arXiv:2302.00628v2 [cs.LG] 2 Feb 2023

Trai使用精确率-召回率散度训练归一化流

Alexandre Verine [1] **Benjamin Negrevergne** [1] **Muni Sreenivas Pydi** [1] **Yann Chevaleyr e** [1]

摘要

生成模型可能会出现不同类型的失效，例如模式丢失和样本质量低下，而单一标量指标无法捕捉这些问题。为了解决这一问题，近期研究提出使用精确率和召回率来评估生成模型，其中精确率衡量样本质量，召回率衡量对目标分布的覆盖程度。尽管人们使用各种目标分布与估计分布之间的差异度量来训练生成模型，但不同差异度量能够实现怎样的精确率-召回率权衡尚不明确。本文表明，实现指定的精确率-召回率权衡，相当于最小化某个族中的 f-散度，我们将其称为 *PR* 散度。反过来，任何f-散度都可以写成 PR 散度的线性组合，因此对应于最小化加权的精确率-召回率权衡。此外，我们提出了一种新型生成模型，能够训练归一化流，使其最小化任意 f-散度，尤其能够实现给定的精确率-召回率权衡。

1. 引言

生成模型已成为机器学习中的强大工具。近年来，人们提出了大量生成模型，例如生成对抗网络（GANs）和归一化流，它们在学习图像和文本数据中的复杂高维分布方面表现出色。尽管取得了这些成功，生成模型的评估仍然是一个棘手的问题。基于生成对抗网络的方法通常能够生成高保真图像，但会遭遇众数丢失（无法生成来自真实分布所有众数的数据）以及似然较差的问题。基于归一化流的模型经过显式训练，旨在实现较高的似然，但往往生成质量较低的样本。为应对这些复杂问题，近期研究提出了一种双重评估方法，同时考量精确率和召回率（Sajjadi等人，2018年；Simon 等人，2019）。

给定一个目标分布 P 以及一组参数化分布 $\{P_\theta|\theta \in \Theta\}$，生成模型会尝试找到^最佳拟合 $P = P_{\theta^*}$，使分布 P与 P 之间的散度或距离度量^最小。通常，归一化流通过最小化 P与 P之间的 Kullback–Leibler 散度（记作 $\mathcal{D}_{\mathrm{KL}}$，^）进行训练。生成对抗网络中的生成器可以使用多种散度/距离进行训练，例如任意一种 f-散度（记作 D_f）（Nowozin 等人，2016），或 Wasserstein 距离（Arjovsky 等人，2017）。众所周知，优化（前向） $\mathcal{D}_{\mathrm{KL}}$ 倾向于产生质量覆盖型模型（Minka，2005），这与我们在大多数其他生成模型中观察到的模式寻求行为形成对比。如图2所示，这会产生召回率较高（样本更多样化）、但精确率较低（离群点更多）的模型。然而，针对一般散度进行优化时隐含着怎样的权衡，目前尚不清楚。这引出了以下问题。

问题1。 最小化^化 $D_f(P\|P)$ 对应于怎样的精确率-召回率权衡?

直观而言，生成模型的精确率衡量其生成样本的质量，而召回率衡量其对目标分布的覆盖程度。根据应用场景的不同，可能需要训练具有高精确率的生成模型（例如用于图像合成和数据增强），或具有高召回率的生成模型（例如用于密度估计和去噪）。然而，现有的生成建模方法无法提供这种灵活性。尽管已有一些研究致力于高效评估现有模型的精确率-召回率曲线，但针对训练能够实现特定精确率-召回率权衡的生成模型的方法，似乎仍鲜有研究。这引出了以下问题。

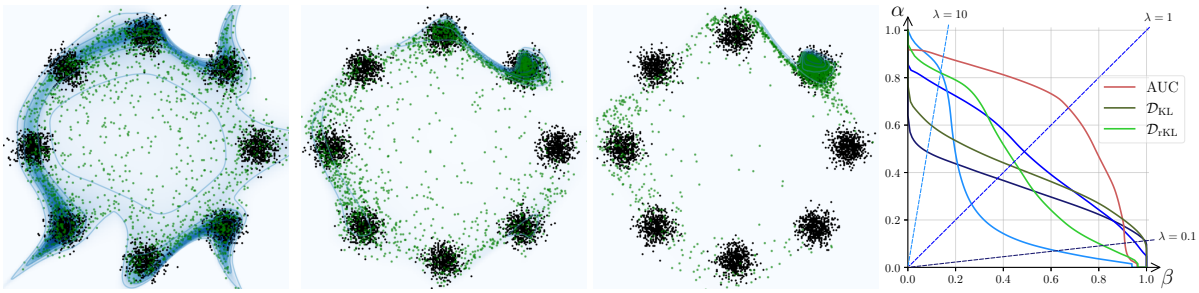
问题 2。 是否可以训练出一个能够实现精确率和召回率之间特定权衡的生成模型?

在本文中，我们通过以下贡献来回答问题1 和 2 。

- 我们证明，实现指定的精确率-召回率权衡等价于最小化某种 f-

[1]LAMSADE，法国国家科学研究中心，巴黎-多芬-PSL大学，法国巴黎。通信作者：Alexandre Verine <alexandre.verine@dauphine.psl.eu>。

初步工作。



(a) $\lambda = 0.1$ (b) $\lambda = 1$ (c) $\lambda = 10$ (d) 精确率-召回率曲线

图 *1*。在从 8 个高斯分布中采样的二维数据点上训练的归一化流（RealNVP）。图 1(a) 至图 1(c)：三个模型采用不同的精确率/召回率权衡进行训练，以最小化精确率-召回率散度。真实$\widehat{\ }$分布 P的样本以黑色表示，训练后模型 P的样本以绿色表示，P的对数似然以蓝色等高线表示（颜色越深，密度越高）。(a) $\lambda = 0.1$，更偏重召回率而非精确率。(b) $\lambda = 1$，精确率与召回率之间的权衡较为均衡。(c) $\lambda = 10$，更偏重精确率而非召回率。图 1(d)：使用不同散度训练的模型的精确率-召回率曲线。蓝色曲线表示图 1 (a) 至 (c) 中的三个模型。深绿色和浅绿色曲线分别表示使用 $\mathcal{D}_{\mathrm{KL}}$ 散度训练的模型，以及使用 $\mathcal{D}_{\mathrm{rKL}}$训练的模型。红色曲线表示使用曲线下面积（AUC）训练的模型。（另见图 5(a)、图 5(b) 和图 5(c)。）

$\widehat{\ }$散度，介于 P 和 P之间。具体而言，在定理 4.3 中，我们给出一族 f-散度（记作 $\mathcal{D}_{\lambda\text{-PR}}$，$\lambda \in [0, \infty]$），它们与生成模型精确率-召回率曲线上的不同点相关联。

- 我们证明，任意 f散度都可以写成$\mathcal{D}_{\lambda\text{-PR}}$ 族中 f散度的线性组合。该结果明确揭示了生成模型在最小化任意 f散度时所隐含的精确率-召回率权衡。

- 我们提出了一种新颖的基于流的生成模型，对于任意给定的 λ，通过最小化$\mathcal{D}_{\lambda\text{-PR}}$，即可实现用户指定的精确率与召回率之间的权衡。

图 1展示了我们的模型在不同 λ设置下的表现。当 λ较高时，我们可以训练模型优先考虑精确率而非召回率，反之亦然。

由于使用 $\mathcal{D}_{\lambda\text{-PR}}$ 训练的判别器在远离原点的区域大多是平坦的，因此直接训练f-GAN来最小化 $\mathcal{D}_{\lambda\text{-PR}}$ 并不可行，否则会遭遇梯度消失问题。我们通过对抗式地联合训练归一化流和判别器来解决这一问题。与现有工作相比，我们的模型具有以下创新之处。

- 我们模型的归一化流最小化的是 $\mathcal{D}_{\lambda\text{-PR}}$，而不是通常使用的 $\mathcal{D}_{\mathrm{KL}}$ 散度。事实上，使用我们的方法，可以训练归一化流来最小化任意 f-散度。

- 为了训练我们模型的判别器，我们使用一种不同于$\mathcal{D}_{\lambda\text{-PR}}$ 的散度 $\mathcal{D}_f$，以实现稳定训练并避免梯度消失。这使我们的训练方式区别于生成对抗网络和 flow-GAN 使用不同的生成器与判别器损失函数，因此我们的训练方式与它们的极大极小训练不同。训练中使用的两种散度满足兼容性的条件见定理5.3。

符号说明：我们用 $\mathcal{P}(\mathcal{X})$ 表示空间 $\mathcal{X}$ 上概率测度的集合。我们用大写字母表示概率测度，用相应的小写字母表示其密度函数。全文中，我们用 $\mathcal{X} \subseteq \mathbb{R}^d$ 表示输入空间，用 $P \in \mathcal{P}(\mathcal{X})$ 表示$\widehat{\ }$我们希望建模的目标分布，用 $P \in \mathcal{P}(\mathcal{X})$ 表示生成模型估计的分布。此外，$\widehat{\ }$我们假设 P和 P在 $\mathcal{X}$中具有相同的支撑集。

2. 相关工作

以往已有一些工作致力于改进归一化流，解决其自然的质量覆盖效应与 $\mathcal{D}_{\mathrm{KL}}$所期望的模式寻求结果之间的权衡。可以使用拒绝采样人为地提高模型的精确率（Stimper 等人，2022年；Issenhuth 等人，2022；Tanielian 等人，2020年）。不过，Midgley 等人Midgley 等人 (2022b)仅关注损失函数，训练归一化流以最小化一种 α-散度，针对 α =2进行估计，并采用退火重要性采样。尽管该工作展示了使用另一种 f-散度进行训练的益处，但它通过固定 α来确定权衡。此外，他们的方法要求已知目标分布的密度P，因为其目标是将流训练成采样器。

