重新审视深度线性判别分析

Maxat Tezekbayev[1], Rustem Takhanov[1], Arman Bolatov[2,4], Zhenisbek Assylbekov[3,4*]
[1]哈萨克斯坦阿斯塔纳纳扎尔巴耶夫大学数学系[2]阿拉伯联合酋长国阿布扎比穆罕默德·本·扎耶德人工智能大学机器学习系[3]美国印第安纳州韦恩堡普渡大学韦恩堡分校数学科学系[4]哈萨克斯坦阿拉木图国家空间研究与技术中心遥感系 maxat.tezekbayev@alumni.nu.edu.kz, rustem.takhanov@nu.edu.kz, arman.bolatov@mbzuai.ac.ae, zassylbe@pfw.edu

arXiv:2601.01619v1 [stat.ML] 4 Jan 2026

我们证明，对于不受约束的深度线性判别分析（LDA）分类器，最大似然训练会产生病态解：类别均值逐渐靠拢，协方差坍缩，学习到的表示几乎不具备判别性。相反，交叉熵训练虽能取得出色的准确率，却会使分类头与底层生成模型脱节，导致参数估计高度不一致。为兼顾生成式结构与判别性能，我们提出判别式负对数似然（DNLL）损失，它在 LDA 对数似然中加入了一个针对混合密度的简单惩罚项。DNLL 可解释为标准 LDA 负对数似然加上一项，用于明确抑制多个类别同时具有较高可能性的区域。使用 DNLL 训练的深度 LDA 能够产生清晰且彼此分离的潜在空间，在合成数据和标准图像基准上达到与 Softmax 分类器相当的测试准确率，并显著改善预测概率的校准效果，从而恢复深度判别模型连贯的概率解释。

1. 引言

经典线性判别分析（LDA）可追溯至 Fisher [1] 和 Rao [2]的开创性研究。它是最简单且应用最广泛的分类方法之一，并具有清晰的概率解释。设 $X \in \mathbb{R}^d$ 为输入向量，$Y \in \{1, 2, \ldots, C\}$ 为相应的类别标签。LDA 假设如下生成模型：

- 类别标签服从分类分布，其概率为 $\pi = (\pi_1, \ldots, \pi_C)$[2]，定义在概率单纯形 Δ^{C-1}上：

$$\Pr[Y = c] = \pi_c, \qquad \pi_c \geq 0, \sum_{c=1}^{C} \pi_c = 1, \tag{1}$$

- 类条件分布为高斯分布：

$$X \mid Y = c \ \sim \ \mathcal{N}(\mu_c, \Sigma), \tag{2}$$

其中，协方差矩阵 Σ 对所有类别都相同。

以这种方式指定 Y 的边缘分布和 $X \mid Y$ 的条件分布，便定义了 (X, Y) 上的联合分布。因此，模型参数[3] $\theta := (\pi, \{\mu_c\}_{c=1}^{C}, \Sigma)$ 可以

*通讯作者。[2]为避免与常数 π混淆，我们用 π（以及 π_c）表示类别先验概率，而 π 表示数学常数。[3]在全文中，θ 表示所讨论模型的所有可学习参数的集合；其确切内容将根据上下文明确。

预印本。进行中的工作。

通过最大似然进行估计。模型拟合完成后，对于新的输入 x ，可以通过以下方式进行预测：

$$\hat{y}(x) \;=\; \arg\max_{c} p_{\hat{\theta}}(x, Y = c),$$

其中 $p_{\theta}(x,\ y)$ 是学得的 $(X,\ Y)$ 联合密度；等价地，这是 $\arg\max_c p_{\hat{\theta}}(Y = c \mid X = x)$。使用相应的判别函数 $\{\delta_c(x)\}_{c=1}^{C}$ 通常会更方便，其定义如下：

$$\delta_c(x) \;:=\; \log \pi_c - \frac{1}{2}\log\left(\det \Sigma\right) - \tfrac{1}{2}(x-\mu_c)^{\top}\Sigma^{-1}(x-\mu_c). \tag{3}$$

$\hat{}$ 分类器因此可写为 $\hat{y}(x) \;= \arg\max_c \delta_c(x)$。

经典线性判别分析的一项主要局限是会产生线性决策边界。当类别在输入空间中并非线性可分时，这种限制可能过于严格。为克服这一局限，人们提出了几种方法。核方法（例如核费歇尔判别分析）[3], 会将数据隐式映射到高维特征空间，在那里线性分离更有可能实现。另一种方法是不再用单个高斯分布对每个类别建模，而是使用高斯混合，从而得到混合判别分析 [4]。

本文聚焦于深度线性判别分析
编码器中，并在
Dorfer 等人 [5] 并未优化 LDA模型的似然；相反，他们试图最大化费歇尔广义特征值目标。直接优化经典的费歇尔目标被证明存在问题，因此他们不得不采用一种启发式方法，强调判别能力较弱的特征方向。因此，他们最终的训练目标偏离了原始的费歇尔准则。

与此同时，LDA 本质上是一种概率模型，因此，通过最大似然训练其深度变体，同时保留连贯的生成式解释，将非常理想。然而，正如我们在第2节中所示，当既不对编码器也不对 LDA 头施加约束时，深度

[5]， 将输入馈送到神经
得到的潜在空间中应用 LDA 头（图1）。

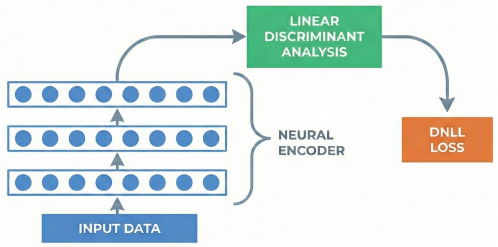


图1：深度 LDA 架构。神经编码器将输入数据映射到潜在空间，并在该空间中应用 LDA；网络则使用所提出的判别式负对数似然（DNLL）损失进行优化。

线性判别分析可能会与判别任务几乎完全脱节。最大似然会被一些病态解最大化：嵌入紧密地坍缩在各自类别均值附近，共享协方差矩阵变得近乎奇异，而不同类别的均值则逐渐彼此靠近。这些配置虽然能获得很高的似然，却会导致簇严重重叠、分类准确率低下，以及不可靠的置信度估计。

从生成式范式转向判别式范式——将线性判别分析判别函数（3）视为逻辑值，并优化交叉熵——确实能提高分类准确率。然而，在这种情况下，线性判别分析参数不再能作为底层生成模型的参数得到一致估计；这种现象在经典线性判别分析中就已可见。

由此引出了本文的核心问题：

> 我们能否通过最大似然训练深度线性判别分析，同时不牺牲参数一致性或判别能力？

在第3节中，我们通过引入一种惩罚似然目标来回答这个问题，并将其称为**判别式负对数似然（DNLL）**。具体而言，我们在 LDA 对数似然中加入一个惩罚项，明确促进类别分离：它既能防止退化解使似然值无限增大，也能鼓励编码器将输入映射到类别均值周围良好分离的簇中。我们对合成数据的分析进一步表明，

对于分离良好的类别，优化这种惩罚损失所得的参数估计仍接近最大似然解。

在第4节中，我们在图像分类任务上使用 DNLL 损失训练深度线性判别分析。最终模型的测试准确率与使用交叉熵训练的 softmax头模型相当。至关重要的是，它还能产生校准效果显著更好的预测概率，这一点可通过可靠性图和期望校准误差（ECE）进行量化。这支持了这样一种观点：采用显式生成式头进行基于似然的训练，可以在不牺牲准确率的情况下提升置信度估计的质量。复现实验所需的代码可在 https://github.com/zh3nis/DNLL 获取。

据我们所知，此前没有研究分析过深度线性判别分析在最大似然训练下的失效模式，也没有提出一种原则明确、基于似然的训练方法，既能强制实现判别式分离，又能保留线性判别分析的生成式结构。

2. 最大似然训练：初探

在转向深度架构之前，我们首先检验在一个简单的合成数据集上，能否通过基于梯度的最大似然方法准确拟合经典（非深度）线性判别分析。

2.1. 预热：经典线性判别分析

我们采用公式 (1)–(2) 中的生成模型。给定一个从公式 (1)–(2) 中抽取的带标签样本 $\{(x_1,\ y_1).....(x_n,\ y_n)\}$，负对数似然为

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\Big(\pi_{y_i}\, \phi(x_i;\, \mu_{y_i}, \Sigma)\Big), \tag{4}$$

其中，对于 $x \in \mathbb{R}^d$,

$$\phi(x; \mu, \Sigma) = (2\pi)^{-d/2}\,(\det \Sigma)^{-1/2} \exp\big\{-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\big\}$$

是 $\mathcal{N}(\mu,\ \Sigma)$ 的密度。

***** 优化 我们使用 Adam [6]最小化 $\mathcal{L}_{\mathrm{NLL}}(\theta)$。为使优化器不受约束，我们采用以下重参数化方法：(i) 逻辑值 $\alpha \in \mathbb{R}^C$ with $\pi = \mathrm{Softmax}(\alpha)$，以及 (ii) 乔列斯基分解 $\Sigma = LL^\top$，其中 L 为对角线元素为正的下三角矩阵。这保证了 $\pi \in \Delta^{C-1}$ 和 $\Sigma \succ 0$，同时支持标准的自动微分。

