器，从而进一步改进

[8]然而，在更复杂、更多样化的数据集（如 ImageNet）上，研究发现较大的批量大小有所帮助：BigGAN（Brock 等人，2019年）得出了这一结论。近期将 StyleGAN 扩展到多样化数据集的工作——StyleGAN-XL（Sauer 等人，2022年）和 GigaGAN（Kang 等人，2023年）——也证实了这一结果，发现将批量大小分别扩大到 2048 和 1024 能带来改进。

隐私 ε	方法	发表于	MNIST		FashionMNIST	
			FID	准确率（%）	FID	准确率（%）
$\varepsilon = \infty$	真实数据	本文工作	1.0	99.2	1.5	92.5
	GAN	本文工作	3.4 ± 0.1	97.0 ± 0.1	16.5 ± 1.7	79.5 ± 0.8
$\varepsilon = 10$	DP-MERF	Cao 等人（2021年）	116.3	82.1	132.6	75.5
	DP-Sinkhorn	Cao 等(2021年)	48.4	83.2	128.3	75.1
	PSG[9]	Chen 等(2022年)	-	95.6	-	77.7
	DPDM	Dockhorn 等(2022年)	5.01	97.3	18.6	84.9
	DPGAN [10]	Chen 等(2020年)	179.16	63	243.80	50
		Long 等(2021年)	304.86	80.11	433.38	60.98
	GS-WGAN	Chen 等人(2020年)	61.34	80	131.34	65
	PATE-GAN	Long 等人 (2021年)	253.55	66.67	229.25	62.18
	G-PATE	Long 等人 (2021年)	150.62	80.92	171.90	69.34
	DataLens	Wang 等（2021年）	173.50	80.66	167.68	70.61
$\varepsilon = 9.32$*	我们的 DPGAN	本文工作	18.5 ± 0.9	93.0 ± 0.6	85.9 ± 6.4	71.7 ± 1.0
	+ 大批次	本文工作	13.2 ± 1.1	94.0 ± 0.6	70.9 ± 6.3	73.0 ± 1.1
	+ 自适应 $n_{\mathcal{D}}$	本文工作	12.8 ± 0.3	95.1 ± 0.1	62.3 ± 8.7	74.7 ± 0.4
$\varepsilon = 1$	DP-MERF[11]	Vinaroz 等人（2022年）	-	80.7	-	73.9
	DP-HP	Vinaroz 等人（2022年）	-	81.5	-	72.3
	PSG	Chen（2022年）	-	80.9	-	70.2
	DPDM	Dockhorn 等人（2022年）	23.4	95.3	37.8	79.4
	DPGAN	Long 等人（2021年）	470.20	40.36	472.03	10.53
	GS-WGAN	Long 等人（2021年）	489.75	14.32	587.31	16.61
	PATE-GAN	Long 等（2021年）	231.54	41.68	253.19	42.22
	G-PATE	Long 等（2021年）	153.38	58.80	214.78	58.12
	DataLens	Wang 等（2021年）	186.06	71.23	194.98	64.78
$\varepsilon = 0.912$*	我们的 DPGAN	本文工作	111.1 ± 17.9	76.9 ± 0.6	155.3 ± 7.1	64.9 ± 0.8
	+ 大批次	本文工作	106.2 ± 64.0	67.5 ± 7.8	158.9 ± 6.0	67.2 ± 1.0
	+ 自适应 $n_{\mathcal{D}}$	本文工作	52.6 ± 3.2	81.3 ± 0.8	126.4 ± 4.1	69.1 ± 0.1

表 2：我们汇总了文献中先前报告的各种方法在 MNIST 和 FashionMNIST 带标签生成任务上的性能结果，并与我们的结果进行比较。对于我们的结果，我们使用 3 个随机种子运行实验，并报告均值 $\pm$ 标准差。请注意， *Reported In* 指的是数值结果的来源，而非该方法的提出者。对于下游准确率，我们报告所用分类器中的最佳准确率，并与我们的卷积神经网络分类器准确率进行比较。**(*)** 对于我们的结果，我们以 $\varepsilon = 10/\varepsilon = 1$ 为目标，使用 Opacus 进行隐私核算，并额外报告 ε，该结果采用了 Gopi 等人 (2021年)提出的改进隐私核算方法。

生成器。不过，从训练一开始就采用较高的判别器更新频率是浪费的——正如低 $n_{\mathcal{D}}$ 在训练早期就取得最佳 FID 和准确率所表明的那样。因此，我们提议从较低的判别器更新频率（$n_{\mathcal{D}} = 1$）开始，并在判别器表现不佳时逐步提高。真实数据上的准确率必须以差分隐私方式发布。虽然这可行，但会带来一个额外问题：必须找到最佳性能所需的隐私预算分配方式。我们观察到，判别器准确率仅与其在假样本上的准确率有关（假样本可通过后处理免费评估）。因此，我们将其用作评估判别器性能的代理指标。

我们提出了一种自适应步频，由 β 和 d参数化。 β 是衰减参数，用于在每次生成器更新前计算判别器在假批次上的准确率的指数移动平均（EMA）。 d 是准确率下限，达到该下限后，我们便转到下一个更新频率$n_{\mathcal{D}} \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, ...\}$。此外，在转到下一个更新频率前，我们承诺给予 $2/(1-\beta)$ 个生成器步数的宽限期——这是基于以下事实：β-EMA 的值主要

[9]由于 PSG 仅生成包含 200 个样本（每类 20 个）的核心集，其 InceptionNet 提取特征的协方差矩阵是奇异的，因此无法计算 FID 分数。[10]我们将按类别划分的无条件生成对抗网络与条件生成对抗网络一并归入 DPGAN。[11]Vinaroz 等人Vinaroz 等人(2022年的结果以图形形式呈现在论文中。具体数值可在其代码中找到。

隐私	方法	发表于	FID	准确率（%）
$\varepsilon = \infty$	真实数据	本文工作	1.1	96.6
	GAN		30.0 ± 1.6	92.0 ± 0.4
$\varepsilon = 10$	DP-MERF	Cao 等人（2021年）	274.0	65
	DP-Sinkhorn	Cao 等人（2021年）	189.5	76.3
	DPDM	Dockhorn 等人（2022年）	21.1	-
	DPGAN	Long 等（2021年）	-	54.09
	GS-WGAN	Long 等（2021年）	-	63.26
	PATE-GAN	Long 等（2021年）	-	58.70
	G-PATE	Long 等人（2021年）	-	70.72
$\varepsilon = 9.39$*	我们的 DPGAN	本文工作	170.8 ± 20.3	82.4 ± 4.4
$\varepsilon = 10$	DPGAN	Long 等人（2021年）	485.41	52.11
	GS-WGAN	Long 等人 (2021年)	432.58	61.36
	PATE-GAN	Long 等人 (2021年)	424.60	65.35
	G-PATE	Long 等人(2021年)	305.92	68.97
	DataLens	Wang 等人 (2021年)	320.84	72.87

表 3：表格上半部分：在 32×32 CelebA-Gender 上与最先进结果的比较，目标是 $(\varepsilon, 10^{-6})$-差分隐私（DP）（Long 等人（2021年）报告的结果除外，这些结果针对较弱的 $(\varepsilon, 10^{-5})$-差分隐私（DP））。我们使用 3 个随机种子运行实验，并报告均值 $\pm$ 标准差。（*）对于我们的结果，我们使用 Opacus 隐私核算，目标为 $\varepsilon = 10$，并额外报告采用 Gopi 等人（2021年）改进的隐私核算所得的 ε。DPDM 的 FID 分数远优于我们的 DPGAN（而我们的 DPGAN 本身也优于此前的结果）。我们的 DPGAN 达到了已报告的最佳准确率分数。表格下半部分：Long 等人（2021年）和 Wang 等人（2021年）报告的基于 GAN 的方法的结果；由于这些方法针对 $(10, 10^{-5})$-差分隐私（DP）并使用 64×64 CelebA-Gender，因此无法直接比较。

由最近的 $2/(1-\beta)$ 次观测决定。

在所有设置中，我们都使用 $\beta = 0.99$，并尝试了 $d = 0.6$ 和 $d = 0.7$。自适应步频的另一项优势是，无需搜索最佳更新频率。尽管自适应步频引入了阈值 d 这一额外超参数，但我们发现，这两种设置（$d = 0.6$ 和 $d = 0.7$）足以超越通过更大规模的超参数搜索得到的结果；该搜索针对 $n_{\mathcal{D}}$ 进行，其最优值会随噪声水平 σ 和预期批量大小 B 显著变化。

6.3 与文献中已有结果的比较

6.3.1 MNIST 和 FashionMNIST

表2 总结了我们在 MNIST 和 FashionMNIST 上表现最佳的实验设置，并将其置于该任务此前已报告结果的背景下。我们还在图7中进行了直观比较。图9 和10 展示了 $\varepsilon = 10$的一些生成图像示例；图11 和12 则展示了 $\varepsilon = 1$。

朴素的差分隐私随机梯度下降优于所有替代的 GAN 隐私保护方案。我们在第 4 节中的基准 DPGAN，在适当选择 $n_{\mathcal{D}}$（且尚未采用本节所述的修改）后，优于文献中提出的所有其他基于 GAN 的方法（GS-WGAN、PATE-GAN、G-PATE 和 DataLens），并且在两个指标、两个数据集和两个隐私级别上均表现如此。

更大的批量大小和自适应判别器步频能够改善 GAN 训练。总体而言，在两个隐私级别和两个数据集上，我们发现增大批量大小可以带来改进，而采用自适应步频后还能进一步提升。