为了将这种方法推广到在无监督设定下使用任意f-散度训练流，我们的方法借鉴了Nowozin 等人(2016年) 关于生成对抗网络的研究。Grover 等人Grover 等人(2018年) 已将这一框架应用于归一化流。研究表明，流

可以通过判别器进行训练，而且对抗训练的模型能从对数似然项中获益。然而，作者仅将评估范围限定在 Wasserstein 距离上，并未探讨 f-散度如何影响流。

此外，我们建立了 f-散度与精确率和召回率之间的联系。已有多项研究尝试通过降维空间中的聚类方法来评估这些指标（Sajjadi等人，2018年；Tanielian 等人，2020年；Kynk¨a¨anniemi 等人，2019）。我们使用似然比估计来计算这些指标（Simon 等人，2019），但也为估计质量提供了保证。最后，虽然我们将精确率表述为一种 f-散度，Djolonga 等人 （2020年）提出了一种基于 R´enyi 散度的不同评估框架。上述研究都只关注模型评估，而我们则构建了一种基于该指标训练模型的方法。

3. 背景

3.1. 归一化流预备知识

设 $\mathcal{Z} \subseteq \mathbb{R}^d$ 为潜在空间。令 $Q \in \mathcal{P}(\mathcal{Z})$ 表示标准正态分布 $\mathcal{N}(0, \mathcal{I}_d)$。给定目标分布 $P \in \mathcal{P}(\mathcal{X})$，我们希望找到一个归一化流（Rezende和Mohamed，2016），即一个双射 $F: \mathcal{X} \to \mathcal{Z}$，使得对于任意可测集 $A \subseteq \mathcal{X}$，都有 $P(A) = Q(F^{-1}(A))$。这样，我们就可以通过将 F^{-1} 应用于来自 Q 的样本，利用 F 生成来自 P 的样本；也可以使用变量替换公式计算 $x \in \mathcal{X}$ 处的密度 $p(x)$：$p(\boldsymbol{x}) = q(F(\boldsymbol{x}))|\det \mathrm{J}\,\mathrm{ac}_F(\boldsymbol{x})|$，其中 $\det \mathrm{Jac}_F(\boldsymbol{x})$ 是 F 在 $\boldsymbol{x}$ 处的雅可比矩阵行列式。

实践中，F 未知，我们的目标是找到映射 $\frown F'$，使其能够最佳地近似 P，即得到 $P_{F'}$。为此，我们考虑一类可逆函数 $\{F_\theta : \theta \in \Theta\}$，通常使用可逆神经网络（INN）架构来表示，例如 GLOW（Kingma和Dhariwal，2018）、RealNVP（Rezende和Mohamed，2016）或 ResFlow（Behrmann等人，2019；Chen等人，2020）。随后，我们可以训练 INN，使 P 与 P_{F_θ} 之间的 KL散度最小化；等价地，也可以通过最大化对数似然 $\mathbb{E}_{\boldsymbol{x}\sim P}[\log \hat{} p_{F_\theta}(\boldsymbol{x})]$ 来进行估计，其中使用来自 P 的样本构成的数据集 D。

$$\begin{aligned}\min_\theta \mathcal{D}_{\mathrm{KL}}(P\|\widehat{P}_{F_\theta}) &= \min_\theta \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\hat{p}_{F_\theta}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= H(P) - \max_\theta \mathbb{E}_{\boldsymbol{x}\sim P}\left[\log \hat{p}_{F_\theta}(\boldsymbol{x})\right],\end{aligned} \tag{1}$$

其中，$H(P)$是 P的熵。在本文其余部分，$\widehat{}$当上下文含义明确时，我们省略 P_{F_θ} 中对 F_θ 的依赖。

为了在实践中取得良好效果，可逆神经网络必须满足若干性质（Kobyzev 等人，2020年）：前向传播和逆向传播都必须能够高效计算；雅可比矩阵的行列式必须易于计算；最后，$\{F_\theta : \theta \in \Theta\}$ 应具有足够强的表达能力来建模P。虽然已有许多架构被提出，以满足前两项要求，但由于可逆性约束，表达能力仍然是一项挑战。因此，$\widehat{}$在大多数实际场景中，$P \neq P$，用于衡量 P与 P之间距离的散度会对最终模型产生关键影响。

3.2. f-散度

沿用Nowozin 等人（2016年）的研究，我们考虑一族 f-散度，用于衡量真实分布 P 与估计分布$\widehat{}$ P之间的差异。给定一个满足条件f：$\mathbb{R}^+ \to \mathbb{R}$的凸下半连续函数 $f(1) = 0$，两个概率分布 f之间的$\widehat{}$-散度 P和 P定义如下。

$$\mathcal{D}_f(P\|\widehat{P}) = \int_{\mathcal{X}} \hat{p}(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{\hat{p}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}. \tag{2}$$

许多广为人知的统计散度，例如 Kullback-Leibler 散度（$\mathcal{D}_{\mathrm{KL}}$）、反向 Kullback-Leibler 散度（$\mathcal{D}_{\mathrm{rKL}}$）或全变差（$\mathcal{D}_{\mathrm{TV}}$），都是 f-散度（见表1）。D_f 也具有对偶变分形式（Nguyen 等人，2009年）。

$$\mathcal{D}_f(P\|\widehat{P}) = \sup_{T\in\mathcal{T}} \left(\mathbb{E}_P\left[T(\boldsymbol{x})\right] - \mathbb{E}_{\widehat{P}}\left[f^*(T(\boldsymbol{x}))\right]\right), \tag{3}$$

其中，$\mathcal{T}$是可测函数的集合 $\mathcal{X} \to \mathbb{R}$ 且 $f^*: \mathbb{R} \to \mathbb{R}$ 是 f的凸共轭（或 Fenchel 变换），定义为 $f^*(t) = \sup_{u\in\mathbb{R}}\{tu - f(u)\}$。我们用 $T_{\mathrm{opt}} \in \mathcal{T}$表示在（3）中取得上确界的函数。

定义如下：

$$\mathcal{D}^{\mathrm{dual}}_{f,T}(P\|\widehat{P}) = \mathbb{E}_{\boldsymbol{x}\sim P}\left[T(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\widehat{P}}\left[f^*(T(\boldsymbol{x}))\right], \tag{4}$$

通过求解以下问题，可以训练一个针对 P的生成模型：$\min_F \max_T \mathcal{D}^{\mathrm{dual}}_{f,T}(P\|\widehat{P})$，其中 F和 T由神经网络表示（Nowozin 等人，2016年）。

3.3. 生成模型的精确率-召回率曲线

通常，生成模型的评估采用 Inception Score（Salimans 等人，2016年）或 Fr´echet Inception Distance（Heusel 等人，2017）等单维度指标，但这些指标无法区分两种不同的失效模式：精确率较低（即无法生成高质量样本）和召回率较低（即未能覆盖 P的所有众数）。Sajjadi等人（Sajjadi等人（2018年）以及后来由Simon 等人（2019）扩展提出的以下定义解决了这个问题。

定义 3.1（PRD 集，改编自Simon 等人（2019））。$\widehat{}$对于 P，$P \in \mathcal{P}(\mathcal{X})$，精确率-召回率集$\mathrm{PRD}(P, \widehat{P})$

表 1. 常见 f-散度列表。生成函数 f 及其 Fenchel 共轭如示。为计算 $p(\boldsymbol{x})/\hat{p}(\boldsymbol{x})$，给出了最优判别器 T^{opt}。然后给出 $f''(1/\lambda)/\lambda^3$，以计算作为精确率-召回率散度组合的 f-散度。

散度	记号	$f(u)$	$f^*(t)$	$T^{\text{最优}}(\boldsymbol{x})$	$f''(1/\lambda)/\lambda^3$
KL	$\mathcal{D}_{\text{KL}}(P\|\widehat{P})$	$u\log u$	$\exp(t-1)$	$1+\log p(\boldsymbol{x})/\hat{p}(\boldsymbol{x})$	$1/\lambda^2$
反向 KL 散度	$\mathcal{D}_{\text{rKL}}(P\|\widehat{P})$	$-\log u$	$-1-\log -t$	$-\hat{p}(\boldsymbol{x})/p(\boldsymbol{x})$	$1/\lambda$
χ^2-皮尔逊	$\mathcal{D}_{\chi^2}(P\|\widehat{P})$	$(u-1)^2$	$\frac{1}{4}t^2+t$	$2\,(p\,(\boldsymbol{x})\,/\hat{p}\,(\boldsymbol{x})\,-1)$	$2/\lambda^3$
全变差	$\mathcal{D}_{\text{TV}}$	$\lvert u-1\rvert/2$	t	符号 $(p\,(\boldsymbol{x})\,/\hat{p}\,(\boldsymbol{x})\,-1)\,/2$	NA