***** 结果 在根据公式 (1)–(2) 生成的数据上，基于梯度的优化能够可靠地恢复最大似然解；参见图2。具体而言，学习得到的先验、均值和协方差与真实参数高度一致，决策边界也与通过 LDA闭式估计得到的结果重合。这证实了经典模型可以通过基于梯度的似然最大化进行训练。

***** 交叉熵训练与不一致性 或者，我们也可以舍弃生成式解释，将判别函数 (3) 视为逻辑值，对 LDA 头进行判别式训练。事实上，在 LDA模型下，根据贝叶斯定理可得

$$p_\theta(Y = c \mid x) = \frac{\pi_c\, \varphi(x; \mu_c, \Sigma)}{\sum_{j=1}^{C} \pi_j\, \varphi(x; \mu_j, \Sigma)} = \frac{\exp(\delta_c(x))}{\sum_{j=1}^{C} \exp(\delta_j(x))} \;=\; \mathrm{softmax}\big(\delta(x)\big)_c, \tag{5}$$

其中 $\delta(x)\ = (\delta_1(x) \ldots, \delta_C(x))^\top$，并且等式成立是因为 $\delta_c(x)\ =$ 对数 $\pi_c + $ 对数 $\varphi(x;\ \mu_c, \Sigma)$ 相差一个与 c 无关的加性常数（该常数在归一化时会抵消）。

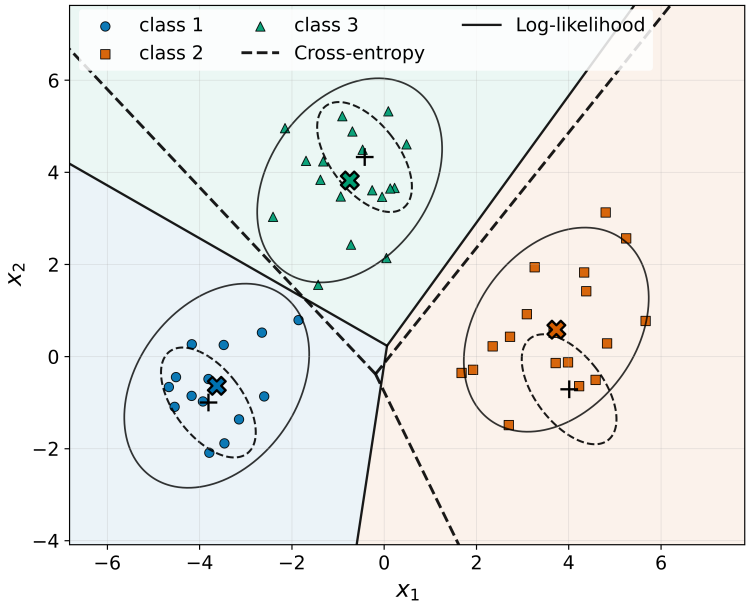


图2：对合成数据进行经典三分类线性判别分析拟合，分别采用最大似然法（实线椭圆、实线决策边界，均值以“×”标记）和将判别得分用作逻辑值、通过最小化交叉熵进行拟合的方法（虚线椭圆、虚线边界，均值以“+”标记）。高斯分量（均值和协方差）存在明显差异，这表明交叉熵训练无法恢复生成式 LDA 模型的最大似然参数。椭圆表示拟合所得高斯类别的 90% 等高线。

这表明，可以通过最小化交叉熵（负条件对数似然）来优化 θ，其中逻辑值为 $\delta_c(x)$:

$$\ell_{\mathrm{CE}}(x, y; \theta) := -\log p_\theta(y \mid x) = -\log \frac{\exp(\delta_y(x))}{\sum_{c=1}^{C} \exp(\delta_c(x))}$$

$$= -\delta_y(x) + \log \sum_{c=1}^{C} \exp(\delta_c(x)), \tag{6}$$

$$\mathcal{L}_{\mathrm{CE}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{CE}}(x_i, y_i; \theta). \tag{7}$$

对于与上述相同的合成数据，我们将此过程与基于梯度的最大似然估计进行比较。两种目标函数都能得到完美的训练准确率，但估计得到的高斯分量存在显著差异；见图2。具体而言，交叉熵解会使类别均值发生偏移并缩小协方差，从而产生更紧凑、与数据生成模型不再吻合的椭圆。因此，尽管交叉熵训练能得到良好的分类器，但它无法恢复底层 LDA 模型的真实参数，因此在这种生成式设定下是不一致估计量。

2.2. 一个警示性实验：深度线性判别分析的似然训练

现在我们转向深度情形。对于深度线性判别分析实验，我们再次根据模型 (1)–(2) 生成数据集$\{(x_1, y_1) \ldots, (x_n, y_n)\}$，但将输入通过神经编码器传递。具体而言，我们令 $f_\psi(\cdot)$ 为一个具有 ReLU激活和 32 个隐藏单元的两层前馈网络。由此得到的线性判别分析负对数似然为

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log(\pi_{y_i}\, \phi(f_\psi(x_i);\, \mu_{y_i}, \Sigma)), \tag{8}$$

参数为 $\theta := (\psi, \pi, \{\mu_c\}_{c=1}^{C}, \Sigma)$。

我们绘制了 20，000 个训练点和 4，000 个测试点。使用 Adam 和 PyTorch 的默认超参数进行 100 个训练轮数，小批量大小为 256（评估时的批量大小为 1024）。学习到的嵌入 $z_i = f_\psi(x_i)$ 如图3所示。

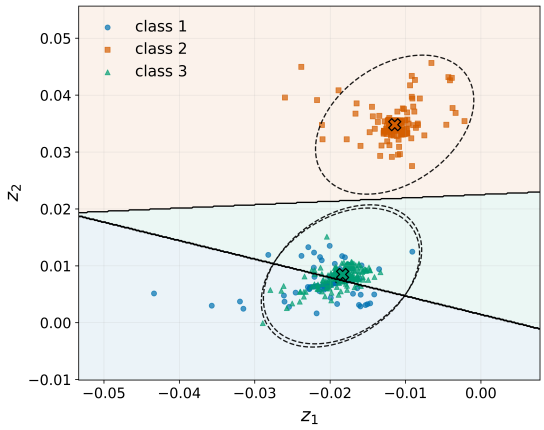


图3：似然训练后的深度线性判别分析嵌入 $z_i = f_\psi(x_i)$。两个类别合并为单个簇；样本紧密聚集在各自的类别质心 $\{\mu_c\}$ 周围，而共享协方差 Σ变得近乎奇异（$|\Sigma| \approx 3 \cdot 10^{-10}$）。训练准确率：67.2%；测试准确率：67.7%。

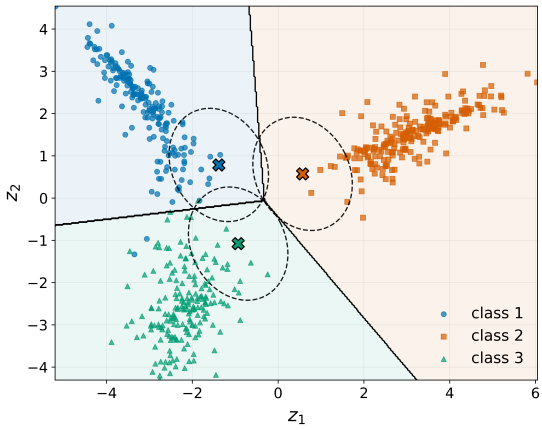


图 4：交叉熵训练后的深度线性判别分析嵌入 $f_\psi(x_i)$。高斯分量位于边界的交点附近，而不是经验类别云周围，从而导致参数估计不一致。训练准确率：99.4%；测试准确率：99.6%。

这种失效模式显而易见：优化器并未强制所有类别彼此分离——其中两个类别簇坍缩为一个。此外，对于每个带标签样本对（x，y），编码器将 $f_\psi(x)$ 推得极其接近其类别均值 μ_y，使马氏距离平方

$$(f_\psi(x) - \mu_y)^\top \Sigma^{-1} (f_\psi(x) - \mu_y)$$

趋近于零，同时将 Σ 压缩至接近奇异的状态。这两种效应都会增大高斯项 $\phi\big(f_\psi(x);\ \mu_y,\ \Sigma\big)$，因此，高度集中的、几乎重叠的簇会使对数似然达到最大。

尽管似然值非常高，分类器的表现却很差：最终训练准确率和测试准确率分别仅为 67.2% 和 67.7%。这说明，深度线性判别分析的无约束最大似然目标与学习良好分离的判别式表示这一目标存在根本性偏差。

2.3. 使用交叉熵损失的深度线性判别分析

上述实验表明，对深度线性判别分析进行无约束的最大似然训练会导致高度退化的解。一种自然的替代方案是以判别式方式训练相同的架构。具体而言，我们保留编码器 f_ψ 和 LDA 头，但不再最大化对数似然，而是将判别分数 $\delta_c(f_\psi(x))$ 视为逻辑值，并最小化交叉熵损失。