与最先进方法的比较。除 DPDM 外，我们表现最佳的 DPGAN 与差分隐私合成数据领域的最先进方法相比具有竞争力，尤其是在 FID 分数方面。



图 7： MNIST 和 FashionMNIST 在不同方法下（10， 10^{-5}）-差分隐私的结果。其他方法的图像来自 Cao 等人 （2021年）和 Dockhorn 等人 （2022年）。

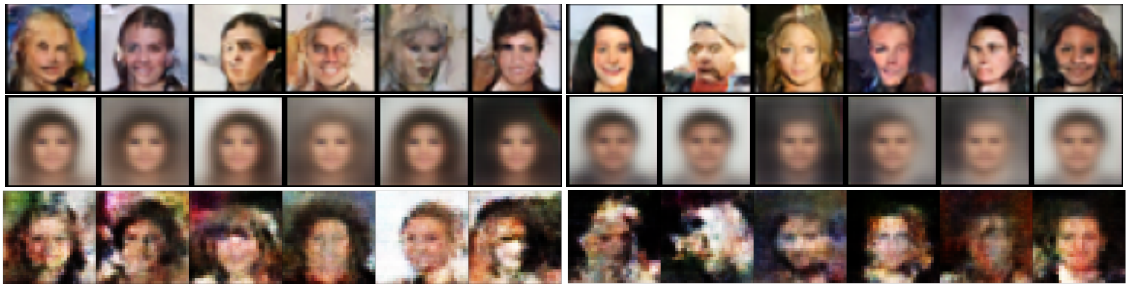


图8： 32×32 CelebA-Gender 在（10， 10^{-6}）-差分隐私下的结果。 **从上到下：** DPDM（无条件生成）、DP-Sinkhorn 和我们的 DPGAN。其他方法的图像来自 Cao 等人 （2021年）和 Dockhorn 等人 （2022年）。

6.3.2 CelebA-Gender

我们还报告了在（10 32×32 , 10^{-6})-差分隐私下按性别条件生成CelebA， 10^{-6}的结果。在这些实验中，我们使用了大批次（$B = 2048$）以及阈值为$d = 0.6$的自适应判别器步频。完整的实现细节见附录C。结果汇总于表3 ，并在图8中展示。更多生成示例请见图13。

7 结论

我们重新审视了差分隐私 GAN，并表明，只要对训练过程进行适当调整，其表现就能远超以往的预期。一些关键改动包括提高判别器步频、增大批量大小，以及引入自适应判别器步频。我们探讨了这样一种假设：以往 DPGAN 表现不佳，是因为判别器的分类准确率较低。更广泛地说，我们的工作支持了一项反复得到验证的发现：在传统架构上精心调整差分隐私随机梯度下降，能够在差分隐私机器学习中取得出色成果。

致谢

我们感谢 TML 感谢匿名审稿人 R 和责任编辑 Yu-Xiang Wang 提供的建设性反馈 .

参考文献

Martin Abadi、Andy Chu、Ian Goodfellow、H. Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。具有差分隐私的深度学习。载于 *CCS'16*：*2016*年 *ACM SIGSAC*计算机与通信安全大会，2016年。Rohan Anil、Badih Ghazi、Vineet Gupta、Ravi Kumar 和 Pasin Manurangsi。大规模差分隐私 BERT。载于计算语言学协会研究成果：*EMNLP'22*，2022年。Martin Arjovsky、Soumith Chintala 和 L´eon Bottou。Wasserstein生成对抗网络。载于第*34*届国际机器学习会议（*ICML'17*）论文集，2017年。Raef Bassily、Adam Smith 和 Abhradeep Thakurta。隐私经验风险最小化：高效算法与紧致误差界。载于 第*55*届*IEEE*计算机科学基础年会（*FOCS'14*），2014年。Brett K. Beaulieu-Jones、Zhiwei Steven Wu、Chris Williams、Ran Lee、Sanjeev P. Bhavnani、James Brian Byrd 和 Casey S. Greene。保护隐私的生成式深度神经网络支持临床数据共享。《循环：心血管质量与结局》，12(7)，2019年。Andrew Brock、Jeff Donahue 和 Karen Simonyan。用于高保真自然图像合成的大规模生成对抗网络训练。载于第*7*届国际学习表征会议（*ICLR'19*），2019年。Tianshi Cao、Alex Bie、Arash Vahdat、Sanja Fidler 和 Karsten Kreis。别生成我：使用 Sinkhorn散度训练差分隐私生成模型。载于 神经信息处理系统进展第*34*卷（*NeurIPS'21*），2021年。Tatjana Chavdarova、Matteo Pagliardini、Sebastian U. Stich、Fran¸cois Fleuret 和 Martin Jaggi。通过前瞻极大极小驯服生成对抗网络。载于 第*9*届国际学习表征会议（*ICLR'21*），2021年。Dingfan Chen、Tribhuvanesh Orekondy 和 Mario Fritz。GS-WGAN：一种用于学习差分隐私生成器的梯度净化方法。载于 神经信息处理系统进展第*33*卷（*NeurIPS'20*），2020年。Dingfan Chen、Raouf Kerkouche 和 Mario Fritz。利用判别信息生成隐私集合。载于 神经信息处理系统进展第*35*卷（*NeurIPS'22*），2022年。Qingrong Chen、Chong Xiang、Minhui Xue、Bo Li、Nikita Borisov、Dali Kaafar 和 Haojin Zhu。差分隐私数据生成模型。*CoRR*，abs/1812.02274，2018年。Soumith Chintala、Emily Denton、Martin Arjovsky 和 Michael Mathieu。如何训练生成对抗网络？让生成对抗网络发挥作用的技巧与诀窍。https://github.com/soumith/ganhacks，2016年。Soham De、Leonard Berrada、Jamie Hayes、Samuel L Smith 和 Borja Balle。通过尺度调整，解锁高精度差分隐私图像分类。*CoRR*，abs/2204.13650，2022年。Tim Dockhorn、Tianshi Cao、Arash Vahdat 和 Karsten Kreis。差分隐私扩散模型。*CoRR*，abs/2210.09929，2022年。Cynthia Dwork 和 Aaron Roth。差分隐私的算法基础。理论计算机科学基础与趋势，9(3-4):211–407，2014年。Cynthia Dwork、Krishnaram Kenthapadi、Frank McSherry、Ilya Mironov 和 Moni Naor。我们的数据，我们自己：通过分布式噪声生成实现隐私保护。载于第*25*届密码技术理论与应用国际年会（*EUROCRYPT'06*），2006年a。Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith。在隐私数据分析中根据敏感度校准噪声。载于第*3*届密码学理论会议论文集（*TCC'06*），2006年b。

Tanner Fiez 和 Lillian J. Ratliff。有限时间尺度分离下梯度下降上升的局部收敛分析。载于第*9*届国际学习表征大会 *(ICLR'21)*，2021年。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。载于神经信息处理系统进展第*27*卷 *(NIPS'14)*，2014年。Sivakanth Gopi、Yin Tat Lee 和 Lukas Wutschitz。差分隐私的数值组合。载于神经信息处理系统进展第*34*卷 *(NeurIPS'21)*，2021年。Frederik Harder、Kamil Adamczewski 和 Mijung Park。DP-MERF：利用随机特征实现差分隐私均值嵌入，以进行实用的隐私保护数据生成。载于第*24*届国际人工智能与统计会议 *(AISTATS'21)*，2021年。Frederik Harder、Milad Jalali Asadabadi、Danica J. Sutherland 和 Mijung Park。差分隐私数据生成需要更优特征。$CoRR$，abs/2205.12900，2022年。Moritz Hardt、Katrina Ligett 和 Frank Mcsherry。一种简单实用的差分隐私数据发布算法。载于神经信息处理系统进展第*25*卷 *(NIPS'12)*，2012年。Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。采用双时间尺度更新规则训练的生成对抗网络会收敛到局部纳什均衡。载于神经信息处理系统进展第*30*卷 *(NIPS'17)*，2017年。James Jordon、Jinsung Yoon 和 Mihaela van der Schaar。PATE-GAN：生成具有差分隐私保证的合成数据。载于第*7*届国际学习表征大会 *(ICLR'19)*，2019年。Minguk Kang、Jun-Yan Zhu、Richard Zhang、Jaesik Park、Eli Shechtman、Sylvain Paris 和 Taesung Park。扩展生成对抗网络以实现文本到图像合成。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR'23)*，2023年。Tero Karras、Samuli Laine 和 Timo Aila。用于生成对抗网络的基于风格的生成器架构。载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR'19)*，2019年。Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。分析并提升StyleGAN的图像质量。载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议论文集 *(CVPR'20)*，2020年。Tero Karras、Miika Aittala、Samuli Laine、Erik H¨ark¨onen、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。无混叠生成对抗网络。载于神经信息处理系统进展第*34*卷 *(NeurIPS'21)*，2021年。Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于第三届国际学习表征会议 *(ICLR'15)*，2015年。Yann LeCun、L´eon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。《*IEEE*论文集》，86(11):2278–2324，1998 年。Xuechen Li、Florian Tram`er、Percy Liang 和 Tatsunori Hashimoto。大语言模型可以成为强大的差分隐私学习器。载于第*10*届国际学习表征大会 *(ICLR'22)*，2022年。Jingcheng Liu 和 Kunal Talwar。从隐私候选项中进行隐私选择。载于第*51*届*ACM SIGACT*计算理论年会论文集 *(STOC'19)*，2019年。Ziwei Liu、Ping Luo、Xiaogang Wang 和 Xiaoou Tang。野外人脸属性的深度学习。载于*2015*年*IEEE*国际计算机视觉会议论文集 *(ICCV'15)*，2015年。