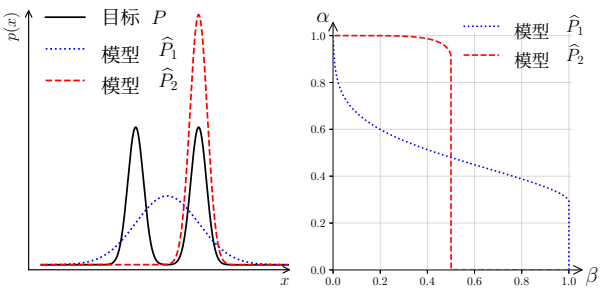
定义为精确率-召回率对 $(\alpha,\ \beta)\ \in\mathbb{R}^+\times\mathbb{R}^+$ 的集合，其中存在 $\mu\in\mathcal{P}\ (\mathcal{X})$ 使得 $P\geq_{\sim}\beta\mu$且 $P\geq\alpha\mu$。精确率-召回率曲线（或 *PR* 曲线）$\widehat{\ }$定义为 $\partial\text{PRD}(P,P)=\{(\alpha,\ \beta)\ \in\text{PRD}(P,\ P)\not\ni(\alpha',\ \beta')$，其中 $\alpha'\geq\alpha$ 且 $\beta'\geq\beta\}$。

Sajjadi等人 (2018年)表明，PR 曲线可参数化为 $\lambda\in\mathbb{R}^+\cup\{+\infty\}$ 如下：

$$\partial\text{PRD}(P,\widehat{P})=\left\{\alpha_\lambda(P\|\widehat{P}),\beta_\lambda(P\|\widehat{P})\mid\lambda\in\mathbb{R}^+\cup\{+\infty\}\right\}$$

$$\text{with}\begin{cases}\alpha_\lambda(P\|\widehat{P})=\int_{\mathcal{X}}\min\left(\lambda p(\boldsymbol{x}),\hat{p}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x},\\ \beta_\lambda(P\|\widehat{P})=\int_{\mathcal{X}}\min\left(p(\boldsymbol{x}),\hat{p}(\boldsymbol{x})/\lambda\right)\mathrm{d}\boldsymbol{x}.\end{cases}\tag{5}$$

$\widehat{\ }$$\widehat{\ }$注意，$\beta_\lambda(P\|\ P)\ =\alpha_\lambda(P\|\ P)/\lambda$。我们将 λ 称为权衡参数，因为它可用于调节对精确率（或召回率）的敏感度。图 2 展示了该 PR 曲线，其中目标分布 P 是由$\widehat{\ }$ $\widehat{\ }$两个高斯分布混合而成，另有两个候选模型 P_1 和 P_2。我们$\widehat{\ }$可以从图 2 中看到，P_1 在$\widehat{\ }$取值较大时表现更好 λ（对精确率敏感度较高），而 P_2在 λ 取值较小时表现更好（对召回率敏感度较高）。



(a) 分布　　$\widehat{\ }$(b) $\partial\text{PRD}(P,P)$ 曲线

图2. 两个模型的精确率-召回率曲线 $\widehat{P}_1$和 $\widehat{P}_2$的 P。图 2(a)展示了 $\widehat{P}_1$、$\widehat{P}_2$和 P。图 2(a)展示了 $\widehat{P}_1$、$\widehat{P}_2$相对于 P的精确率-召回率曲线。 $\widehat{P}_1$的召回率较高，因为它覆盖了 P的两个众数；但精确率较低，因为它会在两个众数之间生成点。 $\widehat{P}_2$的精确率较高，因为它不会在 P之外生成样本；但召回率较低，因为它只能从一个众数生成样本。

4. 精确率与召回率的权衡作为 af-散度

在本节中，我们将形式化精确率与召回率的权衡和 f-散度之间的联系，并回答问题 1。我们将在第 5 节利用这一联系，训练能够优化精确率与召回率之间特定权衡的模型。

4.1. 精确率-召回率作为 f-散度

我们首先介绍如下的*PR*-散度。

定义4.1 (精确率-召回率散度)。 给定权衡参数 $\lambda\in\mathbb{R}^+\cup+\infty$，精确率-召回率散度 （记作 $\mathcal{D}_{\lambda\text{-PR}}$）定义为 f_λ 的 f_λ-散度： $\mathbb{R}^+\to\mathbb{R}$ 定义为

$$f_\lambda:u\mapsto\max(\lambda u,1)-\max(\lambda,1),$$

for $\lambda\in\mathbb{R}^+$ *and* $f_\lambda:u\mapsto 0$ *for* $\lambda=+\infty$.

注意， f_λ 是连续的、凸的，并且满足 $f_\lambda(1)\ =0$对所有 λ 均成立。附录 A.2中给出了 f_λ 的图示。以下命题给出了 $\mathcal{D}_{\lambda\text{-PR}}$ 的一些性质。

命题 4.2（精确率-召回率散度的性质）。

- *Fenchel*共轭 f_λ^* *of* f_λ *is defined on*$\text{dom}\,(f_\lambda^*)=[0,\lambda]$ *and given by,*

$$f_\lambda^*(t)=\begin{cases}t/\lambda & \text{if } \lambda\leq 1,\\ t/\lambda+\lambda-1 & \text{otherwise.}\end{cases}\tag{6}$$

- 在$\widehat{\ }$的$\mathcal{D}_{\lambda\text{-精确率-召回率散度}}(P\|P)$ 变分表示中达到上确界的判别器为 $T^{\text{最优}}$

$$T^{\text{opt}}(\boldsymbol{x})=\lambda\text{sign}\left\{\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}-1\right\}.\tag{7}$$

- $\mathcal{D}_{\lambda\text{-PR}}(\widehat{P}\|P)=\lambda\mathcal{D}_{\frac{1}{\lambda}\text{-PR}}(P\|\widehat{P})$.
- $\mathcal{D}_{1\text{-PR}}(P\|\widehat{P})=\mathcal{D}_{\text{TV}}(P\|\widehat{P})/2$.

定义了精确率-召回率散度后，我们现在可以证明，相对于 λ 的精确率和召回率可以表示为$\widehat{}$与 P和 P之间散度的函数。

定理 4.3（精确率和召回率关于 $\mathcal{D}_{\lambda\text{-PR}}$ 的函数）。 给定 P、$\widehat{P} \in \mathcal{P}(\mathcal{X})$ 和 $\lambda \in \mathbb{R}^+ \cup \{+\infty\}$，*PR* 曲线 $\widehat{\ \ \ \

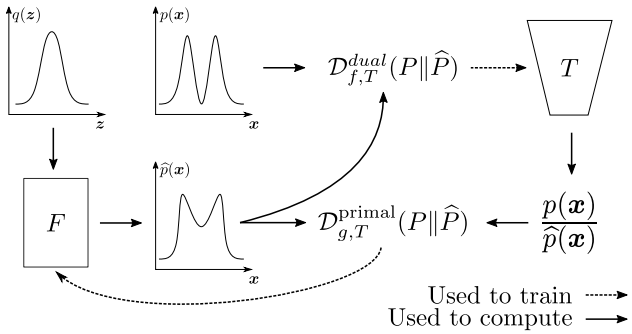


图3. 用于训练归一化流的模型架构 F，该模型最小化 $\mathcal{D}_g$；同时，判别器 T最小化 $\mathcal{D}_f$。

这一点，我们使用 D_f 的对偶形式中的 T最优。根据对偶形式，我们有如下结果。

$$\mathcal{D}_f(P\|\widehat{P}) = \sup_{T\in\mathcal{T}} \int_{\mathcal{X}} \Big(p(x)T(x) - \widehat{P}(x)f^*(T(x))\Big)\,dx.$$

对积分内的表达式关于 $T(x)$ 求导，可得如下结果。

$$\frac{p(x)}{\widehat{p}(x)} = \nabla f^*(T^{\text{opt}}(x)). \tag{14}$$

我们利用前述方程，提出以下 f-散度的原始估计 定义。

定义 5.1 (原始估计 $\mathcal{D}_{g\,T}^{\text{primal}}$,). 设 $P, \widehat{P} \in \mathcal{P}(\mathcal{X})$。对于函数 $T\colon \mathcal{X} \to \mathbb{R}$ 和 $f, g\colon \mathbb{R}^+ \to \mathbb{R}$，我们将散度 $\mathcal{D}_g$的原始估计定义为 $\mathcal{D}_{g\,T}^{\text{primal}}$,，如下所示。

$$\mathcal{D}_{g,T}^{\text{primal}}(P\|\widehat{P}) = \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) g\left(r(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}, \tag{15}$$

其中，$r\colon \mathcal{X} \to \mathbb{R}^+$ 定义为：$r(\boldsymbol{x}) = \nabla f^*(T(\boldsymbol{x}))$。

关键在于，当 T^{opt} 如（14）所给出时，我们有 $\mathcal{D}_{g\,T}^{\text{primal}} = \mathcal{D}_g$ $r\ x$ 。另请注意，() 充当似然 $p(x)/\widehat{p}(x)$ 的估计值。基于此，我们提出以下方案来训练归一化流 F，使其最小化$\mathcal{D}_g$，方法是使用一个判别器 T，使其最小化 $\mathcal{D}_f$。

- 判别器 T 使用对偶变分形式进行训练，以最小化 $\mathcal{D}_f$。

$$\begin{aligned}&\operatorname*{argmin}_T \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P})\\ &= \operatorname*{argmin}_T \mathbb{E}_P\left[T(\boldsymbol{x})\right] - \mathbb{E}_{\widehat{P}}\left[f^*(T(\boldsymbol{x}))\right].\end{aligned}$$

- 生成器是一种归一化流 F，通过原始形式训练以最大化 $\mathcal{D}_g$。

$$\begin{aligned}&\operatorname*{argmin}_T \mathcal{D}_{g,T}^{\text{primal}}\\ &= \operatorname*{argmin}_T \mathbb{E}_{\widehat{P}}\left[g(\nabla f^*(T(\boldsymbol{x})))\right].\end{aligned}$$