我们使用相同的数据生成模型、编码器架构、优化器和训练计划，重复上一小节中的合成实验，仅将目标函数从负对数似然改为交叉熵。所得嵌入 $z_i = f_\psi(x_i)$ 和决策边界如图 4 所示。与采用似然训练的模型不同，三个类别形成了彼此 distinct 的簇。训练和测试准确率都接近完美，证实交叉熵训练能够得到出色的分类器。

然而，LDA 头的参数不再与经验类别结构相吻合。学到的均值 $\{\mu_c\}$ 位于决策边界的交点附近，而图中虚线椭圆

图 4——这些椭圆本应反映共享的协方差结构——与编码器生成的细长类别簇相去甚远。换言之，判别式训练目标仅将 LDA 头用作逻辑值的一种便捷参数化形式，而所得的 $(\pi, \{\mu_c\}, \Sigma)$ 与底层高斯混合分布的最大似然参数相差甚远。这与浅层设置中观察到的不一致现象如出一辙，也说明即使分类性能出色，交叉熵训练仍会破坏模型参数的概率解释，深度线性判别分析亦是如此。

3. 判别式负对数似然

受第2节中失败模式的启发，我们提出了一种简单的损失函数，在保留基于似然的训练的同时，明确鼓励类别分离。

3.1. 定义

设 $z = f_\psi(x)$ 为输入 x 的嵌入，$\{\delta_c(z)\}_{c=1}^{C}$ 为潜在空间中 LDA 头的相应判别函数。对于一个带标签的样本 $(x,\ y)$，生成式 LDA 模型的经典负对数似然（NLL）除一个加性常数外为，

$$\ell_{\mathrm{NLL}}(z, y) \;=\; -\delta_y\big(z\big). \tag{9}$$

为了促进　判别能力，我们在（9）中加入了一个针对未归一化类别密度的惩罚项　　，

$$R(z) \;=\; \sum_{c=1}^{C} \exp\big(\delta_c(z)\big). \tag{10}$$

提出的判别式负对数似然（DNLL）定义为

$$\ell_{\mathrm{DNLL}}(z, y; \theta) := -\delta_y\big(z\big) + \lambda \sum_{c=1}^{C} \exp\big(\delta_c(z)\big), \qquad \lambda \geq 0, \tag{11}$$

$$\mathcal{L}_{\mathrm{DNLL}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{DNLL}}(f_\psi(x_i), y_i; \theta) \tag{12}$$

其中 λ 控制判别式正则化的强度。当 $\lambda = 0$时，我们得到纯粹的最大似然训练；而 $\lambda > 0$ 则会惩罚那些位于多个类别密度同时较大的区域中的嵌入。

对于线性判别分析，我们可以用高斯密度重写式（11）。由于

$$\exp\big(\delta_c(z)\big) \;=\; \pi_c\, \phi(z; \mu_c, \Sigma) \cdot (2\pi)^{d/2},$$

惩罚项（10）等于 z 处的混合密度，二者仅相差一个全局常数:

$$R(z) \;\propto\; \sum_{c=1}^{C} \pi_c\, \phi\big(z; \mu_c, \Sigma\big).$$

因此，DNLL 可分解为

$$\ell_{\mathrm{DNLL}}(z, y) \;=\; \underbrace{-\log\big(\pi_y \phi\big(z; \mu_y, \Sigma\big)\big)}_{\mathrm{NLL}} + \lambda \underbrace{\sum_{c=1}^{C} \pi_c\, \phi\big(z; \mu_c, \Sigma\big)}_{\text{mixture density}} + \text{const.} \tag{13}$$

第一项恰好是 LDA模型的负对数似然，而第二项则惩罚位于总体混合密度较高区域的点。直观地说，优化器会被鼓励将每个样本映射到自身类别密度较高、而其他类别密度较低的区域，从而减少高斯分量之间的重叠。

密度惩罚如何分离类别。 使用 LDA 头时，边缘嵌入密度是一个共享协方差高斯混合模型 $p_\theta(z) = \sum_{c=1}^{C} \pi_c \, \phi(z; \mu_c, \Sigma)$。除一个全局常数外，DNLL正则项等于在嵌入处计算的该混合密度，$R(z) \propto p_\theta(z)$（见式 (13)），因此，DNLL 会惩罚落在总体 模型密度较大区域中的训练嵌入。

更准确地说，将经验 DNLL 目标函数中的惩罚项写出，可得

$$\frac{\lambda}{n} \sum_{i=1}^{n} R(z_i) \;\propto\; \lambda \, \widehat{\mathbb{E}}\big[p_\theta(Z)\big] \;:=\; \frac{\lambda}{n} \sum_{i=1}^{n} p_\theta(z_i), \tag{14}$$

也就是说，DNLL 会最小化在训练嵌入上计算的混合密度平均值。当许多训练点落在多个混合分量同时赋予不可忽略概率质量的区域（高重叠）时，这个量会很大；因此，降低该量会抑制嵌入落入此类高重叠区域，从而促进分离。

衡量混合分布中重叠程度的一种便捷内在（分布层面）代理指标是信息势$C(p_\theta) = \int [p_\theta(z)]^2 \, dz = \mathbb{E}_{Z \sim p_\theta}\big[p_\theta(Z)\big]$，它衡量边缘密度 [7] 的自重叠。因此，DNLL 惩罚项 $\mathbb{E}[p_\theta(Z)]$可视为 $C(p_\theta)$ 的数据驱动类比：如果编码器诱导的嵌入边缘分布 $p(z)$ 接近模型混合分布 $p_\theta(z)$（似然项会促使这一点），那么 $\mathbb{E}_{Z \sim p}[p_\theta(Z)] \approx \mathbb{E}_{Z \sim p_\theta}[p_\theta(Z)] = C(p_\theta)$。（附录 A 中给出了关联这两个期望的定量界。）

对于共享协方差高斯混合模型，$C(p_\theta)$有闭式解：

$$C(p_\theta) = \sum_{i,j=1}^{C} \pi_i \pi_j \int \phi(z; \mu_i, \Sigma) \, \phi(z; \mu_j, \Sigma) \, dz = \sum_{i,j=1}^{C} \pi_i \pi_j \, \phi(\mu_i - \mu_j; 0, 2\Sigma), \tag{15}$$

等价地，

$$C(p_\theta) = (4\pi)^{-d/2} |\Sigma|^{-1/2} \sum_{i,j=1}^{C} \pi_i \pi_j \exp\Big(-\tfrac{1}{4} \|\mu_i - \mu_j\|^2_{\Sigma^{-1}} \Big). \tag{16}$$

(16) 中的分解明确揭示了重叠的影响。对角项 $i = j$ 贡献 $(4\pi)^{-d/2} |\Sigma|^{-1/2} \sum_i \pi_i^2$，它与均值无关，但会在 $\det \Sigma$ 趋近于零时发散。因此，任何与 $C(p_\theta)$相关的目标函数都会形成一道屏障，防止协方差坍缩。非对角项 $i \neq j$ 贡献 $(4\pi)^{-d/2} |\Sigma|^{-1/2} \pi_i \pi_j \exp(-\|\mu_i - \mu_j\|^2_{\Sigma^{-1}}/4)$，当$\mu_i$ 和 μ_j 彼此接近时，该项最大，并随它们之间的马氏距离增大而呈指数衰减。因此，最小化这些项会在不同类别中心之间产生明确的排斥，促使均值彼此充分分离。

最后，DNLL 会在数据所在的空间中应用这种减少重叠的机制。对于类别 y的一个训练嵌入 z，负对数似然项会促使 z 落在其所属成分 $\phi(\cdot; \mu_y, \Sigma)$的高密度区域，而惩罚项 $\lambda \, p_\theta(z)$则会阻止 z落入其他成分也具有不可忽略密度的区域。从几何角度看，这会引导嵌入远离决策边界区域，转向正确类别占主导的区域，从而减少各成分之间的重叠，同时保留生成式结构。

（15）–（16）的详细推导及补充讨论见附录 A。

3.2. 使用合成数据训练的 DNLL 深度线性判别分析

现在，我们重新审视第2节中的合成深度线性判别分析设置，并用 DNLL 损失 (11) 替换负对数似然目标 (8)。数据生成模型、编码器架构、优化器和训练计划均保持不变：我们在输入空间中使用相同的三类高斯混合，采用具有 32 个隐藏单元的双层 ReLU 网络作为编码器 f_ψ，并使用 Adam 在 20，000 个样本上训练 100 个训练轮数（训练时批量大小为 256，评估时为 1024）。正则化系数设为 $\lambda = .01$。