Yunhui Long、Boxin Wang、Zhuolin Yang、Bhavya Kailkhura、Aston Zhang、Carl Gunter 和 Bo Li。G-PATE：通过教师判别器的私有聚合实现可扩展的差分隐私数据生成器。载于神经信息处理系统进展第*34*卷 *(NeurIPS'21)*，2021年。Ryan McKenna、Daniel Sheldon 和 Gerome Miklau。用于差分隐私的基于图模型的估计与推断。载于第*36*届国际机器学习会议论文集 *(ICML'19)*，2019年。Ilya Mironov、Kunal Talwar 和 Li Zhang。抽样高斯机制的 Rényi 差分隐私。$CoRR$，abs/1908.10530，2019年。Takeru Miyato、Toshiki Kataoka、Masanori Koyama 和 Yuichi Yoshida。用于生成对抗网络的谱归一化。载于第*6*届国际学习表征会议 *(ICLR'18)*，2018年。Shubhankar Mohapatra、Sajin Sasy、Xi He、Gautam Kamath 和 Om Thakkar。自适应优化器在诚实的隐私超参数选择中的作用。载于第三十六届*AAAI*人工智能会议 *(AAAI-22)*，2022年。Marcel Neunhoeffer、Steven Wu 和 Cynthia Dwork。私有后GAN提升。载于第*9*届国际学习表征大会 *(ICLR'19)*，2021年。Kobbi Nissim、Sofya Raskhodnikova 和 Adam Smith。私有数据分析中的平滑敏感度与抽样。载于第*39*届*ACM*计算理论年度研讨会 *(STOC'07)*，2007年。´Nicolas Papernot、Mart´ın Abadi、Ulfar Erlingsson、Ian J. Goodfellow 和 Kunal Talwar。利用私有训练数据进行深度学习的半监督知识迁移。载于第*5*届国际学习表征大会 *(ICLR'17)*，2017年。Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas K¨opf、Edward Z. Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。PyTorch：一种命令式风格的高性能深度学习库。载于神经信息处理系统进展第*32*届 *(NeurIPS'19)*，2019年。Natalia Ponomareva、Hussein Hazimeh、Alex Kurakin、Zheng Xu、Carson Denison、H. Brendan McMahan、Sergei Vassilvitskii、Steve Chien 和 Abhradeep Guha Thakurta。如何让机器学习实现差分隐私：差分隐私机器学习实用指南。人工智能研究期刊，77:1113–1201，2023年。Alec Radford、Luke Metz 和 Soumith Chintala。利用深度卷积生成对抗网络进行无监督表征学习。载于第*4*届国际学习表征会议 *(ICLR'16)*，2016年。Axel Sauer、Katja Schwarz 和 Andreas Geiger。StyleGAN-XL：将StyleGAN扩展至大型多样化数据集。载于计算机图形学与交互技术特别兴趣小组会议 *(SIGGRAPH'22)*，2022年。Shuang Song、Kamalika Chaudhuri 和 Anand D. Sarwate。采用差分隐私更新的随机梯度下降。载于*2013*年*IEEE*全球信号与信息处理会议，2013年。Yuchao Tao、Ryan McKenna、Michael Hay、Ashwin Machanavajjhala 和 Gerome Miklau。差分隐私合成数据生成算法的基准测试。$CoRR$，abs/2112.09238，2021年。Reihaneh Torkzadehmahani、Peter Kairouz 和 Benedict Paten。DP-CGAN：差分隐私合成数据与标签生成。载于*IEEE/CVF*计算机视觉与模式识别会议研讨会，*(CVPR Workshops'19)*，2019年。Florian Tram`er 和 Dan Boneh。差分隐私学习需要更优特征（或更多得多的数据）。载于第*9*届国际学习表征大会 *(ICLR'21)*，2021年。

Margarita Vinaroz、Mohammad-Amin Charusaie、Frederik Harder、Kamil Adamczewski 和 Mi Jung Park。用于私有数据生成的厄米多项式特征。载于第*39*届国际机器学习大会（*ICML'22*）论文集，2022年。Chris Waites 和 Rachel Cummings。用于隐私保护密度估计的差分隐私归一化流。*CoRR*，abs/2103.14068，2021年。Boxin Wang、Fan Wu、Yunhui Long、Luka Rimanic、Ce Zhang 和 Bo Li。DataLens：通过梯度压缩与聚合实现可扩展的隐私保护训练。载于*CCS'21*：*2021*年*ACM SIGSAC*计算机与通信安全大会，2021年。Yu-Xiang Wang、Borja Balle 和 Shiva Prasad Kasiviswanathan。子采样 R´ényi 差分隐私与解析矩会计方法。载于第 *22* 届国际人工智能与统计大会（*AISTATS'19*），2019 年。Han Xiao、Kashif Rasul 和 Roland Vollgraf。Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。*CoRR*，abs/1708.07747，2017年。Liyang Xie、Kaixiang Lin、Shu Wang、Fei Wang 和 Jiayu Zhou。差分隐私生成对抗网络。*CoRR*，abs/1802.06739，2018年。Ashkan Yousefpour、Igor Shilov、Alexandre Sablayrolles、Davide Testuggine、Karthik Prasad、Mani Malek、John Nguyen、Sayan Gosh、Akash Bharadwaj、Jessica Zhao、Graham Cormode 和 Ilya Mironov。Opacus：PyTorch 中易于使用的差分隐私库。*CoRR*，abs/2109.12298，2021年。Da Yu、Saurabh Naik、Arturs Backurs、Sivakanth Gopi、Huseyin A. Inan、Gautam Kamath、Janardhan Kulkarni、Yin Tat Lee、Andre Manoel、Lukas Wutschitz、Sergey Yekhanin 和 Huishuai Zhang。语言模型的差分隐私微调。载于第*10*届国际学习表征大会（*ICLR'22*），2022年。Jun Zhang、Graham Cormode、Cecilia M. Procopiuc、Divesh Srivastava 和 Xiaokui Xiao。PrivBayes：通过贝叶斯网络发布私有数据。*ACM*数据库系统汇刊，42(4)，2017年。

生成的样本

我们提供了一些未经筛选的 MNIST 和 FashionMNIST 样本，见 $\varepsilon = 10$ 和 $\varepsilon = 1$；此外，还有32×32 CelebA-Gender 样本，见 $\varepsilon = 10$。