算法 1训练

输入： 流 F、判别器 T、数据集 D
对于 周期 $e = 1$、...、E**执行**
  $\mathcal{L}_d$、$\mathcal{L}_p \leftarrow 0$、$0$
  对于 $\boldsymbol{x}_{\text{real}} \in D$ **执行**
    使用 $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}_d,\ \boldsymbol{I}_d)$生成 $\boldsymbol{x}_{\text{fake}} = F^{-1}(\boldsymbol{z})$
    $\mathcal{L}_d \leftarrow \mathcal{L}_d + T(\boldsymbol{x}_{\text{real}}) - f^*(T(x_{\text{fake}}))$
    $\mathcal{L}_d \leftarrow \mathcal{L}_d + g(\nabla f^*(T(\boldsymbol{x}_{\text{fake}}))$
  结束循环
  更新 T 以最大化 $\mathcal{L}_d$
  更新 F 以最小化 $\mathcal{L}_p$
结束循环

注意，对于 $T = T_{\text{opt}}$，我们有 $\mathcal{D}_{f\,T,}^{\text{dual}}(P\|\widehat{P}) = \mathcal{D}_f$，因此，$\mathcal{D}_{g\quad,\ T}^{\text{primal}} = \mathcal{D}_g$。图5.1展示了模型架构，算法1描述了训练过程。

最后，为了训练生成器以最小化 $\mathcal{D}_{\lambda\text{-PR}}$，我们使用 $g = f_\lambda$ 和一个满足 $dom(f^*)$ 无界的 f，例如 KL、rKL 或 Jensen-Shannon 散度。在下文中，我们将讨论 f的选择如何影响模型训练。

5.2. 选择合适的 f 来训练 T

为了使用所提出的模型优化 PR 散度，我们固定 $g = f_\lambda$。这仍然使我们可以灵活地选择满足 $dom(f^*)$ 无界的 f。f的选择会影响似然 $p(x)/\widehat{p}(x)$ 的估计值 $r(x) = \nabla f^*(T(x))$。在下述定理中，我们将说明与 f 相关联的 Bregman 散度在对偶形式 $\mathcal{D}_{f\,T,}^{\text{dual}}(P\|\widehat{P})$ 的近似质量中起着关键作用。

定理 5.2 （在对偶形式下对 f 散度进行估计的误差。）。对于任意判别器 $T\colon \mathcal{X} \to \mathbb{R}$ 以及 $r(\boldsymbol{x}) = \nabla f^*(T(\boldsymbol{x}))$，

$$D_f(P\|\widehat{P}) - D_{f,T}^{\text{dual}}(P\|\widehat{P}) = \mathbb{E}_{\widehat{P}}\left[\mathrm{Breg}_f\left(r(\boldsymbol{x}), \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right)\right]. \tag{16}$$

在下一个定理中，我们给出 $\mathcal{D}_{g\,T}^{\text{primal}}$,$(P\|\widehat{P})$ 的近似误差界。

定理 5.3 （使用另一种 f 散度对 f 散度进行估计的界）。设 f、$g\colon \mathbb{R}^+ \to \mathbb{R}$ 满足：f 是 μ-强凸的，且 g 是 σ-*Lipschitz* 的。对于判别器 $T\colon \mathcal{X} \to \mathbb{R}$，令 $r(\boldsymbol{x}) = \nabla f^*(T(\boldsymbol{x}))$。如果

$$\mathcal{D}_f(P\|\widehat{P}) - \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) \le \epsilon,$$

那么，

$$\left|\mathcal{D}_g(P\|\widehat{P}) - \mathcal{D}_{g,T}^{\text{primal}}(P\|\widehat{P})\right| \le \sigma\sqrt{\frac{2\epsilon}{\mu}}.$$

注意，$g = f_\lambda$ 对于任意 λ，确实是 Lipschitz 的；并且 $\mathcal{D}_f$ 可以选取为使得 f 是强凸的（例如，χ^2散度）。因此，定理 5.2和 5.3 共同为我们所提出模型的收敛性提供了理论支撑。在图 4中，我们展示了原始形式和对偶形式收敛性的实证证据。

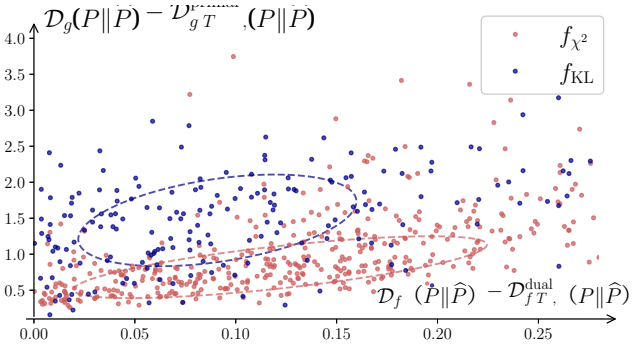
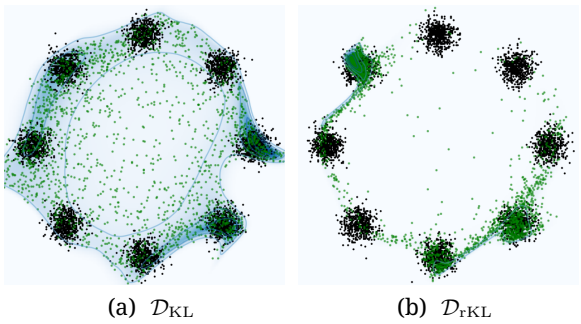


图 *4*。使用由 15 个高斯分布混合而成的 200 个不同的 P和 $\widehat{P}$ 训练了两个判别器：一个使用 $\mathcal{D}_{\text{KL}}$ 训练，另一个使用 $\mathcal{D}_{\chi^2}$训练。对于每个判别器，报告了$\mathcal{D}_g$ 与其原始形式之间的距离，以及 $\mathcal{D}_f$ 与其对偶形式之间的距离。在 y轴上，点越低，估计效果越好。红点表示使用 f_{χ^2} 训练的判别器，蓝点表示使用 f_{KL}训练的判别器。椭球表示每组数据的马氏距离。



(a) $\mathcal{D}_{\text{KL}}$ (b) $\mathcal{D}_{\text{rKL}}$

(c) 曲线下面积

图 *5.* 在二维 8个高斯分布混合模型上训练的 RealNVP。来自 P 的样本以黑色表示，而从 $\widehat{P}$ 中抽取的样本以绿色表示。$\widehat{P}$的对数似然以蓝色等高线表示，颜色越深，数值越高。模型的训练目标（从左到右）依次为最小化 $\mathcal{D}_{\text{KL}}$、$\mathcal{D}_{\text{rKL}}$ 以及曲线下面积（AUC）。对应的 PR曲线绘制在图 1(d)中。

6. 实际考量

6.1. 一种计算 $\partial\text{PRD}(P, \widehat{P})$的新方法

为了评估我们的模型，必须计算前沿 $\partial\text{PRD}(P, \widehat{P})$。Sajjadi 等人（2018）的原始研究提出了一种聚类方法，用于估计每个点 $(\alpha(\lambda), \beta(\lambda))$，但后来证明，这种方法无法估计密集排列的数据点（Kynk¨a¨anniemi 等人，2019）。近期的一种方法（Simon 等人，2019）提出训练一个二元分类器，以区分来自 P 和 $\widehat{P}$ 的点，然后使用假阳性率和假阴性率来估计前沿。既然我们已经证明，对于每个 $\lambda \in \mathbb{R}^+$，都可以通过 $\mathcal{D}_{\lambda\text{-PR}}$ 的原始形式来估计前沿，因此我们只需使用判别器 T 得到的估计比率 $\widehat{r}(\boldsymbol{x}) = p(\boldsymbol{x})/\widehat{p}(\boldsymbol{x})$。因此，将 $r(\boldsymbol{x}) = \nabla f^*(T(\boldsymbol{x}))$ 作为估计比率时，前沿上的估计精确率可写为。通过使用 $\mathcal{D}_{\chi^2}$ 训练 T，与先前方法使用的交叉熵（即 $\mathcal{D}_{\text{KL}}$）相比，我们可以为估计精确率提供有界误差。

6.2. 针对最优AUC进行训练

受分类领域相关工作的启发，我们可以计算甚至训练模型，以优化曲线下面积 （AUC）。这样得到的模型能够在所有可能的权衡范围内都表现良好，取值 λ 而非只针对某一种特定权衡最大化性能。曲线下面积可以根据命题 6.1 中给出的表达式计算，也可以用这一损失来训练模型。

命题 6.1 (在 $\partial\text{PRD}(P, \widehat{P})$)**下的曲线下面积**。 曲线下面积为：

$$\text{AUC} = \int_0^{+\infty} \alpha_\lambda(P\|\widehat{P})^2 \mathrm{d}\lambda \tag{17}$$

7. 实验

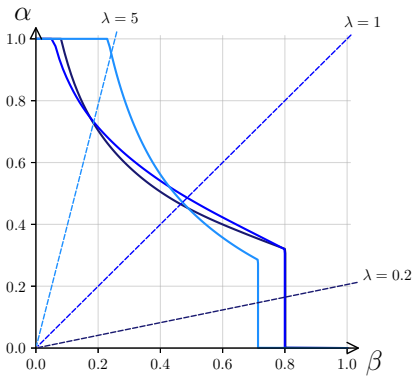
本节报告一系列实验，旨在说明使用精确率-召回率散度的优势。第一组实验在二维数据集（8个高斯分布）上进行，另一组则使用更高维的图像数据集 MNIST 和 CelebA。我们在第一个数据集上训练 RealNVP（Rezende和Mohamed，2016年），并在高维数据集上训练 GLOW（Kingma和Dhariwal，2018年）。对于每个模型，判别器 T都经过训练以最大化 $\mathcal{D}^{\text{dual}}_{\chi^2}(P\|\widehat{P})$，然后训练模型以最小化使用 T估算的不同损失，具体如第 5.1 节所述。