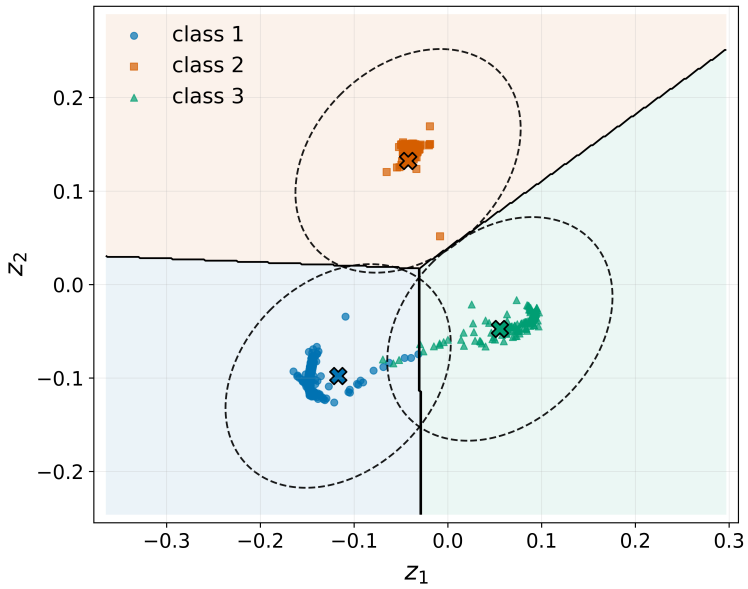


图 5：使用所提出的判别式负对数似然训练后得到的深度线性判别分析嵌入 $z_i = f_\psi(x_i)$。与纯似然训练（图 3）不同，DNLL 产生了分离良好的簇，高斯分量与经验类别云对齐，同时实现了近乎完美的训练和测试准确率：分别为 99.4% 和 99.5%。

图5展示了由此得到的嵌入 $z_i = f_\psi(x_i)$、学得的高斯分量以及相应的决策区域。与纯似然训练的病态行为（图3）形成鲜明对比的是，现在三个类别在潜在空间中形成了区分度更高的簇。类别均值 $\{\mu_c\}$ 位于经验样本云的中心附近。同时，分类性能与交叉熵训练相当：在这项合成任务中，训练和测试准确率均超过 99%。

这些结果表明，DNLL 成功协调了 LDA 头的生成与判别作用。似然项保留了模型的概率结构，并使参数接近底层高斯混合的参数；而密度惩罚则防止编码器利用退化的非判别式解。在本文余下部分，我们将通过在真实图像分类基准上评估这一损失，进一步分析它。

4. 图像分类实验

现在，我们在图像分类基准上评估使用第3节的判别式负对数似然训练的深度线性判别分析模型。与合成实验一样，该模型由一个编码器 f_ψ 和一个 LDA 头组成。编码器和分类头的参数通过最小化经验判别式负对数似然损失进行联合训练。我们将该模型与一个结构相同、配备线性 Softmax 头并使用交叉熵训练的编码器进行比较。

与合成数据设置的一个关键区别在于，对于图像实验，我们对 LDA 头施加几何约束，以确保稳定性，并与线性 Softmax 层进行公平比较。在主要配置中，我们使用 $d = C - 1$ 和球形协方差 $\Sigma = \sigma^2 I_d$。此外，我们还在表 1 中报告了针对协方差参数化方式（球形、对角（协方差）和完整（协方差））的消融实验。具体而言，我们使用：

- 在 $\mathbb{R}^d$ 中的嵌入，具有 $d = C - 1$，其中 C 是类别数量；
- 一个共享协方差矩阵 Σ，其参数化方式可以是：(i) 球形 $\sigma^2 I_d$（主要设定）；(ii) 对角 $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$；或 (iii) 完整（协方差） $\Sigma \succ 0$ （ Cholesky 参数化）。

从经典线性判别分析的角度来看，这是一个自然的选择：对于具有共享协方差的 C 类高斯模型，类间散布矩阵的秩至多为 $C-1$，而贝叶斯决策规则仅取决于投影到一个（$C-1$ 维）判别子空间上的结果。让编码器直接在 $\mathbb{R}^{C-1}$ 中生成嵌入，相当于端到端地学习这个判别子空间。

球形协方差进一步简化了分类头，并使其可训练参数的数量与线性 Softmax 层处于同一数量级：当 $\mu_1, \ldots\ldots, \mu_C \in \mathbb{R}^d$且 $d = C-1$时，两者都具有 $\mathcal{O}(Cd)$ 个参数。球形协方差使分类头保持轻量，而对角（协方差）和完整（协方差） Σ则以增加参数数量和潜在过拟合为代价，提高了灵活性。

在这些约束下，潜在空间中的线性判别分析判别函数形式为

$$\delta_c(z) \;=\; \log \pi_c - \frac{1}{2\sigma^2}\,\|z-\mu_c\|_2^2 \quad \text{for } z = f_\psi(x) \in \mathbb{R}^{C-1},$$

式（11）中的 DNLL 损失则简化为类条件平方距离与 $\mathbb{R}^{C-1}$ 中混合密度的组合。

***** 数据集 我们在三个标准图像分类数据集上进行评估。Fashion-MNIST [8] 包含 70 000 张大小为 28×28 的灰度图像，来自 $C = 10$ 个类别，其中 60 000 张用于训练，10 000 张为测试样本。CIFAR-10 和 CIFAR-100 [9] 分别包含 32×32 张彩色图像，具有 $C = 10$ 个和$C = 100$ 个类别，每个数据集均包含 50 000 张训练图像和 10 000 张测试图像。按照常见做法，我们对 CIFAR-10/100 应用标准数据增强（随机水平翻转，以及带填充的随机裁剪），而在 Fashion-MNIST 上仅使用逐像素归一化。

***** 编码器架构与潜在维度 所有实验均使用相同的卷积编码器，仅输入通道数不同。该编码器由三个卷积块组成，通道宽度为 $\{64, 128, 256\}$。每个卷积块包含两个 3×3 Conv–BN–ReLU 层；前两个卷积块后接 2×2最大池化。最后一个卷积块后通过自适应平均池化进行全局聚合，生成一个 256 维向量，然后将其投影到判别空间 $\mathbb{R}^d$，使用 $d = C-1$:

$$\begin{aligned}\mathrm{Conv}(C_{\text{in}}\to64) \to \mathrm{Conv}(64\to64) \to \mathrm{Pool} \to \mathrm{Conv}(64\to128)\\ \to \mathrm{Conv}(128\to128) \to \mathrm{Pool} \to \mathrm{Conv}(128\to256)\\ \to \mathrm{Conv}(256\to256) \to \mathrm{AdaptiveAvgPool} \to \mathrm{Linear}(256\to d),\end{aligned}$$

其中， $C_{\text{in}} = 1$ 适用于 Fashion-MNIST， $C_{\text{in}} = 3$适用于 CIFAR-10/100。所有实验均使用这一编码器架构，不作任何改动。对于 Softmax 基线，最后的线性层也输出 $d = C-1$ 个特征，随后将其输入线性分类器 $\mathbb{R}^d \to \mathbb{R}^C$。

***** 分类头与训练 对于每个数据集，我们比较基于相同编码器构建的模型 r:

1. 线性 Softmax 层 $Wz + b \in \mathbb{R}^C$，使用交叉熵训练。2. LDA 头，使用 DNLL 训练，并采用不同的协方差参数化：

- 球形： $\Sigma = \sigma^2 I_d$,
- 对角： $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots\ldots, \sigma_d^2)$,
- 完整： $\Sigma \succ 0$，通过 Cholesky 因子进行参数化。

我们对线性判别分析参数进行初始化，使潜在类别在开始时就能较好地彼此分离。先验设为 $\pi_c = 1/C$。每个类别均值 μ_c 均从 $\mathcal{N}\left(0,\ 6^2/(2d)\right)$中抽取，使这些均值构成一个各向同性云，其预期成对距离约为 6，从而确保不同类别在潜在空间中一开始就位于不同区域。共享方差设为 $\sigma^2 = 1$，以提供非退化的初始尺度。

***** 结果 表 1 报告了三个随机种子的平均测试准确率。在全部三个数据集上，使用 DNLL 训练的深度线性判别分析与 Softmax 基线相比表现不俗。此外，协方差消融表明，将协方差灵活性从球形提升到对角或完整（协方差）并未带来

稳定的准确率提升：对角（协方差）和完整（协方差） Σ 的表现与球形（协方差） Σ相近。这表明，DNLL 的主要优势源自似然与分离之间的相互作用，而非高表达能力的协方差模型；轻量级球形分类头则是稳健的默认选择。我们在后续实验中采用该分类头。

表 1：图像分类：测试准确率（3 次运行的均值 $\pm$ 2标准差）。

Head	Loss	Fashion-MNIST	CIFAR-10	CIFAR-100
Softmax	交叉熵	93.61 ± 0.76	89.07 ± 0.40	64.67 ± 0.60
深度线性判别分析（球形 Σ）	DNLL	92.82 ± 0.30	90.33 ± 0.56	66.09 ± 1.30
深度线性判别分析（对角（协方差） Σ）	DNLL	92.84 ± 0.52	90.09 ± 0.68	65.95 ± 0.62
深度线性判别分析（完整（协方差） Σ）	DNLL	93.00 ± 0.46	90.24 ± 0.48	65.51 ± 0.42