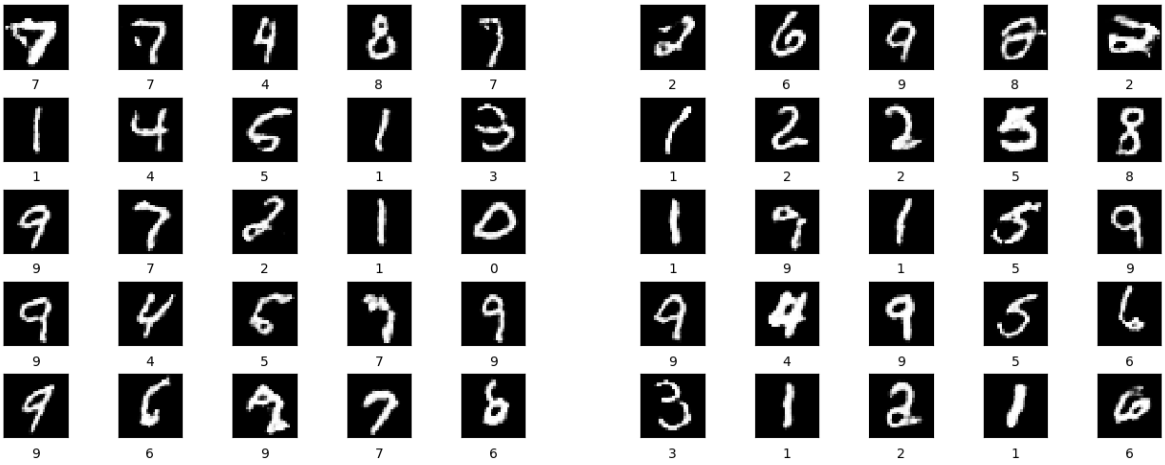


图 9： 我们的方法生成的一些未经挑选的 MNIST 样本， $\varepsilon = 10$。

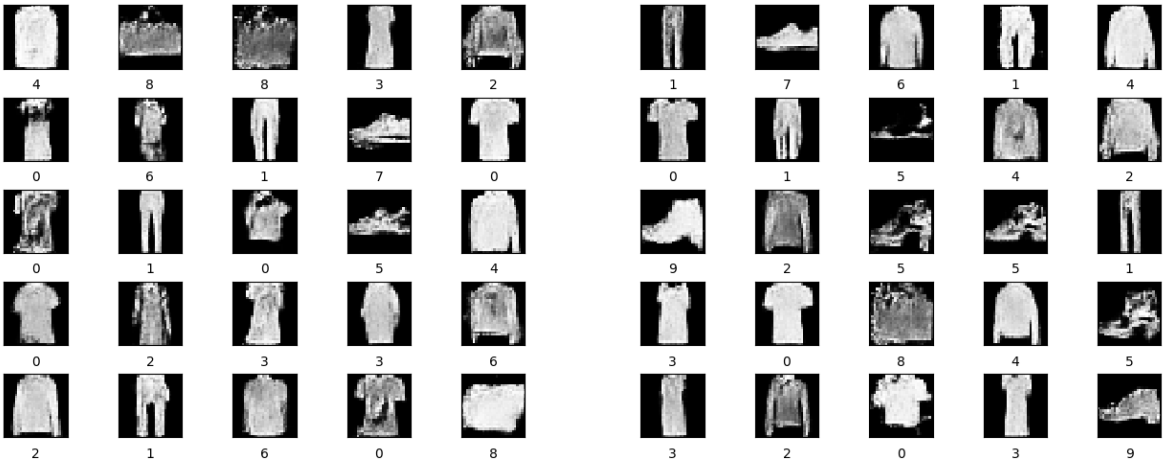


图 10： 我们的方法生成的一些未经挑选的 FashionMNIST 样本， $\varepsilon = 10$。

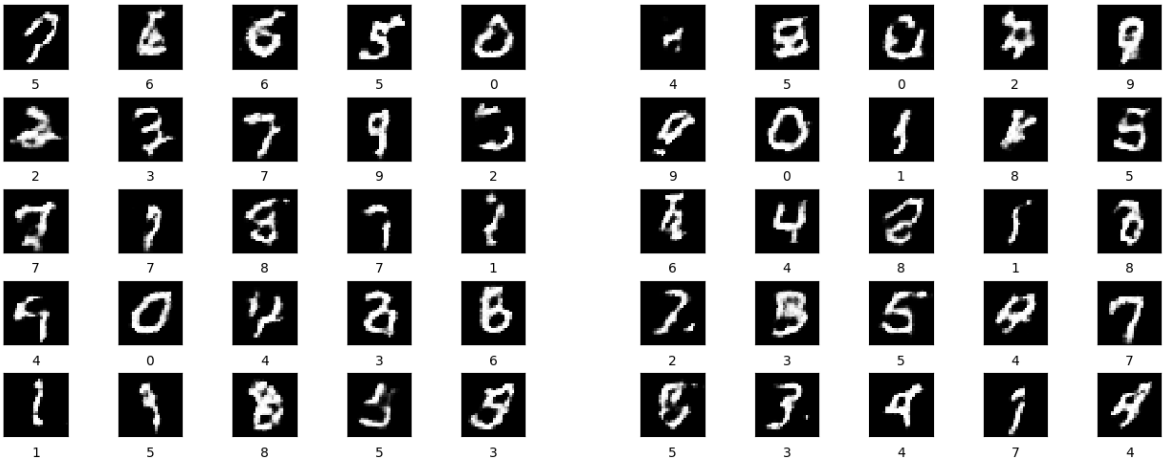


图 11： 我们的方法生成的一些未经挑选的 MNIST 样本， $\varepsilon = 1$。

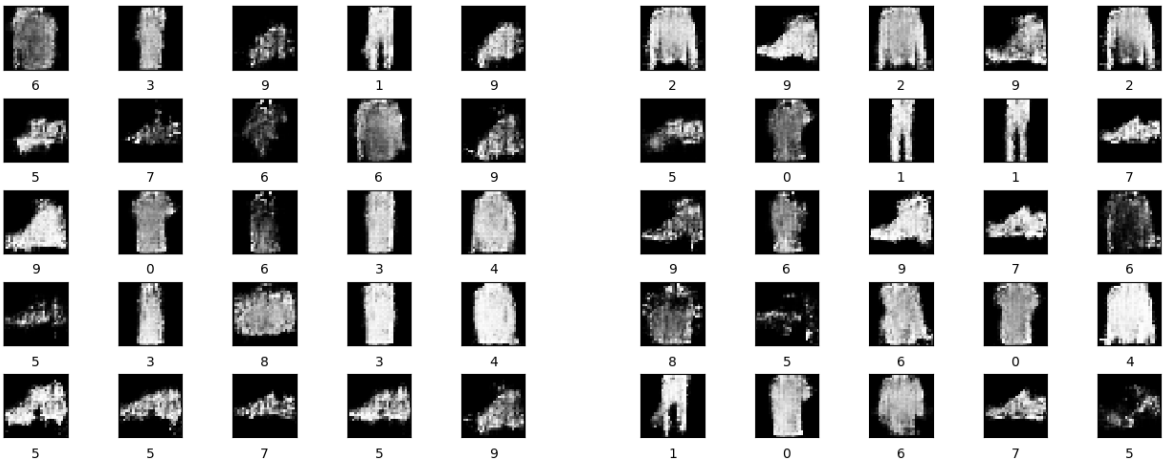


图 12： 我们的方法生成的一些未经挑选的 FashionMNIST 样本， $\varepsilon = 1$。

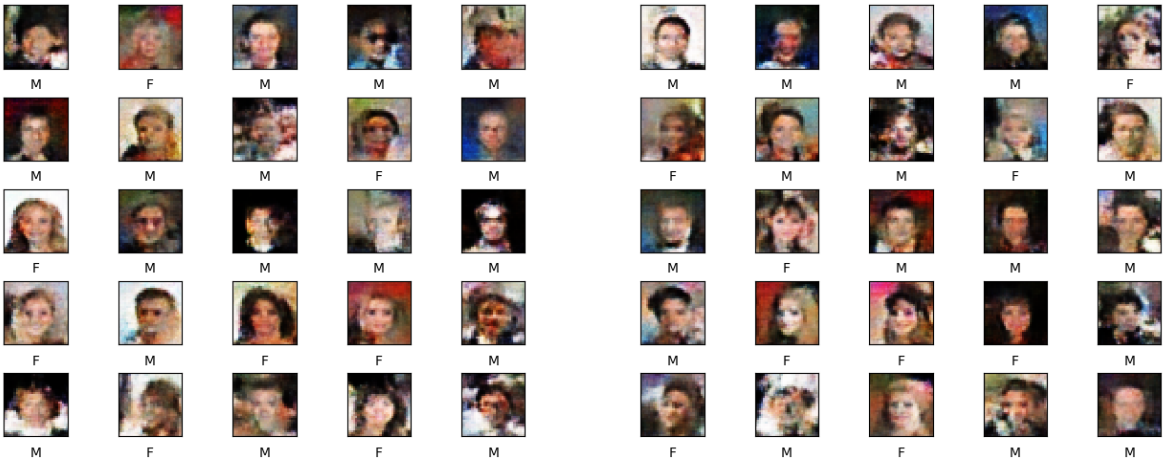


图13：我们的方法生成的一些未经挑选的 CelebA 样本，$\varepsilon = 10$。

B MNIST 和 FashionMNIST 实现细节

B.1 训练方案

对于 MNIST 和 FashionMNIST，我们基于一个开源的 PyTorch 深度卷积生成对抗网络实现（Radford 等人，2016年）（可在此处链接获取），该实现具有良好的非隐私性能，并沿用其训练方案。方案包括：批量大小 $B = 128$，Adam 优化器（Kingma 和 Ba，2015年），其参数为（$\alpha = 0.0002$，$\beta_1 = 0.5$，$\beta_2 = 0.999$），分别用于 $\mathcal{G}$ 和 $\mathcal{D}$，非饱和 GAN 损失（Goodfellow 等人，2014年），以及一个宽度参数为 $d = 128$ 的五层全卷积架构。

为使其适用于我们的目的，我们对架构进行了三项修改：在 $\mathcal{G}$ 和 $\mathcal{D}$ 中，我们（1）移除所有批归一化层（这些层与差分隐私随机梯度下降不兼容）；（2）添加标签嵌入层，以支持带标签生成；（3）调整卷积/转置卷积的步幅和卷积核尺寸，并移除最后一层，以便处理 $1 \times 28 \times 28$ 图像而无需调整大小。最后，我们移除其自定义权重初始化，改用 PyTorch 默认设置。

我们的基准非隐私生成对抗网络训练了 45K 步。我们也采用泊松抽样训练非隐私生成对抗网络：在每一步判别器训练中，以概率 B/n 独立地将数据集中的每个元素纳入抽样，其中 n 是数据集的大小。然后，我们添加从 $\mathcal{G}$ 抽取的 B 个假样本，组成假样本与真实样本的混合批次。

裁剪假样本梯度。使用差分隐私随机梯度下降私密训练判别器时，我们抽取 B个假样本，并在真实样本与假样本组成的整个混合批次上计算经过裁剪的逐样本梯度（见算法1）。此前的 Torkzadehmahani 等人(2019年的工作采用了这种方法。需要说明的是，这纯粹是一种设计选择——没有必要裁剪假样本的梯度，也不必在同一批次中一并处理这些样本。只要保留梯度查询相对于真实数据的敏感度，相同程度的噪声就足以保障隐私。

B.2 大批量大小超参数搜索

我们扩大批量大小，考虑 $B \in \{64, 128, 512, 2048\}$，并搜索最优噪声水平 σ和$n_{\mathcal{D}}$。对于以 $\varepsilon = 10$ 为目标的 $B = 128$，我们在三个噪声水平 $\Sigma^{10}_{128} = \{0.6, 1.0, 1.4\}$ 中进行搜索。其他批量大小的候选噪声水平选择如下：当考虑批量大小 $B = 128n$ 时，我们搜索

$$\Sigma^{10}_{128n} := \{\sqrt{n} \cdot \sigma : \sigma \in \Sigma^{10}_{128}\}.$$

我们也因此，目标是高隐私（$\varepsilon = 1$）区间。对于 $\varepsilon = 1$，我们将所有噪声水平乘以 5，

$$\Sigma^{1}_{B} = \{5\sigma : \sigma \in \Sigma^{10}_{B}\}.$$

对于 $(B,\ \sigma)$ 的每种设置，我们都会对 $n_{\mathcal{D}} \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ 进行网格搜索。由于计算资源有限，我们省略了一些确信会失败的取值（例如，当 $n_{\mathcal{D}} = 5$ 发生模式崩溃时尝试 $n_{\mathcal{D}} = 1$ ）。

C CelebA 实现细节

CelebA 数据集（Liu 等人，2015年）由 202，599 178×218 张名人面孔的 RGB 图像组成，每张图像都标注了 40 个二元属性。我们使用的数据集版本 32×32 CelebA-Gender（发表于 Cao 等人 （2021年）提出的基准），通过将图像缩放至 32×32 并标注性别属性得到。202,599 张图像被划分为包含 182,637 张图像的训练集和包含 19,962 张图像的测试集。