8 高斯数据集 图 1(a)、1(b) 和1(c) 展示了在 8个高斯分布数据集上训练的模型，其中 $\mathcal{D}_{\lambda\text{-PR}}$ 使用$\lambda = 0.1$, $\lambda = 1$ 和 $\lambda = 10$。如图所示，增大 λ 会显著提高精确率，但会牺牲召回率，反之亦然。图 5(a)、5(b) 和 5(c)展示了分别经过训练以最小化 $\mathcal{D}_{\mathrm{KL}}$、 $\mathcal{D}_{\mathrm{rKL}}$以及曲线下面积的模型。图中清楚展示了 $\mathcal{D}_{\mathrm{KL}}$ 和 $\mathcal{D}_{\mathrm{rKL}}$ 的质量覆盖与模式寻求行为。不过请注意，得益于精确率-召回率散度的灵活性，调整 λ 即可在这两种极端行为之间选择任意模型。

相应的PR曲线绘制于图5.2中。^两条绿色曲线是 $\partial \mathrm{PRD}(P、 P)$ 的前沿， $\mathcal{D}_{\mathrm{KL}}$和 $\mathcal{D}_{\mathrm{rKL}}$ 可以作为参考。对于 $\lambda = 1$，该模型位于两条曲线之间，其召回率低于 $\mathcal{D}_{\mathrm{KL}}$，精确率低于 $\mathcal{D}_{\mathrm{rKL}}$ ，但精确率高于 $\mathcal{D}_{\mathrm{KL}}$ ，召回率高于 $\mathcal{D}_{\mathrm{rKL}}$。由于 λ的值较高，我们赋予精确率相对较高的重要性，并且该模型在精确率方面优于 $\mathcal{D}_{\mathrm{rKL}}$。最后，AUC模型并不总是像更专业的模型那样优秀，但它具有最高的曲线下面积和最佳的 $\mathcal{D}_{\mathrm{TV}}$。



(a) $\lambda = \frac{1}{5}$ (b) $\lambda = 1$ (c) $\lambda = 5$



(d) PR曲线

图 6。 在 MNIST 上训练的 Glow 模型。给出了不同模型生成的小批量样本。

MNIST 和 CelebA 为在高维图像数据集上测试该框架，我们在 MNIST 数据集（Yann LeCun 等人，2010 年）和 CelebA 64×64 （Liu 等人，2015）上训练了不同的多尺度 Glow 模型。针对每个数据集，我们训练了三个模型，通过最小化 $\mathcal{D}_{\lambda\text{-PR}}$ 和 $\lambda = 1/5$， $\lambda = 1$ 以及 $\lambda = 5$。MNIST 生成图像的示例见图 6 ， CelebA 的示例见图 7。（更多样本批次见附录 B。）对于 MNIST，可以看到，当 $\lambda \leq 1$取值为这些数值时，生成图像的质量较差，这一点在 PR 曲线上表现为两条曲线都迅速下降。不过，使用 $\lambda = 10$ 训练的模型生成的样本方差较低（大多来自类别 1、9 8、6 和 4），但质量更好。因此，在 λ取值较高时，PR 曲线更高；在 λ取值较低时则更低。对于 CelebA，大多数正规化流在生成具有多样背景的高质量样本时都遇到困难。这里可以看到，将权衡设在召回率一侧时，模型会生成背景丰富多样的图像，但人脸质量较差。相反，将权衡设在精确率一侧时，模型会对人脸进行标准化，并提高其精确率。



(a) $\lambda = \frac{1}{5}$ (b) $\lambda = 1$ (c) $\lambda = 5$

图 7。 在 MNIST 上训练的 Glow 模型。给出了不同模型生成的小批量样本。

8. 结论

本文提出了一种在 f-散度框架内使用一种新型精确率/召回率（PR）散度训练归一化流的方法。与其他散度相比，我们的方法具有独特优势，因为它能显式控制生成模型中的精确率-召回率权衡。PR 散度可产生从极端模式寻求（高精确率）到极端权重覆盖（高召回率）的一系列模型，也能产生更均衡、可能更适用于各种应用场景的模型。我们的框架还为 KL 散度和反向 KL 散度等其他知名散度提供了新的见解。实验表明，针对高精确率训练的模型（使用反向 KL 散度）往往精确率较高，但召回率较差。鉴于通过目视检查模型样本来发现模型的低召回率，比发现低精确率更困难，我们怀疑现有的评估方法过分强调了高精确率模型。因此，我们希望这项研究最终能够推动开发更好的工具，以评估生成模型的整体质量。

致谢

我们感谢 GENCI 批准我们使用 IDRIS Jean Zay 集群的计算资源，资源分配编号为 AD011011296。

参考文献

Arjovsky、M.、Chintala、S. 和 Bottou、L. Wasserstein GAN。发表于 第*34*届国际机器学习会议论文集，澳大利亚悉尼，，2017年12月。网址：http://arxiv.org/abs/1701.07875。arXiv：1701.07875。

Behrmann、J.、Grathwohl、W.、Chen、R. T. Q.、Duvenaud、D. 和 Jacobsen、J.-H. 可逆残差网络。发表于第*36*届国际机器学习会议论文集，长滩，加利福尼亚州，*PMLR 97*，*2019*年，2019年5月。arXiv：1811.00995。

Chen、R. T. Q.、Behrmann、J.、Duvenaud、D. 和 Jacobsen、J.-H. 用于可逆生成建模的残差流。发表于 第*33*届神经信息处理系统大会（*NeurIPS 2019*），加拿大温哥华。，2020年7月。arXiv：1906.02735。

Djolonga, J., Lucic, M., Cuturi, M., Bachem, O., Bousquet, O. 和 Gelly, S. 使用信息散度前沿绘制精确率-召回率曲线，2020年6月。网址http://arxiv.org/abs/1905.10768。arXiv:1905.10768[cs, stat]。

Grover, A., Dhar, M. 和 Ermon, S. Flow-GAN：在生成模型中结合最大似然与对抗学习，2018年1月。网址http://arxiv.org/abs/1705.08868。arXiv:1705.08868[cs, stat]。

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. 和 Hochreiter, S. 通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。载于 神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。网址 https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html。

Issenhuth, T., Tanielian, U., Picard, D., 和 Mary, J. 潜在重加权：一种几乎零成本的生成对抗网络改进方法。载于 *2022*年*IEEE/CVF*冬季计算机视觉应用会议（*WACV*），第 3574–3583 页，怀科洛亚，夏威夷州，美国，2022年1月。IEEE。ISBN 978-1-66540-915-5。doi: 10.1109/WACV51458.2022.00363。网址：https://ieeexplore.ieee.org/document/9706934/。

Kingma, D. P. 和 Dhariwal, P. Glow：采用可逆1×1卷积的生成流。载于 第*32*届神经信息处理系统大会（*NeurIPS 2018*），蒙特利尔，加拿大。，第 31 卷，2018年。

Kobyzev, I., Prince, S. J. D., 和 Brubaker, M. A. 归一化流：导论与当前方法综述。 *IEEE*模式分析与机器智能汇刊，第 1–1 页，2020年。ISSN 0162-8828, 2160-9292, 1939-3539。doi: 10.1109/TPAMI.2020.2992934。网址：http://arxiv.org/abs/1908.09257。arXiv:1908.09257。

Kynk¨a¨anniemi, T., Karras, T., Laine, S., Lehtinen, J. 和 Aila, T. 用于评估生成模型的改进精确率与召回率指标。载于第*33*届神经信息处理系统大会（*NeurIPS 2019*），加拿大温哥华。，2019年10月。arXiv：1904.06991。

Liu, Z., Luo, P., Wang, X. 和 Tang, X. 野外人脸属性深度学习，2015年9月。网址http://arxiv.org/abs/1411.7766。arXiv:1411.7766[cs]。

Midgley, L. I., Stimper, V., Simm, G. N. C. 和 Hern´andez-Lobato, J. M. 自举你的流，2022年3月。网址 http://arxiv.org/abs/2111.11510。arXiv:2111.11510 [cs,stat]。

Midgley, L. I., Stimper, V., Simm, G.N. C., Sch¨olkopf, B.,and Hern´andez-Lobato, J. M. 流退火重要性采样自举，2022年11月。网址：http://arxiv.org/abs/2208.01893。arXiv:2208.01893[cs, q-bio, stat]。

Minka, T. 散度度量与消息传递。第17页，2005年。

Nguyen, X., Wainwright, M. J., and Jordan, M. I. 关于替代损失函数和 f-散度。《统计年鉴》，37(2)，2009年4月。ISSN0090-5364。doi: 10. 1214/08-AOS595。网址：http://arxiv.org/abs/math/0510521。arXiv:math/0510521。

Nowozin, S., Cseke, B. 和 Tomioka, R. f-GAN：使用变分散度最小化训练生成式神经采样器，2016年6月。网址 http://arxiv.org/abs/1606.00709。arXiv:1606.00709 [cs,stat]。

Rezende, D. J. 和 Mohamed, S. 使用归一化流进行变分推断。发表于*arXiv:1505.05770 [cs, stat]*，2016年6月。

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O. 和 Gelly, S.《通过精确率和召回率评估生成模型》。发表于 第 *32* 届神经信息处理系统大会（*NeurIPS 2018*），蒙特´利尔，加拿大，2018年10月。网址 http://arxiv.org/abs/1806.00035。arXiv:1806.00035。