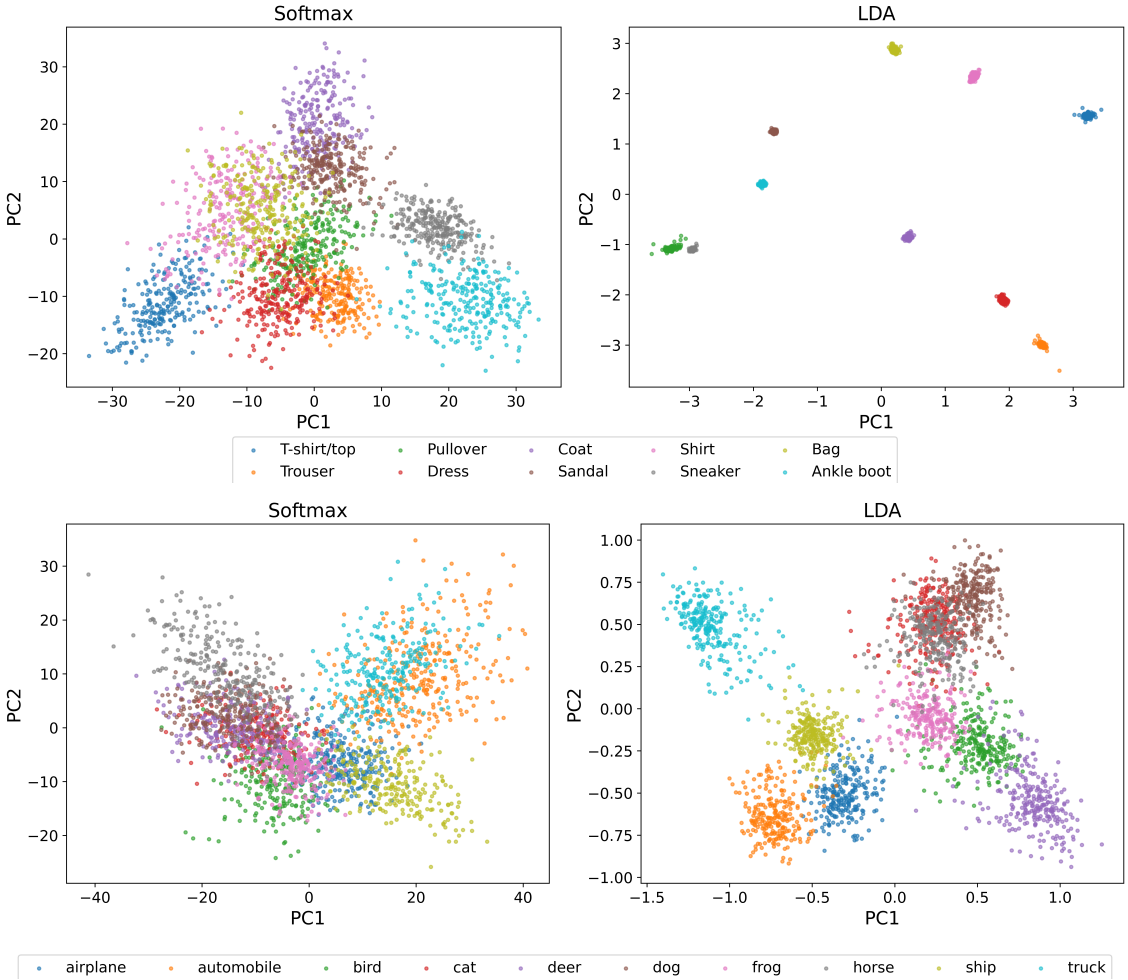


图6：使用两个分类头学习得到的 Fashion-MNIST（上）和 CIFAR-10（下）嵌入的主成分分析投影。即使在二维投影中，深度线性判别分析也能生成更紧凑、分离效果更好的簇。

***** 嵌入几何。为了探究深度线性判别分析所诱导的几何结构，我们将模型在 Fashion-MNIST 和 CIFAR-10 上学到的嵌入可视化。训练完成后，我们收集编码器输出 $z_i = f_\psi(x_i) \in \mathbb{R}^{C-1}$，这些输出来自随机选取的训练样本，并使用主成分分析将其投影到二维空间。得到的散点图如图6所示。在 softmax 头下，投影后的簇

明显依赖于类别，但存在部分重叠。相比之下，使用 DNLL 训练的深度线性判别分析模型形成的簇更紧密，即使在二维投影中也能保持更好的分离。

这些发现与第2节和第3节的合成实验一致：DNLL 会促使嵌入集中在自身类别密度较高、混合密度不过高的区域，从而在不牺牲分类准确率的情况下，得到清晰且易于解释的潜在表示。

对 DNLL 权重的敏感性 λ（准确率）。本文中的所有深度线性判别分析实验都使用固定的 DNLL 权重 $\lambda = 0.01$。为评估该方法对这一选择是否敏感，我们在 CIFAR-100 上对跨越多个数量级的 λ 进行遍历，并报告每个 λ 取值下 5 个随机种子的平均测试准确率的 95%置信区间（图 7）。测试准确率变化

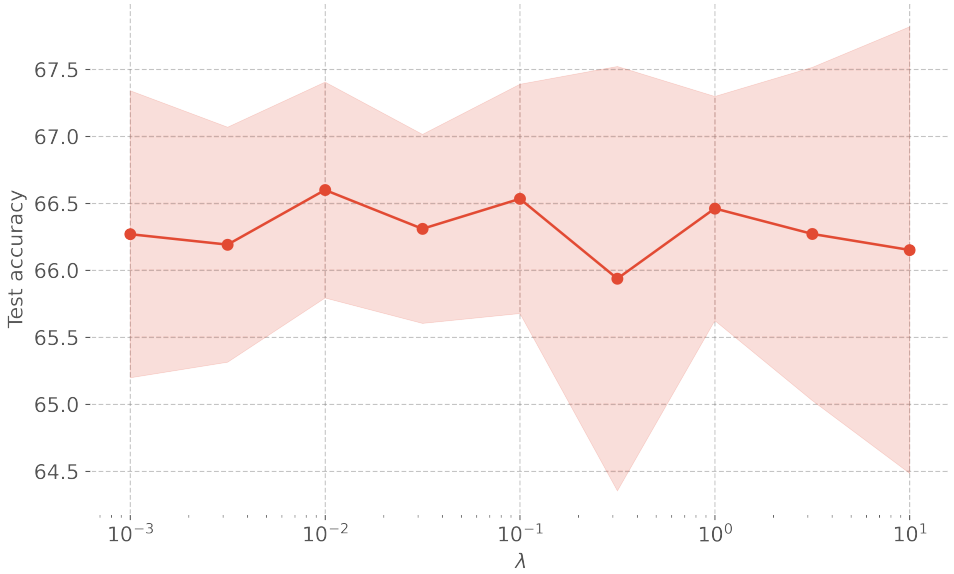


图 7：CIFAR-100 上深度线性判别分析测试准确率对正则化权重 λ 的敏感性（每个 λ 取值运行 5 次的均值和标准差 ± 2）。准确率在跨越多个数量级的范围内保持稳定，这支持我们始终使用固定的 $\lambda = 0.01$。

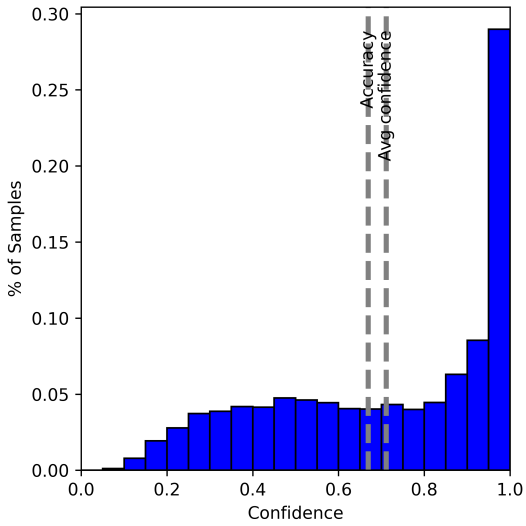
在整个扫描过程中仅略有变化，这表明 DNLL 无需对 λ 进行精细调参，也能获得具有竞争力的分类性能。因此，为了简便并便于比较，我们在所有实验中都使用 $\lambda = 0.01$。

附录C报告了在 CIFAR-10 和 Fashion-MNIST 上进行的其他 λ 扫描，结果显示了相同的定性行为。

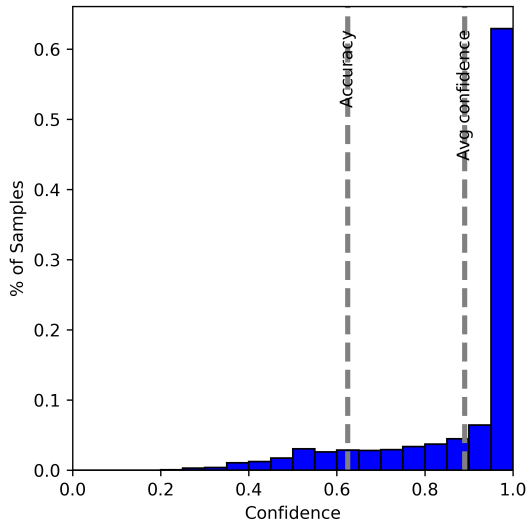
4.1. 校准与置信度行为

除了分类准确率，我们还评估预测概率的校准 情况，即预测置信度是否与经验正确率相符。对于测试样本 x，令 $\hat{y}(x) = \arg\max_c p_\theta(c \mid x)$，并令相应的置信度 为 $\hat{p}(x) = \max_c p_\theta(c \mid x)$。置信度直方图展示了测试集上 $\hat{p}(x)$ 的分布；它揭示了模型是否倾向于给出高度自信的预测（例如，$\hat{p}(x) \approx 1$），或产生较为温和的概率。