对于 CelebA，我们基本沿用用于 MNIST 和 FashionMNIST 的模型架构：用于两者的带标签嵌入层的 4 层全卷积网络 $\mathcal{D}$ 和 $\mathcal{G}$。我们调整卷积 / 转置卷积的步幅和卷积核大小，以便处理 $3 \times 32 \times 32$ 图像而无需缩放。CelebA 的 $\mathcal{D}$ 和 $\mathcal{G}$ 稍大，分别有 2.64M 和 3.16M 个可训练参数。

根据我们在 MNIST 和 FashionMNIST 上的实验结果，我们采用了较大的批量大小（$B = 2048$）和自适应判别器更新，阈值为 $d = 0.6$。我们对噪声水平 $\sigma \in \{2, 3, 4\}$ 尝试了几种设置。效果最好的是最大的噪声 $\sigma = 4$ ，在目标为 $\varepsilon = 10$ 时产生了 385K 个判别器步数。

D 消融实验

D.1 改变判别器大小

我们按照第 4 节中的设置在 MNIST 上训练 DPGAN：噪声水平为 $\sigma = 1$，批量大小为 $B = 128$，目标为 $\varepsilon = 10$ ，产生了 450K 个判别器步数。通过调整 $d_{\mathcal{D}}$ （判别器第一个卷积层中的滤波器组数量，该参数控制整个网络的宽度），我们可以得到参数数量约为 0.25×、0.5× 和 2× 的判别器（表 4）。在这些实验中，我们改变判别器大小，同时保持生成器大小（2.27M 个参数）不变。

$d_{\mathcal{D}}$	$\mathcal{D}$ 参数数量	比率
64	0.44M	0.26×
96	0.97M	0.57×
128	1.72M	1×
196	3.86M	2.24×

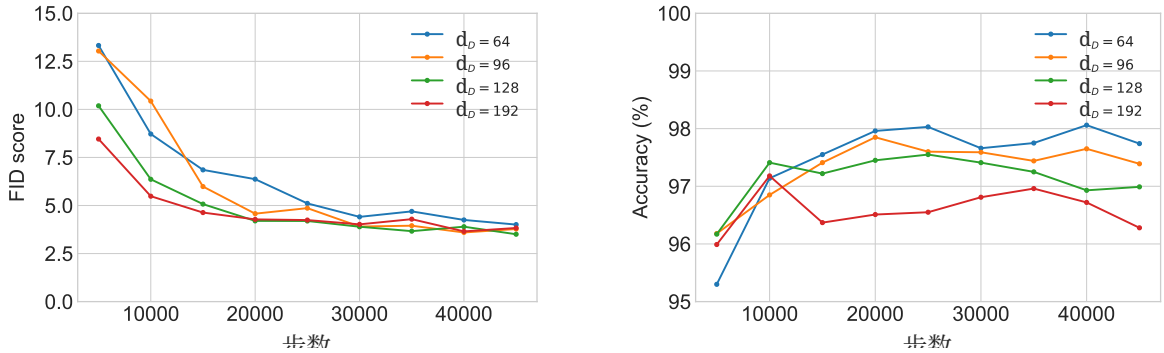
表4：不同判别器大小变体中的可训练参数数量。

结果。在图14中，我们绘制了非隐私训练期间，使用不同大小判别器时，FID 和生成的 MNIST 样本的下游分类器准确率的变化情况。我们观察到，在非隐私训练中，较大的判别器在训练早期的 FID 表现更好，但最终收敛到的准确率略低。

在图15 和 16中，我们分别绘制了以不同判别器更新频率 $n_{\mathcal{D}}$训练的 DPGAN 在 MNIST 上（目标为 $\varepsilon = 10$）时，FID 和准确率的变化情况。在每张图中，我们将 $d_{\mathcal{D}} = 128$ 次运行（绿色）的结果与图 1a 和 2a 中的结果进行比较，并与使用可训练参数数量为其 0.26–2.24× 倍的判别器进行训练的结果进行比较。这些额外设置的表现大多与 $d_{\mathcal{D}} = 128$ 次运行相近。较大的判别器似乎表现略好，尤其是在训练早期的准确率方面。较大的判别器也会显著消耗更多计算资源。

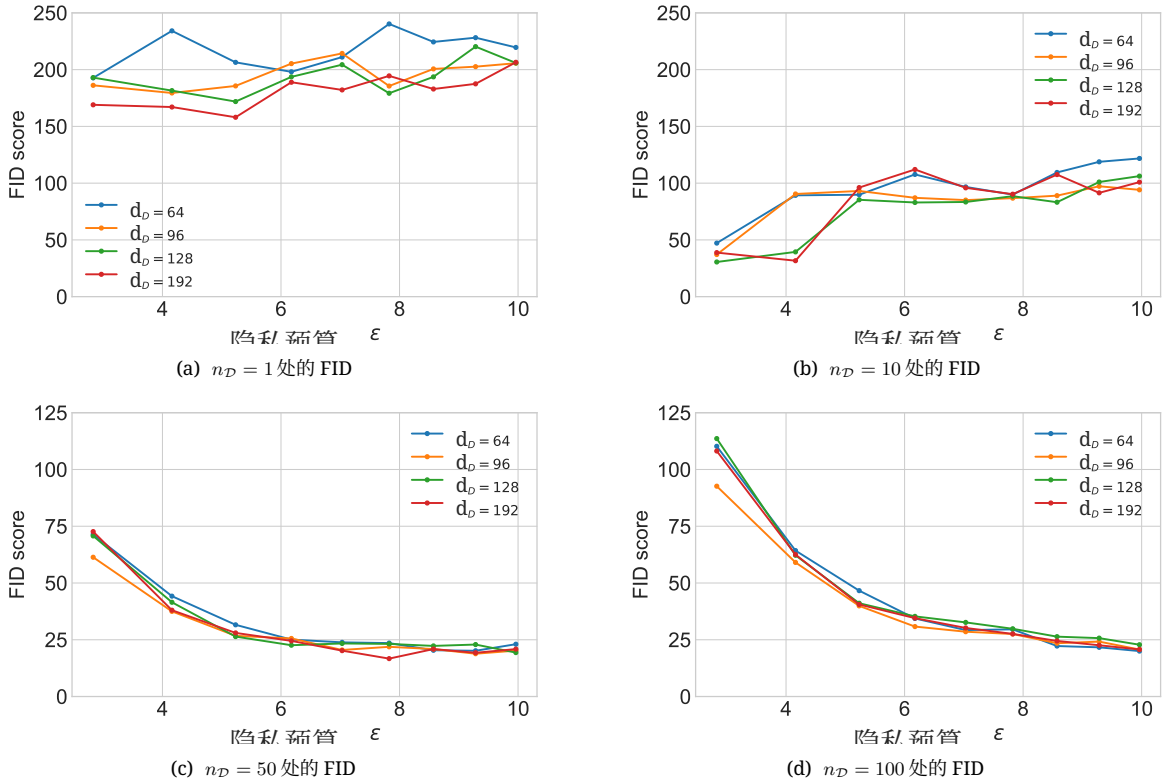
D.2 变化学习率

我们按照第4节的设置，在 MNIST 上训练差分隐私生成对抗网络：使用噪声水平 $\sigma = 1$、批量大小 $B = 128$，并以 $\varepsilon = 10$ 为目标，这对应 450K 个判别器步数。此处，我们将判别器大小 $d_{\mathcal{D}} = 128$



(a) 非隐私 FID　　(b) 非隐私准确率

图14： 不同判别器规模下非隐私 GAN 训练的 MNIST FID 和下游分类器准确率。 绿色 ($d_{\mathcal{D}} = 128$) 曲线对应于以往实验中使用的 1.72M 参数判别器。



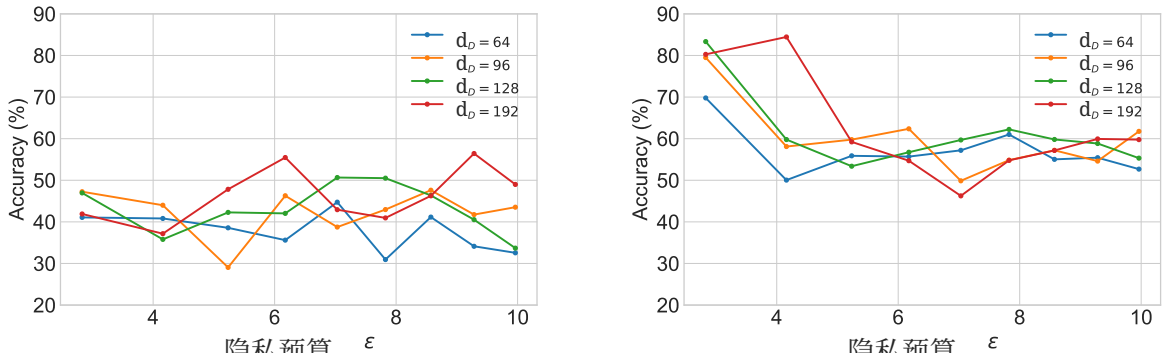
(a) $n_{\mathcal{D}} = 1$ 处的 FID　　(b) $n_{\mathcal{D}} = 10$ 处的 FID