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. 和 Chen, X.《改进生成对抗网络的训练技术》，2016年6月。网址：http://arxiv.org/abs/1606.03498。arXiv：1606.03498 [cs]。

Simon, L., Webster, R. 和 Rabin, J. 重新审视生成建模中的精确率与召回率定义。载于第*36*届国际机器学习大会论文集，第5799–5808页。PMLR，2019年5月。网址：https://proceedings.mlr.press/v97/simon19a.html。ISSN:2640-3498。

Stimper, V., Sch¨olkopf, B. 和Hern´andez-Lobato, J. M. 归一化流的基分布重采样。*arXiv:2110.15828 [cs, stat]*，2022年2月。网址：http://arxiv.org/abs/2110.15828。arXiv: 2110.15828。

Tanielian, U., Issenhuth, T., Dohmatob, E. 和 Mary, J. 学习不连通流形：一个没有生成对抗网络的世界。载于第*37*届国际机器学习大会论文集，奥地利维也纳，*PMLR 119*，*2020*年，2020年12月。arXiv: 2006.04596。

Yann LeCun、Corinna Cortes 和 Burges, C.《MNIST 手写数字数据库》。*ATTLabs*，2，2010年。网址 http://yann.lecun.com/exdb/mnist。

A. 附录 A.

A.1. 定理4.3的证明

我们需要证明，对于任意 $\lambda \in \mathbb{R}^+$， $\alpha(\lambda)$ 都可以表示为某种 f-散度的函数。首先，我们可以推导 $\alpha(\lambda)$ 的表达式：

$$\alpha(\lambda) = \int_{\mathcal{X}} \min\left(\lambda p(\boldsymbol{x}), \widehat{p}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \tag{18}$$

$$= \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) \min\left(\lambda \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}, 1\right) \mathrm{d}\boldsymbol{x} \tag{19}$$

要将此积分视为 f-散度，我们需要 f 先是凸下半连续的，然后满足$f(1) = 0$。然而，对于每个 $a, b \in \mathbb{R}$，最小值满足 $\min(a, b) = a + b - \max(a, b)$。因此，

$$\alpha(\lambda) = \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) \left[\lambda \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} + 1 - \max\left(\lambda \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}, 1\right)\right] \mathrm{d}\boldsymbol{x} \tag{20}$$

$$= \lambda \int_{\mathcal{X}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + 1 - \int_{\mathcal{X}} \max\left(\lambda \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}, 1\right) \mathrm{d}\boldsymbol{x} \tag{21}$$

$$= \lambda + 1 - \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) \max\left(\lambda \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}, 1\right) \mathrm{d}\boldsymbol{x} \tag{22}$$

因此，我们可以取 $f(u) = \max(\lambda u, 1) - \max(\lambda, 1)$，使得 $f(1) = 0$。精确率变为：

$$\alpha(\lambda) = \lambda + 1 - \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) f\left(\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) - \max(\lambda, 1) \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right. \tag{23}$$

$$= \min(\lambda, 1) - \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) f\left(\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} = \min(\lambda, 1) - \mathcal{D}_{\lambda\text{-PR}}(P, \widehat{P}).\right. \tag{24}$$

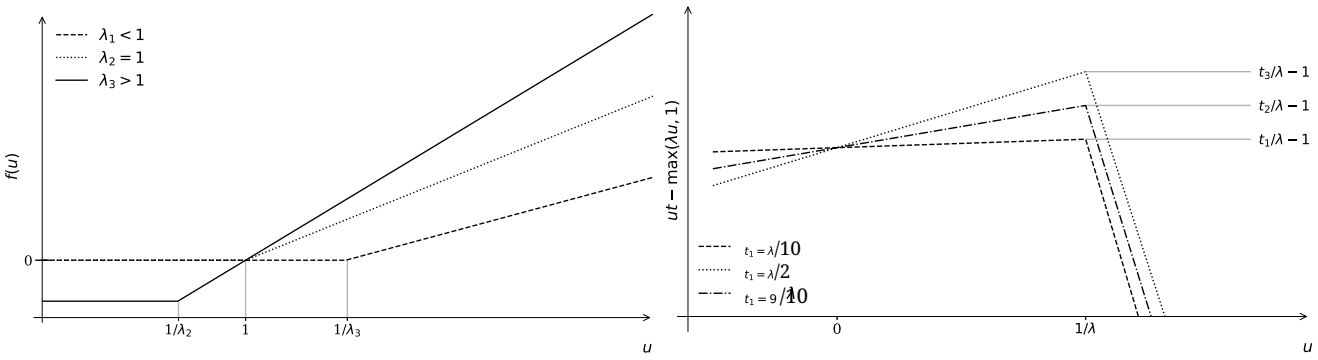
因此， $\alpha(\lambda)$ 可以写成一个 f-散度函数 $\mathcal{D}_{\lambda\text{-PR}}$，其中 $f(u) = \max(\lambda u, 1) - \max(\lambda, 1)$。

A.2. 命题4.2的证明

若精确率-召回率散度的生成函数 f为 $f(u) = \max(\lambda u, 1) - \max(\lambda, 1)$，则其 Fenchel 共轭函数为：

$$f^*(t) = \sup_{u \in \mathrm{dom}(f)} \{tu - f(u)\} = \max(\lambda, 1) + \sup_{u \in \mathbb{R}^+} \{tu - \max(\lambda u, 1)\} \tag{25}$$

如果 $t > \lambda$或 $\lambda < 0$，那么，对于各自的 上确界$_{u \in \mathbb{R}^+}$ $\{tu -$ 最大值（λu、1）$\} = \infty$ 分别为 $u \to \infty$ 和 $u \to -\infty$。因此， f^* 的定义域被限制为 $[0, \lambda]$。因此，对于 $0 \le t \le \lambda$，上确界在 $u = 1/\lambda$ 处取得，因为 0 属于函数 $1/\lambda$ 的次微分，如图8(b)所示。



(a) 不同 λ 取值下的函数 $f(u)$。

(b) t 介于 0 和 λ 之间时的函数 $u \mapsto ut - \max(\lambda u, 1)$。

因此， f 的Fenchel共轭为:

$$\forall t \in [0, \lambda], \quad f^*(t) = \max(\lambda, 1) + t\lambda - 1 = \begin{cases} t/\lambda & \text{if } \lambda \leq 1, \\ t/\lambda - 1 + \lambda & \text{otherwise.} \end{cases} \tag{26}$$

最后，对 T_{opt} 中的 f 关于 $^{p_{(\boldsymbol{x})}\,\widehat{p}(\boldsymbol{x})}$求导，得到最优判别器:

$$T^{\text{opt}}(\boldsymbol{x}) = \nabla f\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) = \begin{cases} \lambda & \text{if } \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} \leq 1/\lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

然后我们可以计算反向$\mathcal{D}_{\lambda\text{-PR}}$:

$$\begin{aligned}
\mathcal{D}_{\lambda\text{-PR}}(\widehat{P}\|P) &= \int_{\mathcal{X}} p(\boldsymbol{x}) f_\lambda\left(\frac{\widehat{p}(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} && (28)\\
&= \int_{\mathcal{X}} \max(\lambda\widehat{p}(\boldsymbol{x}), p(\boldsymbol{x})) - p(\boldsymbol{x}) \max(\lambda, 1)\, \mathrm{d}\boldsymbol{x} && (29)\\
&= \lambda\left(\int_{\mathcal{X}} \max(\widehat{p}(\boldsymbol{x}), p(\boldsymbol{x})/\lambda)\partial\boldsymbol{x} - \max(1, 1/\lambda)\right) && (30)\\
&= \lambda \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) \max(1, \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}/\lambda) - \widehat{p}(\boldsymbol{x}) \max(1, 1/\lambda)\, \mathrm{d}\boldsymbol{x} && (31)\\
&= \lambda \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) f_{1/\lambda}\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} && (32)\\
&= \lambda \mathcal{D}_{\frac{1}{\lambda}\text{-PR}}(P\|\widehat{P}). && (33)
\end{aligned}$$

利用这些结果，我们可以证明:

$$\begin{aligned}
\mathcal{D}_{\text{TV}}(P\|\widehat{P}) &= \int_{\mathcal{X}} |p(\boldsymbol{x}) - \widehat{p}(\boldsymbol{x})|\, \mathrm{d}\boldsymbol{x} && (34)\\
&= \int_{\mathcal{X}} \max(p(\boldsymbol{x}) - \widehat{p}(\boldsymbol{x}), 0) + \max(\widehat{p}(\boldsymbol{x}) - p(\boldsymbol{x}), 0) \mathrm{d}\boldsymbol{x} && (35)
\end{aligned}$$

然后，由于 $\mathcal{D}_{1\text{-PR}}(P\|\widehat{P}) = \int_{\mathcal{X}} \max(\widehat{p}(\boldsymbol{x})\ p(\boldsymbol{x})) - p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}} p(\boldsymbol{x})\ \max(\widehat{p}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ 且 $\mathcal{D}_{1\text{-PR}}(P\|\widehat{P}) = \mathcal{D}_{1\text{-PR}}(\widehat{P}\|P)$, 我们有:

$$\begin{aligned}
\mathcal{D}_{\text{TV}}(P\|\widehat{P}) &= \mathcal{D}_{1\text{-PR}}(P\|\widehat{P}) + \mathcal{D}_{1\text{-PR}}(\widehat{P}\|\P) && (36)\\
&= 2\mathcal{D}_{1\text{-PR}}(P\|\widehat{P}). && (37)
\end{aligned}$$