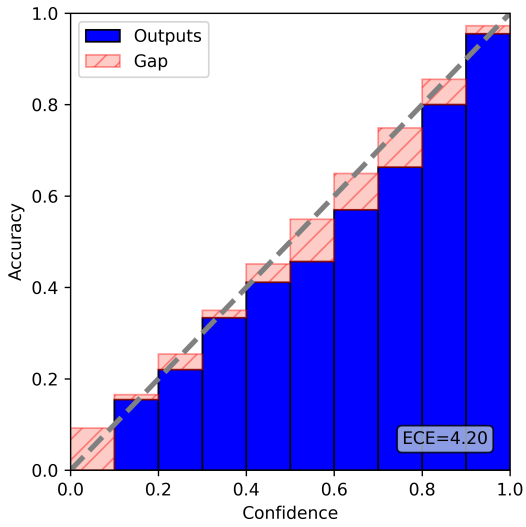
为量化校准情况，我们使用可靠性图 和 [10]的期望校准误差（*ECE*）。可靠性图按置信度对测试点进行分箱，例如，$B_m = \{x : \text{对于置信度区间} I_m,\ \hat{p}(x) \in I_m\}$，并绘制每个分箱的经验准确率 $\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{x \in B_m} \mathbb{I}\{\hat{y}(x) = y\}$ 与平均置信度 $\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{x \in B_m} \hat{p}(x)$ 的对比。完美校准对应于 $\mathrm{acc}(B_m) \approx$



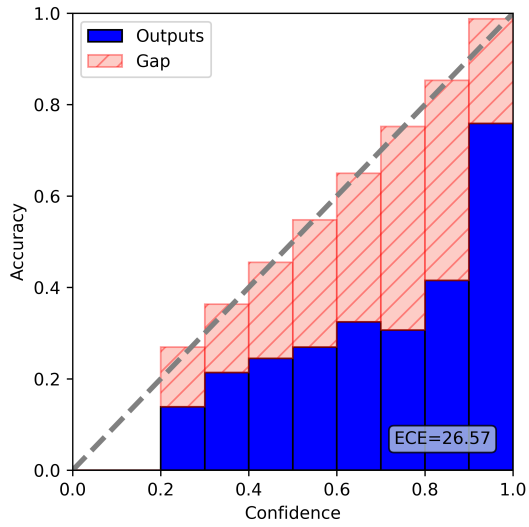
(a) LDA 头



(b) Softmax 头



(c) LDA 头



(d) Softmax 头

图 8：**CIFAR-100 上的校准与置信度行为**。上排：最大预测置信度的直方图，虚线表示平均置信度和测试准确率。下排：可靠性图，阴影部分表示校准误差，并标出了期望校准误差（ECE）。尽管 Softmax 分类器表现出严重的过度自信和较差的校准效果（ECE= 26.6%），采用 DNLL 训练的基于 LDA 的生成式头能产生校准效果显著更好的预测（ECE = 4.2%），且分类准确率相当。

置信度(B_m)在各分箱中的表现（位于对角线上的点）。期望校准误差总结了整体偏差 a s

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{17}$$

其中， n是测试样本数， M 是分箱数（我们使用 $M = 10$个等宽分箱）。

图 8 比较了在 CIFAR-100 上使用 DNLL 训练的基于 LDA 的生成式头与标准 Softmax 分类器。尽管两种模型的分类准确率相近，但其置信度行为存在显著差异。softmax头表现出严重的过度自信：尽管经验准确率低得多，却仍对很大比例的样本赋予接近 1 的置信度，因此期望校准误差达到 26.6%。相比之下，LDA 头产生的置信度分布更宽，置信度与准确率之间的吻合度也显著提高，期望校准误差为 4.2%。

附录C 给出了 CIFAR-10 和 Fashion-MNIST 上相同的置信度直方图和可靠性图。LDA 头的校准效果始终优于 Softmax （ CIFAR-10 上期望校准误差为 3.08%，而 7.73% ； Fashion-MNIST 上为 2.98%，而 5.05% ）。

这些结果表明，在分类头中保留显式的生成式结构，即使在标准监督分类设置下，也能显著改善预测概率的校准效果。

对 λ （校准）的敏感性。 我们使用与图 7 相同的 CIFAR-100 参数扫描，进一步考察校准效果如何随 λ变化。图 9 展示了 ECE 随 λ变化的情况；在很宽的取值范围内，校准效果始终稳定且良好。这表明，DNLL 带来的校准收益并不依赖于对 λ进行精细调节。

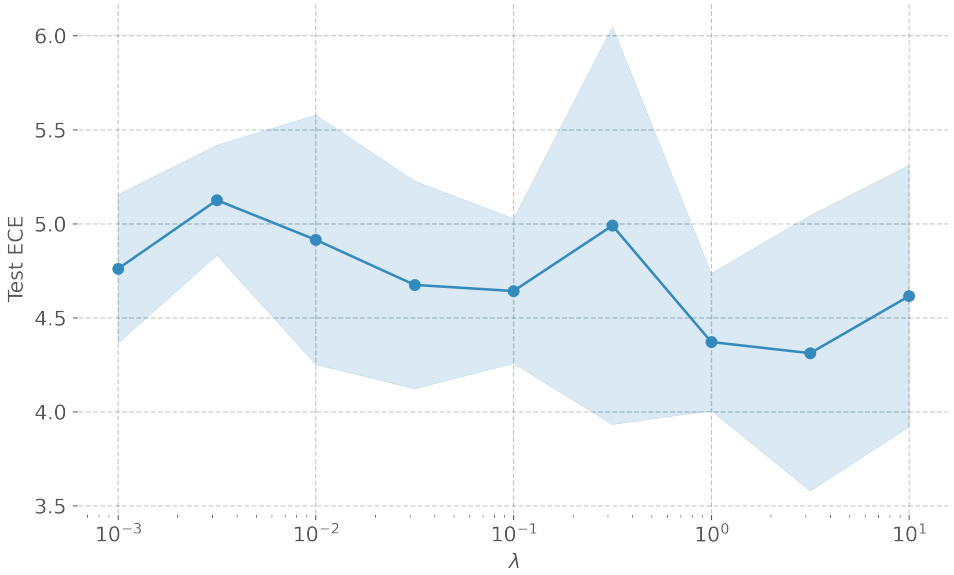


图 9：CIFAR-100 上校准对 DNLL 权重 λ 的敏感性，以期望校准误差（ECE）衡量（每个 λ取值运行 5 次所得的均值 $\pm$ 2 标准差）。在跨越多个数量级的范围内，校准效果保持稳定。

4.2. DNLL 权重 λ 控制什么?

图7和图9显示，在 CIFAR-100 上，测试准确率和校准（期望校准误差）对 DNLL 权重 λ 在跨越多个数量级的变化中都极不敏感。这自然引出了一个问题：如果 λ不会改变准确率或期望校准误差，那么它究竟会影响什么？

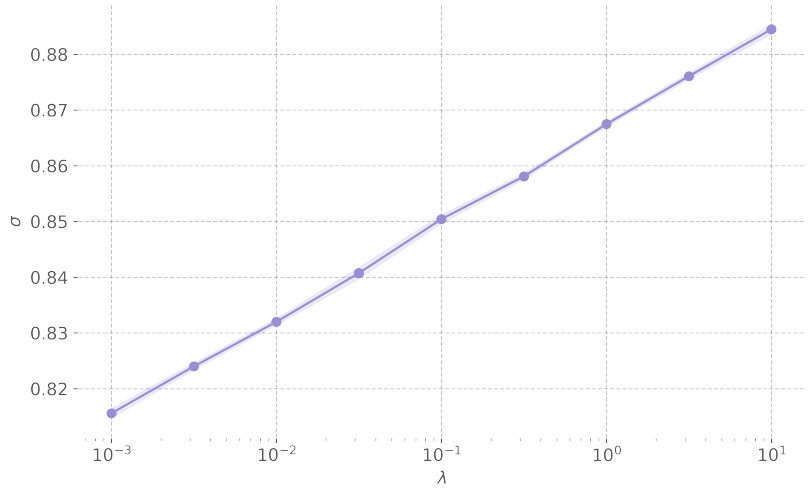


图10：在 CIFAR-100 上，学习到的球形协方差尺度 σ 与 DNLL 权重 λ 的关系（扫描方案与图6和图8相同）。虽然准确率和期望校准误差几乎不受 λ影响，但学习到的 σ 会系统性地变化：减小 λ 会使 σ 变小（类别条件分布更尖锐）。

基于梯度的解释。回顾 DNLL 损失 (11) 的定义。令 $\delta(z) = (\delta_1(z), \ldots, \delta_C(z))^\top$，并定义逐元素的 $\exp(\delta(z))$。直接求导可得关于判别值向量的梯度：

$$\nabla_\delta \, \ell_{\mathrm{DNLL}}(z, y) \;=\; \lambda \, \exp\big(\delta(z)\big) \;-\; e_y, \tag{18}$$