(c) $n_{\mathcal{D}} = 50$ 处的 FID　　(d) $n_{\mathcal{D}} = 100$ 处的 FID

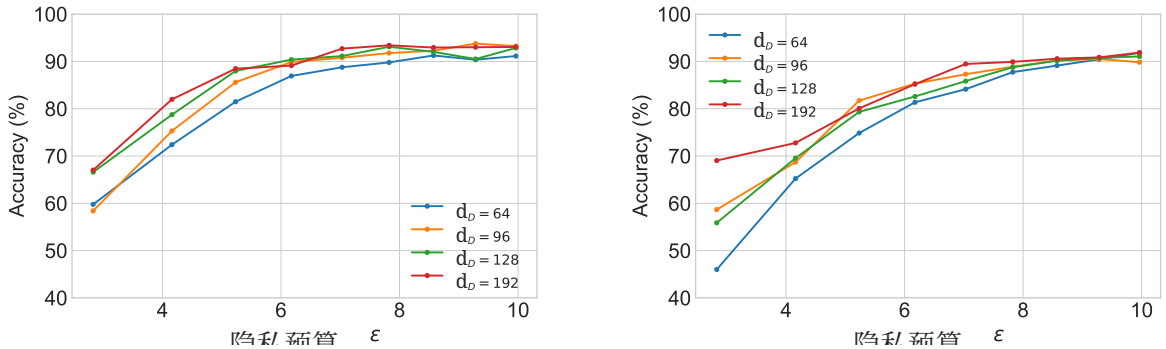
图 15： 不同判别器更新频率 $n_{\mathcal{D}}$下，DPGAN 训练的 MNIST FID（目标值为 $\varepsilon = 10$）。每张图均展示了采用不同判别器规模进行训练的结果。 绿色 ($d_{\mathcal{D}} = 128$) 曲线对应图 1a中的结果。可训练参数数量为原始设置的 0.26–2 倍的判别器，也能跟踪原始 $d_{\mathcal{D}} = 128$ 设置的结果。

固定，并调整 $\mathcal{G}$ 和 $\mathcal{D}$的学习率，同时对 β_1 和 β_2 对应的 Adam 参数$\mathcal{G}$ 和 $\mathcal{D}$ 保持不变。表5 列出了我们考虑的学习率设置。

结果。 在图 17 中，我们绘制了在不同学习率设置下进行非隐私训练时，生成的MNIST样本的FID和下游分类器准确率随训练进程的变化。我们观察到，在训练即将结束时， $5\times$ ($\mathcal{D}$) 学习率设置下的FID和准确率有所下降。$0.2\times$ 学习率设置收敛



(a) $n_{\mathcal{D}} = 1$ 时的准确率　　(b) $n_{\mathcal{D}} = 10$ 时的准确率



(c) $n_{\mathcal{D}} = 50$ 时的准确率　　(d) $n_{\mathcal{D}} = 100$ 时的准确率

图16： 在各种判别器更新频率下，针对 $\varepsilon = 10$ 的 DPGAN 训练所得到的 MNIST 下游分类器准确率 $n_{\mathcal{D}}$。在每个图中，我们展示了使用各种判别器规模进行训练的结果。绿色 （$d_{\mathcal{D}} = 128$）线对应于图 2a 中所示的结果。具有原始 24× 倍可训练参数的判别器，其结果与原始 $d_{\mathcal{D}} = 128$ 设置相符。

设置	$\mathcal{G}$ LR	$\mathcal{D}$ LR
Base	0.0002	0.0002
5× 学习率	0.001	0.001
0.2× 学习率	0.00004	0.00004
5× $\mathcal{D}$ 学习率	0.0002	0.001
0.2× $\mathcal{D}$ 学习率	0.0002	0.00004

表 5： 学习率设置。

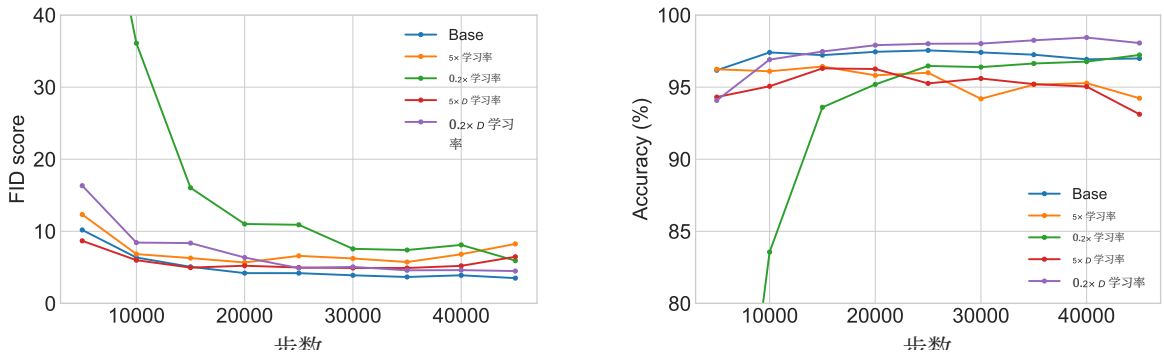
速度慢得多。调整仅 $\mathcal{D}$ 学习率 0.2×，并保持 $\mathcal{G}$ 学习率不变后，这一问题得到了解决（比较图 17 中的绿色曲线与紫色曲线）。

图 18 和图 19 考察了同时调整 $\mathcal{G}$ 和 $\mathcal{D}$ 学习率的情况，分别调整为 5× 和 0.2×。总体而言，我们观察到与第4节相同的现象：随着我们增大 $n_{\mathcal{D}} >> 1$，FID 和下游分类准确率显著提升，直到 $n_{\mathcal{D}}$ 过高，导致生成器无法采取足够的步数来收敛。不过，我们注意到一些差异：（1） $n_{\mathcal{D}}$ 最佳设置的性能全面下降（在 0.2× 学习率设置下的准确率尤为明显；见图 19b）；（2）取得最佳性能的 $n_{\mathcal{D}}$ 有所不同——在第4节中， $n_{\mathcal{D}} = 50$在 $\varepsilon = 10$ 时使 MNIST 取得最佳结果；而对于 5× 学习率， $n_{\mathcal{D}} = 200$ 表现最佳，$n_{\mathcal{D}} = 100$ 则在 0.2× 学习率下表现最佳。请注意，在附录 D.1 中改变判别器大小 $d_{\mathcal{D}}$ 的实验并未观察到这两种差异：具有不同 $d_{\mathcal{D}}$ 的所有运行结果都与 $d_{\mathcal{D}} = 128$ 运行结果十分接近。

在图 20和 21中，我们考察仅调整 $\mathcal{D}$ 学习率的情况，分别调整为 5× 和 0。2× ，并将 $\mathcal{G}$ 学习率固定为 0。0002。同样，我们观察到，随着 $n_{\mathcal{D}} >> 1$增加，效用显著提升，直至

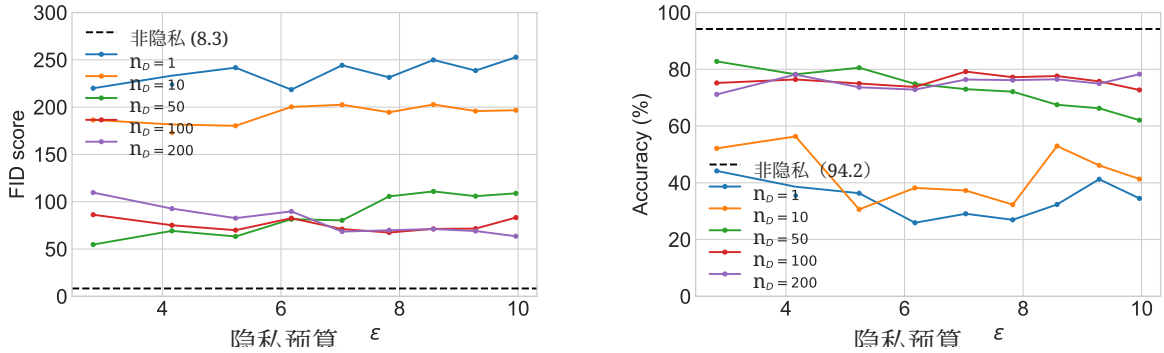
直到 $n_{\mathcal{D}}$ 过高。我们注意到，当将 $\mathcal{G}$ 学习率固定时，0.2× 的设置所带来的改进幅度更接近基准学习率设置下改变 $n_{\mathcal{D}}$ 所观察到的改进幅度。

总之：在其他超参数保持不变的情况下调整学习率，仍然能体现增大 $n_{\mathcal{D}}$ 的益处；但与基准设置相比，无法重现：（1）改进的幅度；（2）该现象的具体表现，即相同的最优 $n_{\mathcal{D}}$。我们暂不深入探讨该现象在不同学习率下究竟如何变化：进一步研究 Adam 的动量参数（β_1，β_2）和差分隐私随机梯度下降的噪声水平 σ 对结果的影响或许会有所帮助；也可以考察较大的 $\mathcal{D}$ LR 对非隐私 GAN 结果造成的性能下降。



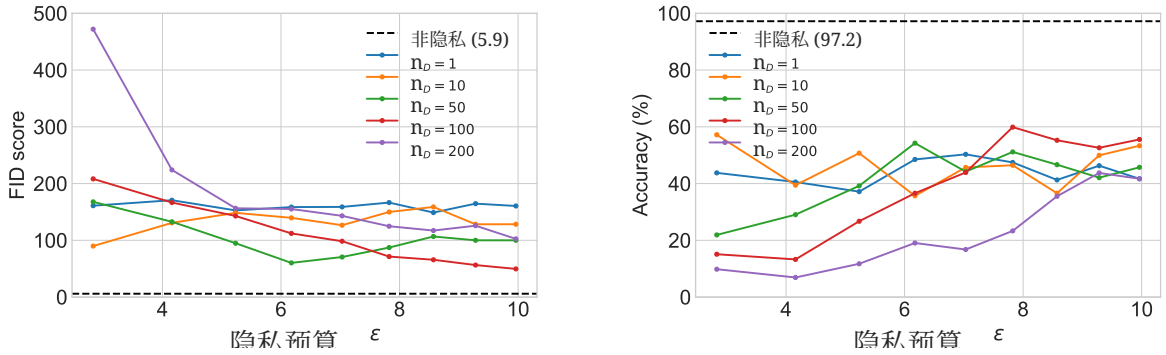
(a) 非隐私 FID　　(b) 非隐私准确率

图 17：在不同学习率设置下进行非隐私 GAN 训练时的 MNIST FID 和下游分类器准确率。蓝色 ($d_{\mathcal{D}} = 128$) 线对应于先前实验中使用的基准学习率设置。