A.3. 定理5.2 的证明

^从现在起，假设 P 和 P 的支撑集相同。对于任意 T: $\mathcal{X} \to \mathbb{R}$,

$$\begin{aligned}
\mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) &= \mathbb{E}_{x\sim P}\left[d\left(x\right)\right] - \mathbb{E}_{x\sim Q}\left[f^*\left(d\left(x\right)\right)\right] \\
&= \mathbb{E}_{x\sim Q}\left[\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} d\left(x\right) - f^*\left(d\left(x\right)\right)\right]
\end{aligned}$$

令 $T_{\text{opt}} \in \arg\sup \mathcal{D}_{f\,T}^{\text{dual}}(P\|\widehat{P})$。对于任意 T: $\mathcal{X} \to \mathbb{R}$

$$\begin{aligned}
\mathcal{D}_f(P\|\widehat{P}) - D_{f,T}^{\text{dual}}(P\|\widehat{P}) &= \mathcal{D}_{f,T^{\text{opt}}}^{\text{dual}}(P\|\widehat{P}) - \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) \\
&= \mathbb{E}_{\widehat{P}}\left[\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\left(T^{\text{opt}}(\boldsymbol{x}) - T(\boldsymbol{x})\right) - f^*\left(T^{\text{opt}}(\boldsymbol{x})\right) + f^*\left(T(\boldsymbol{x})\right)\right]
\end{aligned}$$

已知，对于所有 $x \in \mathcal{X}$，均有 $\nabla f^*(T_{\text{opt}}(\boldsymbol{x})) = {}^{p}{}_{(\boldsymbol{x})}\,\widehat{p}\,(\boldsymbol{x})$：

$$\mathcal{D}_f(P\|\widehat{P}) - \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) = \mathbb{E}_{\widehat{P}}\left[\nabla f^*(T^{\text{opt}}(\boldsymbol{x}))\left(T^{\text{opt}}(\boldsymbol{x}) - T(\boldsymbol{x})\right) - f^*\left(T^{\text{opt}}(\boldsymbol{x})\right) + f^*\left(T(\boldsymbol{x})\right)\right]$$

回顾一下，对于任意连续可微的严格凸函数f，f的 Bregman 散度为 $\text{Breg}_f(a,\ b) = f(a) - f(b) - \langle \nabla f(b)\ a - b\rangle$。因此我们有

$$\mathcal{D}_f(P\|\widehat{P}) - \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) = \mathbb{E}_{\widehat{P}}\left[\text{Breg}_{f^*}\left(T(\boldsymbol{x}), T^{\text{opt}}(\boldsymbol{x})\right)\right]$$

现在让我们使用以下性质：$\text{Breg}_f(a,\ b) = \text{Breg}_{f^*}(a^*,\ b^*)$，其中 $a^* = \nabla f(a)$ 且 $b^* = \nabla f(b)$。

$$\begin{aligned}\mathcal{D}_f(P\|\widehat{P}) - \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) &= \mathbb{E}_{\widehat{P}}\left[\text{Breg}_f\left(\nabla f^*(T(\boldsymbol{x})), \nabla f^*(T^{\text{opt}}(\boldsymbol{x}))\right)\right] \\ &= \mathbb{E}_{\widehat{P}}\left[\text{Breg}_f\left(\nabla f^*(T(\boldsymbol{x})), \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right)\right]\end{aligned}$$

我们定义 $r(\boldsymbol{x}) = \nabla f^* T(\boldsymbol{x})$ 作为 $p(\boldsymbol{x})/\widehat{p}(\boldsymbol{x})$ 的估计量。最终，我们有

$$\mathcal{D}_f(P\|\widehat{P}) - \mathcal{D}_{f,T}^{\text{dual}}(P\|\widehat{P}) = \mathbb{E}_{\widehat{P}}\left[\text{Breg}_f\left(r(\boldsymbol{x}), \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right)\right]$$

A.4. 定理证明5.3

现在假设 f 是 μ-强凸的，则 $\text{Breg}_f(a,\ b) \geq \frac{\mu}{2}\|a-b\|^2$ 如果 $\mathbb{E}_{\widehat{P}}\left[\text{Breg}_f\left(r(\boldsymbol{x})\ {}^{p}{}_{(\boldsymbol{x})}\ \widehat{p}\ (\boldsymbol{x})\right)\right] \leq \epsilon$，并且如果 f 是 μ-强凸的，那么

$$\mathbb{E}_{\widehat{P}}\left[\left(r(\boldsymbol{x}) - \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right)^2\right] \leq \frac{2\epsilon}{\mu}. \tag{38}$$

考虑任意一个 f-散度 $\mathcal{D}_g(P\|\widehat{P}) = \int g\left(\frac{dP}{\mathrm{d}\,\widehat{P}}\right)\mathrm{d}\,\widehat{P}$。定义 $\mathcal{D}_{g\,T}^{\text{primal}}{}_{,}(P\|\widehat{P}) = \int g(r(\boldsymbol{x}))\,\mathrm{d}\,\widehat{P}$。于是，

$$\begin{aligned}\left|\mathcal{D}_g(P\|\widehat{P}) - \mathcal{D}_{g,T}^{\text{primal}}(P\|\widehat{P})\right| &= \left|\mathbb{E}_{\widehat{P}}\left[g\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) - g\left(r(\boldsymbol{x})\right)\right]\right| \\ &\leq \mathbb{E}_{\widehat{P}}\left[\left|g\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) - g\left(r(\boldsymbol{x})\right)\right|\right] \\ &\overset{(a)}{\leq} \mathbb{E}_{\widehat{P}}\left[\sigma\left|e(x)\right|\right] \\ &= \sigma\mathbb{E}_{\widehat{P}}\left[\left|e(x)\right|\right] \\ &\overset{(b)}{\leq} \sigma\sqrt{\mathbb{E}_{\widehat{P}}\left[e(x)^2\right]} \\ &\overset{(c)}{\leq} \sigma\sqrt{\frac{2\epsilon}{\mu}},\end{aligned}$$

其中，(a) 由 σ-Lipschitz 假设可得，该假设适用于 g，(b) 由詹森不等式可得，最后，(c) 由方程（38）可得。

A.5. 定理证明10

令 $c:\ \mathbb{R}^+ \mapsto \mathbb{R}$ 为一个 $\mathcal{C}^2$ 函数，并取 $u_{\min}$ 和 $u_{\max}$。目标是将所有 $u \in [u_{\min}, u_{\max}]$ 对应的 $f(u)$ 表示为以下各项的加权平均：f_λ^{PR}

$$\forall u \in \mathbb{R}_*^+, \int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda) f_\lambda^{\mathrm{PR}}(u)\mathrm{d}\lambda = \int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda)\left[\max(\lambda u, 1) - \max(\lambda, 1)\right]\mathrm{d}\lambda \tag{39}$$

首先，假设 $u_{\min} \leq 1$ 和 $u_{\max} \geq 1$，那么可以将这些项分解，并将积分拆开以求得最大值：

$$\int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda) f_\lambda^{\mathrm{PR}}(u)\mathrm{d}\lambda = \int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda)\max(\lambda u, 1)\mathrm{d}\lambda - \int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda)\max(\lambda, 1)\,\mathrm{d}\lambda \tag{40}$$

$$= \int_{1/u_{\max}}^{1/u} c''(\lambda)\max(\lambda u, 1)\mathrm{d}\lambda + \int_{1/u}^{1/u_{\min}} c''(\lambda)\max(\lambda u, 1)\mathrm{d}\lambda \tag{41}$$

$$- \int_{1/u_{\max}}^{1} c''(\lambda)\max(\lambda, 1)\,\mathrm{d}\lambda - \int_{1}^{1/u_{\min}} c''(\lambda)\max(\lambda, 1)\,\mathrm{d}\lambda$$

$$= \int_{1/u_{\max}}^{1/u} c''(\lambda)\mathrm{d}\lambda + \int_{1/u}^{1/u_{\min}} c''(\lambda)\lambda u\mathrm{d}\lambda - \int_{1/u_{\max}}^{1} c''(\lambda)\mathrm{d}\lambda - \int_{1}^{1/u_{\min}} c''(\lambda)\lambda\mathrm{d}\lambda. \tag{42}$$

通过分部积分，我们得到：$\int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda)\lambda\mathrm{d}\lambda = \left[c'(\lambda)\lambda\right]_{1/u_{\max}}^{1/u_{\min}} - \int_{1/u_{\max}}^{1/u_{\min}} c'(\lambda)\mathrm{d}\lambda$ 因此满足：

$$\int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda) f_\lambda^{\mathrm{PR}}(u)\mathrm{d}\lambda = \int_{1/u_{\max}}^{1/u} c''(\lambda)\mathrm{d}\lambda + u\left[c'(\lambda)\lambda\right]_{1/u}^{1/u_{\min}} - u\int_{1/u}^{1/u_{\min}} c'(\lambda)\mathrm{d}\lambda \tag{43}$$

$$- \int_{1/u_{\max}}^{1} c''(\lambda)\mathrm{d}\lambda - \left[c'(\lambda)\lambda\right]_{1}^{1/u_{\min}} + \int_{1}^{1/u_{\min}} c'(\lambda)\mathrm{d}\lambda$$

$$= \left[c'(\lambda)\right]_{1/u_{\max}}^{1/u} + u\left[c'(\lambda)\lambda\right]_{1/u}^{1/u_{\min}} - u\left[c(\lambda)\right]_{1/u}^{1/u_{\min}} \tag{44}$$

$$- \left[c'(\lambda)\right]_{1/u_{\max}}^{1} - \left[c'(\lambda)\lambda\right]_{1}^{1/u_{\min}} + \left[c(\lambda)\right]_{1}^{1/u_{\min}}$$

$$= c'\left(\frac{1}{u}\right) - c'(0) + uc'\left(\frac{1}{u_{\min}}\right) - uc'\left(\frac{1}{u}\right)\frac{1}{u} - uc\left(\frac{1}{u_{\min}}\right) + uc\left(\frac{1}{u}\right) - c'(1) \tag{45}$$

$$+ c'(0) - c'\left(\frac{1}{u_{\min}}\right)\frac{1}{u_{\min}} + c'(1)\times 1 + c\left(\frac{1}{u_{\min}}\right) - c(1)$$

$$= \left[c'\left(\frac{1}{u_{\min}}\right)\frac{1}{u_{\min}} - c\left(\frac{1}{u_{\min}}\right)\right](u-1) + uc\left(\frac{1}{u}\right) - c(1). \tag{46}$$