其中，e_y 是 $\mathbb{R}^C$中的第 y 个标准基向量。等价地，按分量表示为：

$$\frac{\partial \ell_{\mathrm{DNLL}}(z, y)}{\partial \delta_c(z)} = \begin{cases} -1 + \lambda \exp(\delta_y(z)), & c = y, \\ \lambda \exp(\delta_c(z)), & c \neq y. \end{cases} \tag{19}$$

因此，基于梯度的优化会推动 $\lambda\exp(\delta_y(z))$趋近于 1（对于真实类别），并推动 $\lambda\exp(\delta_c(z))$趋近于 0（对于竞争类别）。尤其是，对于典型的训练嵌入，预期 $\exp(\delta_y(z))$的量级为 $1/\lambda$。

为何这会针对球形分类头中的协方差尺度。对于线性判别分析分类头，

$$\exp\big(\delta_c(z)\big) = \pi_c \, \varphi(z; \mu_c, \Sigma) \cdot (2\pi)^{d/2}, \tag{20}$$

因此，$\exp(\delta_c(z))$与在 z 处取值的类别条件高斯密度成正比。在我们的主要图像分类配置中，我们使用球形共享协方差 $\Sigma = \sigma^2 I_d$。在这种参数化下，全局控制 $\varphi(z; \mu_c, \sigma^2 I_d)$“尖峰程度”（整体尺度）的唯一 标量自由度是 σ^2。因此，（18）所提示的驻点条件 $\exp(\delta_y(z)) \approx 1/\lambda$预示着，减小 λ 时，学习到的 σ也应随之减小（类别条件分布更尖锐）；而增大 λ 则应允许 σ 增大（类别条件分布更宽缓）。

实证验证。为了检验这一假设，我们重新进行了图7和图9中使用的同一组 CIFAR-100 λ 实验，并记录了训练结束时 σ （球形分类头）的最终学习值。图10展示了 σ 随 λ 变化的情况（对多次运行取均值）。与准确率和期望校准误差不同，协方差尺度表现出对 λ明显的单调依赖关系：较小的 λ 总是会产生较小的 σ （更尖锐的高斯分布），而较大的 λ 则会产生较大的 σ （更宽的高斯分布）。这解释了为什么 λ 可以大幅变化，却不影响分类指标：在这一设定下，λ 主要通过 σ充当生成式头的尺度/温度调节旋钮，而不是显著改变所诱导决策规则的参数。

该 在 CIFAR-10 和 Fashion-MNIST（附录C).

5. 结论

我们从基于似然的训练视角重新审视了深度线性判别分析，并展示了无约束的最大似然优化会产生病态解：类别簇彼此重叠，而纯粹的判别式交叉熵训练则破坏了 LDA 头的概率解释。为调和这两种视角，我们提出了判别式负对数似然（DNLL），在 LDA 对数似然中加入密度惩罚，明确抑制多个类别同时具有高概率的区域。在合成数据和标准图像基准数据集上，采用 DNLL 的深度线性判别分析能够生成清晰且分离良好的潜在类别簇，其测试准确率可与 Softmax 分类器相媲美，同时还能产生显著更优的校准后的预测概率。

我们的敏感性分析进一步表明，DNLL 无需对正则化权重 λ进行精细调节：在跨越数个数量级的范围内，准确率和校准（期望校准误差，ECE）均保持稳定。同时，λ并非无关紧要：在球形头设置 $\Sigma = \sigma^2 I$下，改变 λ会系统性地改变学习到的协方差尺度 σ^2（较小的 λ 会导致较小的 σ，即更尖锐的高斯类别条件分布），即使下游指标基本保持不变也是如此。这表明，在实践中，λ 主要充当生成式头的尺度/温度调节旋钮，用于控制拟合得到的类别密度的集中程度，而不是强烈影响所诱导分类器的参数。

这些发现进一步印证了本文的核心观点：简单的生成式结构与显式的判别式损失相结合，可以恢复基于似然的训练在深度判别模型中的可行性，同时生成有意义的预测概率。未来一个值得探索的方向是将这种“尺度”效应显式化，例如，通过重新参数化 DNLL，将 λ吸收到温度（或方差）参数中；或者设计自适应方案，在训练过程中以期望的协方差尺度或置信度分布为目标。

可复现性用于复现我们实验的代码已发布于 https://github.com/zh3nis/ DNLL。

人工智能辅助我们使用了基于人工智能的工具对文本进行编辑润色；作者对内容承担全部责任。

参考文献

[1] Ronald A Fisher. 分类学问题中多重测量的应用。《优生学年鉴》，7(2):179–188，1936。[2] C. Radhakrishna Rao. 生物分类问题中多重测量的应用。《皇家统计学会学报 *B*辑（方法论）》，10(2):159–203，1948。ISSN 00359246。URL http://www.jstor.org/stable/2983775。[3] Sebastian Mika、Gunnar Ratsch、Jason Weston、Bernhard Scholkopf 和 Klaus-Robert Mullers。基于核的 Fisher 判别分析。载于《用于信号处理的神经网络 *IX*：*1999*年 *IEEE* 信号处理学会研讨会论文集（目录号 *98th8468*）》，第41–48页。IEEE，1999。[4] Trevor Hastie 和 Robert Tibshirani。基于高斯混合的判别分析。《皇家统计学会学报 *B*辑：统计方法论》，58(1):155–176，1996。[5] Matthias Dorfer、Rainer Kelz 和 Gerhard Widmer。深度线性判别分析。载于 Yoshua Bengio 和 Yann LeCun（编），《第*4*届国际学习表征会议（*ICLR 2016*），波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议专题论文集》，2016。URL http://arxiv.org/abs/1511.04707。[6] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Bengio 和 Yann LeCun（编），《第*3*届国际学习表征会议，*ICLR*》

*2015*年，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议专题论文集，2015。URL http:// arxiv.org/abs/1412.6980。[7] Ana-Maria Acu、Gülen Başcanbaz-Tunca 和 Ioan Rasa。若干概率密度函数的信息势。《应用数学与计算》，389:125578，2021。ISSN 0096-3003。[8] Han Xiao、Kashif Rasul 和 Roland Vollgraf。Fashion-MNIST：一个用于机器学习算法基准测试的新型图像数据集。*arXiv* 预印本 *arXiv:1708.07747*，2017。[9] Alex Krizhevsky、Geoffrey Hinton 等。学习微小图像中的多层特征。2009。[10] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。现代神经网络的校准。载于《国际机器学习会议》，第1321–1330页。PMLR，2017。

A. 线性判别分析分类头的 DNLL 密度惩罚分析

本附录对论文中使用的线性判别分析分类头进行了详细的信息势分析。我们 (i) 推导了线性判别分析边缘分布$p_\theta(z)$的信息势的闭式表达式，(ii) 将所得项解释为类别均值之间的排斥作用，以及防止协方差坍缩的屏障，并且 (iii) 通过 KL 控制界，将经验 DNLL 密度惩罚 $\mathbb{E}_{Z\sim p}[p_\theta(Z)]$ 与内在信息势 $\int p_\theta(z)^2 dz$ 联系起来。我们还记录了一个球形特例 $\Sigma = \sigma^2 I_d$。

A.1. 符号说明与嵌入空间中的线性判别分析混合模型

在正文中，编码器 f_ψ 将输入映射为嵌入

$$z = f_\psi(x) \in \mathbb{R}^d.$$

线性判别分析分类头是一个共享协方差的高斯类条件模型

$$p_\theta(z \mid y = c) = \mathcal{N}(z; \mu_c, \Sigma), \qquad p_\theta(y = c) = \pi_c, \qquad \Sigma \succ 0, \quad \pi_c > 0, \quad \sum_{c=1}^{C} \pi_c = 1,$$

其中 $\theta = (\pi,\ \{\mu_c\}_{c=1}^C,\ \Sigma)$。论文中使用的判别分数为

$$\delta_c(z) = \log p_\theta(z, y = c) = \log \pi_c + \log \mathcal{N}(z; \mu_c, \Sigma).$$

嵌入空间中的诱导边缘（混合）密度为

$$p_\theta(z) = \sum_{c=1}^{C} \pi_c \, \varphi_c(z), \qquad \varphi_c(z) = \mathcal{N}(z; \mu_c, \Sigma). \tag{21}$$

注意，$\sum_{c=1}^C e^{\delta_c(z)} = \sum_c p_\theta(z,\ c) = p_\theta(z)$，因此正文中的 DNLL 惩罚项（忽略常数后）就是在嵌入处计算的混合密度。

A.2. LDA混合模型的信息势

定义信息势

$$C(p_\theta) := \int_{\mathbb{R}^d} p_\theta(z)^2 \, dz. \tag{22}$$

定理 1（LDA混合模型的信息势（共享协方差））。设 $p_\theta(z)$ 为 *LDA*混合模型（21）。则

$$C(p_\theta) = \sum_{i,j=1}^{C} \pi_i \pi_j \, \mathcal{N}(\mu_i - \mu_j; 0, 2\Sigma). \tag{23}$$

等价地，

$$C(p_\theta) = (4\pi)^{-d/2} \, (\det \Sigma)^{-1/2} \sum_{i,j=1}^{C} \pi_i \pi_j \, \exp\Big(-\tfrac{1}{4} (\mu_i - \mu_j)^\top \Sigma^{-1} (\mu_i - \mu_j) \Big). \tag{24}$$