(a) 5× 学习率下的 FID　　(b) 5×学习率下的准确率

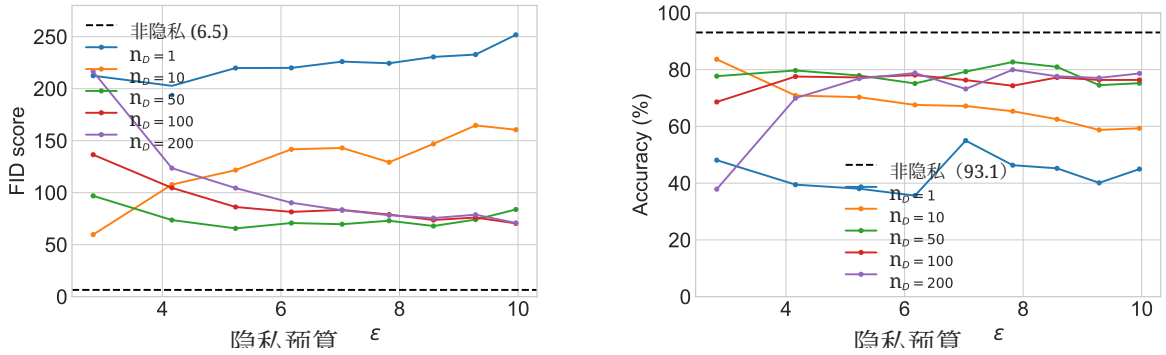
图 18：在各种 $n_{\mathcal{D}}$设置下，以 $\varepsilon = 10$ 为目标、对 $\mathcal{G}$ 和 $\mathcal{D}$均使用5× 基准学习率的 DPGAN 训练运行的 MNIST FID 和下游分类器准确率。



(a) 学习率为 0.2× 时的 FID

(b) 学习率为 0.2× 时的准确率

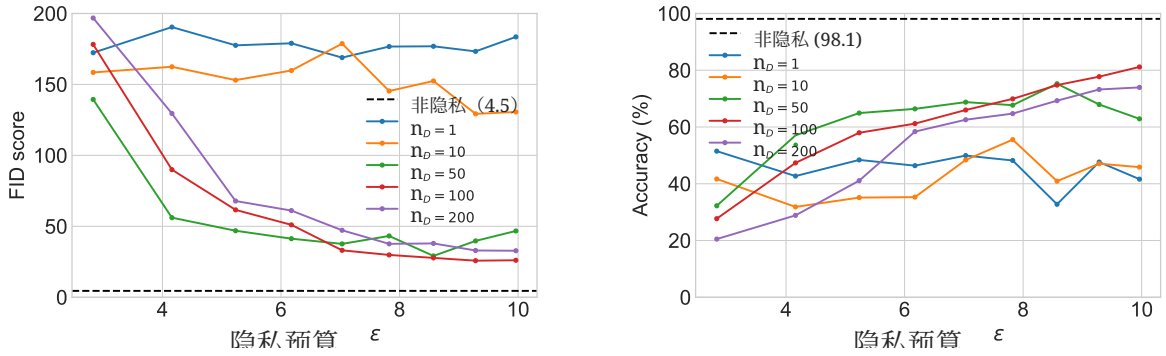
图 19：在各种 $n_{\mathcal{D}}$ 设置下，以 $\varepsilon = 10$ 为目标、$\mathcal{G}$ 和 $\mathcal{D}$ 均采用基准学习率的 0.2× 进行 DPGAN 训练运行所得到的 MNIST FID 和下游分类器准确率。



(a) 学习率为 5× $\mathcal{D}$ 时的 FID

(b) 学习率为 5× $\mathcal{D}$ 时的准确率

图 20：在各种 $n_{\mathcal{D}}$ 设置下，以 $\varepsilon = 10$ 为目标、仅 $\mathcal{D}$ 采用 5× 基准学习率（$\mathcal{G}$ 学习率保持不变）的 DPGAN 训练运行所得到的 MNIST FID 和下游分类器准确率。



(a) 0.2× $\mathcal{D}$ 学习率下的 FID

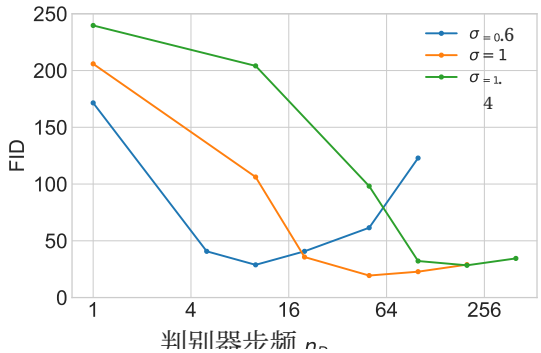
(b) 准确率在 0.2× $\mathcal{D}$ 学习率时

图 21：针对 $\varepsilon = 10$、仅使用 0.2× 倍基准学习率（$\mathcal{G}$ 学习率保持不变），并采用不同 $n_{\mathcal{D}}$ 设置的 DPGAN 训练运行的 MNIST FID 和下游分类器准确率。

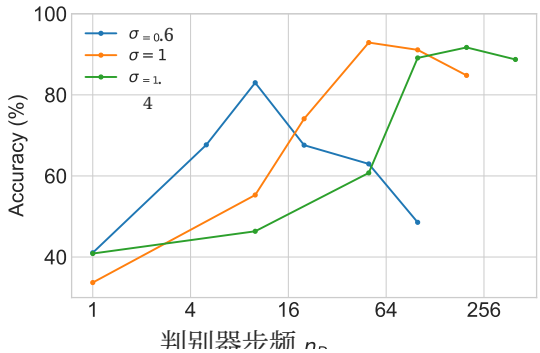
D.3 改变批量大小和噪声水平

固定批量大小 B 和噪声水平 σ 后，可在隐私预算 ε允许的范围内确定判别器步数总预算 T。例如，第 4 节及附录 D.1 和 D.2 中的结果使用 $B = 128$ 和 $\sigma = 1$，在目标隐私预算为 $T = 450$ 时，可获得 $\varepsilon = 10$ K。再次以 MNIST 上的 $\varepsilon = 10$ 为目标，我们取

各种（B， σ）组合，并在一系列 $n_{\mathcal{D}}$范围内，绘制在此设置下训练的差分隐私生成对抗网络的最终 FID 和准确率。结果见图 22、23 和24。

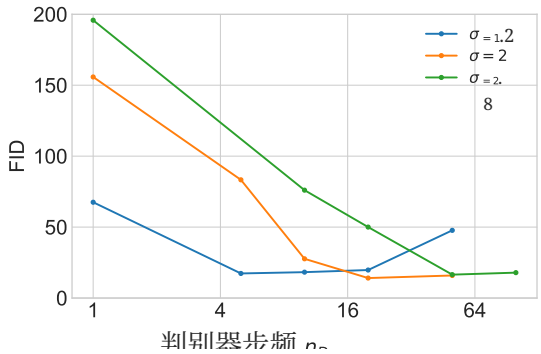


(a) $B = 128$处的 FID

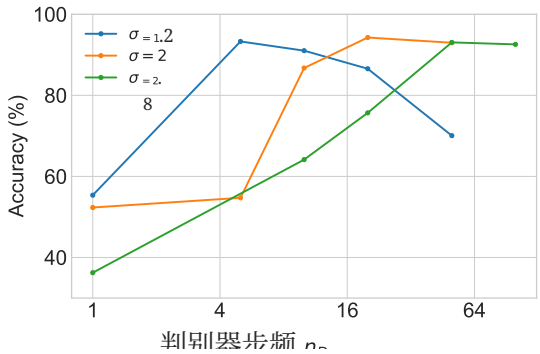


(b) $B = 128$处的准确率

图 22： 针对 $\varepsilon = 10$ 的 $B = 128$ 次运行在 3 个噪声水平下的 MNIST FID 和下游分类器准确率，使用 $\sigma \in \{0.6, 1, 1.4\}$。我们报告了 $n_{\mathcal{D}}$ 范围内的最终效用。横轴采用对数刻度。

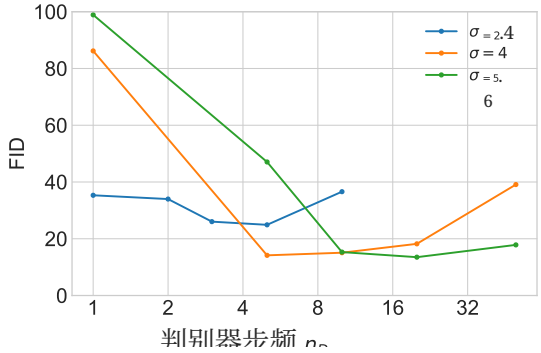


(a) $B = 512$处的 FID

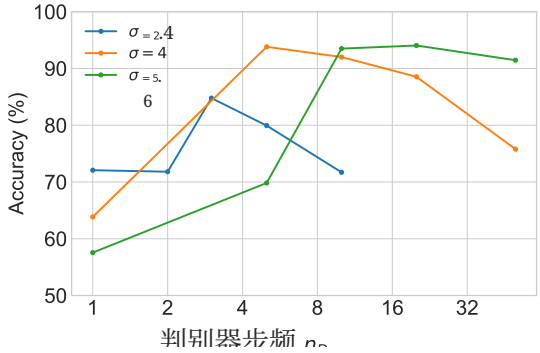


(b) $B = 512$处的准确率

图 23： 针对 $\varepsilon = 10$ 的 $B = 512$ 次运行在 3 个噪声水平下的 MNIST FID 和下游分类器准确率，使用 $\sigma \in \{1.2, 2, 2.8\}$。我们报告了 $n_{\mathcal{D}}$ 范围内的最终效用。横轴采用对数刻度。