我们希望 $\int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda) f_\lambda^{\mathrm{PR}}(u)\mathrm{d}\lambda$ 在 $u_{\min}$ 和 $u_{\max}$之间等于 f。但是，由于由f和 g生成的两个 f-散度相等，当存在一个 $c \in \mathbb{R}$ 使得 $f(u) = g(u) + c(u-1)$时，由 $\int_{1/u_{\max}}^{1/u_{\min}} c''(\lambda) f_\lambda^{\mathrm{PR}}(u)\mathrm{d}\lambda$生成的散度就等于由 $uc\left(\frac{1}{u}\right) - c(1)$生成的散度。因此，我们要求 c满足：

$$\forall u \in [u_{\min}, u_{\max}], \quad f(u) = uc\left(\frac{1}{u}\right) - c(1).$$

对 u求导可得：

$$f'(u) = \lim_{\lambda\to\infty}\left[c'(\lambda)\lambda - c(\lambda)\right] + c\left(\frac{1}{u}\right) - \frac{1}{u}c'\left(\frac{1}{u}\right). \tag{47}$$

最后：

$$f''(u) = -\frac{1}{u^2}c\left(\frac{1}{u}\right) + \frac{1}{u^2}c'\left(\frac{1}{u}\right) + \frac{1}{u^3}c''\left(\frac{1}{u}\right) \tag{48}$$

$$= \frac{1}{u^3}c''\left(\frac{1}{u}\right). \tag{49}$$

因此，对于 $\lambda = 1/u$，我们有：

$$c''(\lambda) = \frac{1}{\lambda^3} f''\left(\frac{1}{\lambda}\right). \tag{50}$$

根据这样的结果，对于 $m = \min_{\mathcal{X}}\left(\frac{\widehat{p}(\boldsymbol{x})}{p(\boldsymbol{x})}\right)$ 和 $M = \max_{\mathcal{X}}\left(\frac{\widehat{p}(\boldsymbol{x})}{p(\boldsymbol{x})}\right)$，我们可以将任意 f-散度写为：

$$\begin{aligned}
\mathcal{D}_f(P\|\widehat{P}) &= \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) \int_m^M \frac{1}{\lambda^3} f''\left(\frac{1}{\lambda}\right) f_\lambda^{\mathrm{PR}}\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) \mathrm{d}\lambda \mathrm{d}\boldsymbol{x} \\
&= \int_m^M \int_{\mathcal{X}} \frac{1}{\lambda^3} f''\left(\frac{1}{\lambda}\right) \widehat{p}(\boldsymbol{x}) f_\lambda^{\mathrm{PR}}\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) \mathrm{d}\lambda \mathrm{d}\boldsymbol{x} \\
&= \int_m^M \frac{1}{\lambda^3} f''\left(\frac{1}{\lambda}\right) \left[\int_{\mathcal{X}} \widehat{p}(\boldsymbol{x}) f_\lambda^{\mathrm{PR}}\left(\frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}\right] \mathrm{d}\lambda \\
&= \int_m^M \frac{1}{\lambda^3} f''\left(\frac{1}{\lambda}\right) \mathcal{D}_{\lambda\text{-PR}}(P\|\widehat{P}) \mathrm{d}\lambda
\end{aligned}$$

A.6. 推论4.5 的证明

特别地，对于 $\mathcal{D}_{\mathrm{KL}}$, $f(u) = u\log u$，因此 $f''(u) = 1/u$，得到：

$$\mathcal{D}_{\mathrm{KL}}(P\|\widehat{P}) = \int_m^M \frac{1}{\lambda^2} \mathcal{D}_{\lambda\text{-PR}}(P\|\widehat{P}) \mathrm{d}\lambda. \tag{51}$$

而对于 $\mathcal{D}_{\mathrm{rKL}}$，我们可以 either use Equation 50 与 $f(u) = -\log u$ 或者利用以下事实：$\mathcal{D}_{\lambda\text{-PR}}(P\|P) = \lambda \mathcal{D}_{1/\lambda\text{-PR}}(P\|P)$：

$$\mathcal{D}_{\mathrm{rKL}}(P\|\widehat{P}) = \int_m^M \frac{1}{\lambda} \mathcal{D}_{\lambda\text{-PR}}(P\|\widehat{P}) \mathrm{d}\lambda. \tag{52}$$

A.7. 命题 6.1 的证明

AUC 可以通过对第一象限内的角度 θ 积分来计算：

$$\mathrm{AUC} = \int_0^{\pi/2} r^2(\theta) \mathrm{d}\theta. \tag{53}$$

因此，令 $\lambda = \tan\theta$，可得 $r(\theta) = \alpha_{\tan(\theta)}(P\|P)/\cos\theta$。于是：

$$\mathrm{AUC} = \int_0^{\pi/2} \frac{\alpha_{\tan(\theta)}(P\|\widehat{P})^2}{\cos^2\theta} \mathrm{d}\theta \tag{54}$$

$$= \int_0^{+\infty} \alpha_\lambda(P\|\widehat{P})^2 \cos^2\theta \mathrm{d}\lambda \quad \text{with} \quad \frac{\mathrm{d}\theta}{\cos^2\theta} = \partial\lambda. \tag{55}$$

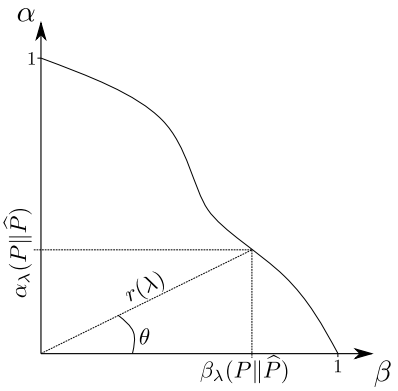
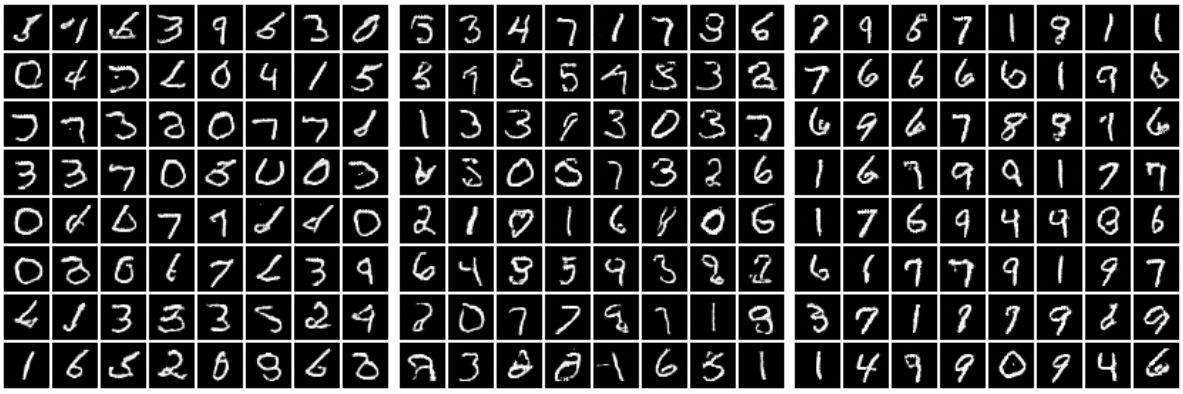


图 *A.* *8* ∂ P $\lambda \in$ 。计算曲线下面积时变量替换的示 $\mathbb{R} \cup \{\infty\}$ $\theta \in [0, \frac{\pi}{2}]$ $\lambda =$ θ 意图。我们不以 为参数化前沿 PRD(, $\widehat{P}$)，而是取 tan 。

B. 实验



(a) $\lambda = \frac{1}{5}$ (b) $\lambda = 1$ (c) $\lambda = 5$



(d) $\lambda = \frac{1}{5}$ (e) $\lambda = 1$ (f) $\lambda = 5$

图 *B.9*。 不同 Glow 模型在 MNIST（(a)、(b) 和 (c)）与 CelebaA（(d)、(e) 和 (f)）上训练后生成的大批样本。模型经过训练以最小化 $\mathcal{D}_{\lambda\text{-PR}}$，图注中给出了不同的 λ。λ 越低，模型越重视召回率；而当 λ取值较高时，模型则会高度重视精确率。

Th实验的详细信息见表 格 **??**

数据集	模型	#参数 F	#参数 D
8个高斯分布	RealNVP	540k	659k
MNIST	Glow	85M	1.7M
CelabA	Glow	188M	13.2M

表 *2.* 不同实验所用模型的详细信息。