(a) $B = 2048$处的 FID



(b) $B = 2048$处的准确率

图24： MNIST FI 在针对 $\varepsilon = 10$ 的 $B = 2048$ 次运行中，MNIST FID 和下游分类器准确率，噪声水平为 $\sigma \in \{2.4, 4, 5.6\}$ 。我们报告最终效用 在 3 种噪声水平下， $n_{\mathcal{D}}$ 范围内的结果。横轴采用对数刻度。

结果。 在所有批量大小和噪声水平下，我们都观察到了第 4 节所描述的相同 U形效用曲线。该曲线预测，对于（σ， B）的任意固定设置，都存在一个最优的 $n_{\mathcal{D}}$ 。对于固定的 B，较小 σ对应的最优 $n_{\mathcal{D}}$ 更低。我们还发现，对于 σ 较低且 B较大的设置，最优 $n_{\mathcal{D}}$ 可能会非常低。对于所有批量大小，选择在相当大的值（$>> 1$）下达到其最优 $n_{\mathcal{D}}$ 的噪声水平，往往优于较小的、较早达到其最优 $n_{\mathcal{D}}$ 的噪声水平。

E 更多结果

E.1 墙钟时间

我们报告了在各种超参数设置下训练运行的实际耗时，这些运行在 1×搭载 NVIDIA A40 显卡、运行 PyTorch 1.11 的环境中执行。0+CUDA 11.3.1 和 Opacus 1.1.3。表6 展示了 MNIST 上的结果，尤其比较了 $n_{\mathcal{D}}$ 对训练时间的影响。判别器步数总量 T由隐私预算和 DPSGD超参数决定。因此，增加 $n_{\mathcal{D}}$ 会减少总 $\mathcal{G}$ 步数，并缩短训练时间。表7展示了针对各种数据集，在自适应判别器步频下的训练时间。

所有隐私设置都比非隐私训练慢得多。事实上，最好的差分隐私结果往往来自长时间使用较高噪声水平进行训练，以计算成本换取效用（De 等人，2022年）。例如，最好的差分隐私扩散模型（Dockhorn 等人，2022年）使用 8 个 V100 训练 1 天，得到其最佳 MNIST 模型。尽管两者并不能直接比较，但我们注意到，我们最好的 $\varepsilon = 10$ 结果在 1 个 A40 上训练 7.5 小时即可获得。

隐私	B	σ	T	$n_{\mathcal{D}}$	FID	实际耗时
$\varepsilon = \infty$	128	-	45K	1	3.4 ± 0.1	44m
$\varepsilon = 10$	128	1	450K	1	205.3 ± 0.9	11小时03分钟
				10	103.4 ± 5.8	6小时33分钟
				50	18.5 ± 0.9	5小时56分钟
				100	21.0 ± 1.6	5小时57分钟
				200	26.6 ± 2.2	5小时54分钟
	2048	5.6	98K	20	13.2 ± 1.0	16小时54分钟
$\varepsilon = 1$	128	5	325K	200	111.1 ± 17.9	4小时40分钟
	512	14	165K	50	106.2 ± 64.0	7小时31分钟

表6： 在不同设置下，MNIST 上的墙钟时间。隐私级别 ε、批量大小 B 和噪声水平 σ 决定了训练期间采取的总步数 $\mathcal{D}$， T。给定 T，判别器更新频率 $n_{\mathcal{D}}$ 决定了训练期间采取的 $\mathcal{G}$ 步数。

隐私	数据集	B	σ	T	FID	实际耗时
$\varepsilon = \infty$	MNIST	128	-	45K	3.4 ± 0.1	44m
	FashionMNIST				16.5 ± 1.7	42m
	CelebA				30.0 ± 1.6	47m
$\varepsilon = 10$	MNIST	512	2	174K	12.8 ± 0.3	7小时35分钟
	FashionMNIST				62.3 ± 8.7	7小时35分钟
	CelebA	2048	4	385K	170.8 ± 20.3	3天17小时49分钟
$\varepsilon = 1$	MNIST	512	14	165K	52.6 ± 3.2	7小时43分钟
	FashionMNIST				126.4 ± 4.1	7小时26分钟

表 7：使用自适应判别器步频进行训练时，在 MNIST、FashionMNIST 和 32×32 CelebA 上的墙钟时间（ $\varepsilon = \infty$ 结果除外，该结果使用 $n_{\mathcal{D}} = 1$）。

E.2 提高判别器学习率，而非步频

在第 4 节针对 MNIST 的实验设置中，$\varepsilon = 10$，我们考虑 $n_{\mathcal{D}} = 1$ 设置，并提高判别器学习率，而不是 $n_{\mathcal{D}}$。$\mathcal{G}$ 学习率固定为 0。0002。结果见表 8。我们没有观察到通过提高 $n_{\mathcal{D}}$ 并保持 $\mathcal{D}$ 学习率 /$\mathcal{G}$ 学习率为 $1\times$ 所获得的同等程度的改进。

$\mathcal{D}$学习率/$\mathcal{G}$ 学习率	FID	准确率（%）
$1\times$	205.9	33.7
$5\times$	251.8	45.0
$10\times$	228.1	37.5
$50\times$	237.2	54.5

表8： 在 $\varepsilon = 10$ 下提高 $\mathcal{D}$ 学习率并保持 $n_{\mathcal{D}} = 1$时，MNIST 上的结果。作为参考，在 $1\times$ 使用 $n_{\mathcal{D}} = 50$ 并保持 $\mathcal{D}$ 学习率/$\mathcal{G}$ 时，FID 分数为 18.5 ± 0.9，准确率为 93.0 ± 0.6%。

F 补充讨论

差分隐私表格数据合成。 我们的研究聚焦于图像数据集，而许多重要的私有数据生成应用涉及表格数据。在这些场景中，基于边际分布的方法（Hardt 等人，2012年；Zhang 等人，2017年；McKenna 等人，2019年）表现最佳。尽管Tao 等人（2021年）发现，基于私有生成对抗网络的方法在这些场景中甚至无法保留基本统计量，但我们推测，我们的技术可能会带来改进。

采取多个判别器步骤。 Goodfellow 等人提出的原始生成对抗网络公式(2014年) 将$n_{\mathcal{D}}$（即在生成器步骤之间执行的判别器步数）作为可调超参数。在 WGAN 框架（Arjovsky 等人，2017年）中，建议在生成器步骤之间尽可能充分地训练判别器，即训练至最优状态，以获得最佳性能。实际上，WGAN 实现采用$n_{\mathcal{D}} = 5$ 以节省计算开销。一些研究通过实证探索了采取多个判别器步骤的影响（Miyato 等人，2018年；Brock 等人，2019年），发现寻找最优的 $n_{\mathcal{D}}$ 可以改善结果。类似的不平衡设置，例如前瞻法和不平衡学习率，也已得到理论分析（Chavdarova 等人，2021年；Fiez 和 Ratliff，2021年）。

在隐私设置下，应用此类策略来改进 DPGAN 训练的研究相对较少。DPGAN 训练方案在很大程度上是对非隐私方法的移植——继承了许多为实现高性能非隐私训练而设计的参数选择，而这些选择在 DP 下并非最优。

非隐私设置下的指导建议（Chintala 等人 （2016年）中的提示14）建议在存在噪声时增加判别器的训练步数（这是一种用于非隐私 GAN 的正则化方法）。对于差分隐私来说，情况正是如此；这也是我们的核心策略，带来了最显著的效用提升。我们发现这一现象时并不知道这条建议，但它为我们的发现提供了佐证。尽管Chintala 等人（2016年）对此着墨不多，但进一步探索这一原则（以及其他策略）在非隐私设置下的应用，或许能为改进 DPGAN 提供指导。例如，Chavdarova 等人（2021年）提出了一种前瞻更新规则，可在存在噪声时实现快速收敛，而无需使用大批次——此类技术或许也有助于隐私设置。

差分隐私机器学习中的超参数调优。 超参数调优对于实现深度学习至关重要。在隐私保护场景下也是如此，但还需考虑两个额外问题：(1) 调优并非没有代价——根据朴素组合，隐私损失会随着运行次数增加而扩大；(2) DPSGD 会改变超参数景观——引入额外的超参数，同时还会 改变现有超参数的相对重要性。(1) 已由 Liu 和 Talwar（2019年）加以解决（不过，在自适应选择很重要的场景中，组合具有竞争力（Mohapatra 等人，2022年））。关于 (2)，Ponomareva 等人 （2023年）提供了关于差分隐私超参数调优的全面指南；对于生成对抗网络训练，我们的工作确定了一个在差分隐私设置下影响尤为显著的重要参数。

为与先前工作进行比较，我们报告了最优的超参数设置。事实上，在实际部署场景中，将搜索成本计算在内时，引入高度依赖数据集的参数可能会导致整体性能更差。我们的自适应判别器更新频率正是出于这一考虑，而直接将 MNIST 的超参数用于 FashionMNIST，则是一项用于检验稳健性的合理性检查。

本文工作中揭示判别器更新频率对 privateGAN 训练至关重要的这一方面，不受隐私超参数搜索相关问题的影响。

在比较算法时，将搜索成本纳入考量的评估方法是一个重要的研究方向，但本文工作并未探讨这一点。对于基准数据集而言，问题更为复杂，因为各种算法的设计选择和“默认”超参数范围中都蕴含着隐含知识。