备注 1　（对角项与非对角项：排斥与坍缩）。 中的对角项 (23) *a* *re*

$$\pi_i^2 \, \mathcal{N}(0; 0, 2\Sigma) = (4\pi)^{-d/2} (\det \Sigma)^{-1/2} \, \pi_i^2,$$

它们不依赖于 μ_i，但当$\det \Sigma \downarrow 0$ 时会发散 。因此，对 $C(p_\theta)$ 进行惩罚可以抑制协方差坍缩。对于 $i \neq j$，(i, j) 项等于

$$\pi_i \pi_j \, (4\pi)^{-d/2} (\det \Sigma)^{-1/2} \, \exp\Big(-\tfrac{1}{4} \|\mu_i - \mu_j\|^2_{\Sigma^{-1}} \Big),$$

它具有排斥性：随着类别中心之间的马氏距离增大，它会呈指数衰减。因此，惩罚信息势可以解释为在类别均值之间引入排斥势（同时也惩罚 Σ的行列式过小）。

定理 *1* 的证明。由线性性可得,

$$\int p_\theta(z)^2\,dz = \int \Big(\sum_{i=1}^{C}\pi_i\varphi_i(z)\Big)\Big(\sum_{j=1}^{C}\pi_j\varphi_j(z)\Big)\,dz = \sum_{i,j=1}^{C}\pi_i\pi_j\int \varphi_i(z)\varphi_j(z)\,dz.$$

因此，我们需要计算

$$I_{ij} := \int_{\mathbb{R}^d}\varphi_i(z)\varphi_j(z)\,dz, \qquad \varphi_i(z)=\mathcal{N}(z;\mu_i,\Sigma),\ \varphi_j(z)=\mathcal{N}(z;\mu_j,\Sigma).$$

步骤 1：化简为零点处的卷积。定义反射密度

$$\tilde{\varphi}_j(z) := \varphi_j(-z).$$

然后，通过变量替换,

$$(\varphi_i * \tilde{\varphi}_j)(0) = \int_{\mathbb{R}^d}\varphi_i(t)\tilde{\varphi}_j(-t)\,dt = \int_{\mathbb{R}^d}\varphi_i(t)\varphi_j(t)\,dt = I_{ij}.$$

因此,

$$I_{ij} = (\varphi_i * \tilde{\varphi}_j)(0). \tag{25}$$

步骤 2：高斯分布的反射。若 $\varphi_j(z)=\mathcal{N}(z;\ \mu_j,\ \Sigma)$，则

$$\tilde{\varphi}_j(z) = \varphi_j(-z) = \mathcal{N}(z;-\mu_j,\Sigma),$$

因为将 z 替换为 $-z$ 会改变均值的符号，但协方差保持不变。

St步骤 3：高斯分布的卷积。一个标准恒等式指出，对于高斯密度 s,

$$\mathcal{N}(\cdot;\mu_1,\Sigma_1) * \mathcal{N}(\cdot;\mu_2,\Sigma_2) = \mathcal{N}(\cdot;\mu_1+\mu_2,\Sigma_1+\Sigma_2).$$

将其应用于 $\varphi_i=\mathcal{N}(\cdot;\ \mu_i,\ \Sigma)$ 和 $\tilde{\varphi}_j=\mathcal{N}(\cdot;\ -\mu_j,\ \Sigma)$，得到

$$(\varphi_i * \tilde{\varphi}_j)(z) = \mathcal{N}(z;\mu_i-\mu_j,2\Sigma).$$

步骤 4：在 $z=0$ 处评估并利用对称性。在 $z=0$ 处求值，得到

$$I_{ij} = (\varphi_i * \tilde{\varphi}_j)(0) = \mathcal{N}(0;\mu_i-\mu_j,2\Sigma).$$

由高斯密度的对称性可知,

$$\mathcal{N}(0;\mu_i-\mu_j,2\Sigma) = \mathcal{N}(\mu_i-\mu_j;0,2\Sigma).$$

将 I_{ij} 代入展开式，即完成了 (23) 的推导。展开高斯密度可得 (24)。

□

A.3. 我们的损失：内在重叠与经验重叠

在总体形式下，考虑以下两个目标（以嵌入空间中的形式表示）：

$$I_1(\theta) = -\mathbb{E}_{(Z,Y)\sim p}\big[\log p_\theta(Z,Y)\big] + \lambda\int_{\mathbb{R}^d} p_\theta(z)^2\,dz, \tag{26}$$

$$I_2(\theta) = -\mathbb{E}_{(Z,Y)\sim p}\big[\log p_\theta(Z,Y)\big] + \lambda\int_{\mathbb{R}^d} p(z)\,p_\theta(z)\,dz, \tag{27}$$

其中， p 表示编码器所诱导的 $(Z,\ Y)$ 的（未知）数据分布，而 $p(z)$ 表示其在嵌入空间中的边缘分布。训练中使用的 DNLL 惩罚项是对 $\int p(z)p_\theta(z)\,dz$ 的经验近似：给定训练集 $\{(z_n,\ y_n)\}_{n=1}^N$,

$$\int p(z)p_\theta(z)\,dz \approx \frac{1}{N}\sum_{n=1}^{N}p_\theta(z_n) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}p_\theta(z_n,c) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}e^{\delta_c(z_n)}.$$

用 KL散度 重写。假设 $p(z)$ 是一个密度，且 $p(y \mid z)$ 是一个条件概率质量函数。忽略一个加性常数（即 $p(z)$ 的微分熵与 $p(y \mid Z)$ 的期望香农熵之和），负对数似然项可以写成 KL散度。具体而言，

$$-\mathbb{E}_{(Z,Y)\sim p}\big[\log p_\theta(Z,Y)\big] = \mathrm{KL}\big(p(z,y)\,\|\,p_\theta(z,y)\big) + \text{const.}$$

因此，除一个加法常数外，(26)–(27) 等价于

$$L_1(\theta) = \mathrm{KL}\big(p(z,y)\,\|\,p_\theta(z,y)\big) + \lambda \int p_\theta(z)^2\,dz, \tag{28}$$

$$L_2(\theta) = \mathrm{KL}\big(p(z,y)\,\|\,p_\theta(z,y)\big) + \lambda \int p(z)\,p_\theta(z)\,dz. \tag{29}$$

我们要回答： 何时最小化 L_2 也会使 L_1达到最小?

A.4. 通过 KL散度得到的一般重叠界

命题 2 (通过 KL散度控制重叠差)。 设 p 和 p_θ 是定义在 $\mathbb{R}^d$上的概率密度。假设 $p_\theta \in L^\infty(\mathbb{R}^d)$，其中 $\|p_\theta\|_\infty < \infty$。则

$$\left|\int_{\mathbb{R}^d} p(z)\,p_\theta(z)\,dz - \int_{\mathbb{R}^d} p_\theta(z)^2\,dz\right| \le \|p_\theta\|_\infty \sqrt{2\,\mathrm{KL}(p(z)\|p_\theta(z))}. \tag{30}$$

证明。 我们首先将差写成

$$\int p(z)p_\theta(z)\,dz - \int p_\theta(z)^2\,dz = \int (p(z) - p_\theta(z))\,p_\theta(z)\,dz.$$

由霍尔德不等式,

$$\left|\int (p - p_\theta)p_\theta\right| \le \|p - p_\theta\|_1\,\|p_\theta\|_\infty = 2\,\mathrm{TV}(p,p_\theta)\,\|p_\theta\|_\infty,$$

其中 $\mathrm{TV}(p,p_\theta) = \frac{1}{2}\|p - p_\theta\|_1$。应用平克斯不等式,

$$\mathrm{TV}(p,p_\theta) \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(p\|p_\theta)},$$

我们得到 (30)。 □

A.5. 对线性判别分析混合模型的特化：显式常数

命题 3 （线性判别分析分类头的显式重叠界）。 设 $p_\theta(z) = \sum_{c=1}^{C} \pi_c \mathcal{N}(z;\,\mu_c,\Sigma)$ 满足$\Sigma \succ 0$。则

$$\|p_\theta\|_\infty \le (2\pi)^{-d/2}(\det\Sigma)^{-1/2}, \tag{31}$$

因此

$$\left|\int p(z)p_\theta(z)\,dz - \int p_\theta(z)^2\,dz\right| \le (2\pi)^{-d/2}(\det\Sigma)^{-1/2}\,\sqrt{2\,\mathrm{KL}(p(z)\|p_\theta(z))}. \tag{32}$$

证明。 对于每个高斯分量,

$$\phi_c(z) := \mathcal{N}(z;\mu_c,\Sigma) = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}\exp\Big(-\tfrac{1}{2}(z-\mu_c)^\top\Sigma^{-1}(z-\mu_c)\Big),$$

因此，$\sup_z \phi_c(z) = \phi_c(\mu_c) = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}$。利用 $\exp(\cdot) \le 1$ 和 $\sum_c \pi_c = 1$，我们对所有 z 得到:

$$p_\theta(z) = \sum_{c=1}^{C} \pi_c\phi_c(z) \le \sum_{c=1}^{C} \pi_c(2\pi)^{-d/2}(\det\Sigma)^{-1/2} = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}.$$

这就证明了 (31)。代入命题 2 得到 (32)。 □

备注 2 （该失效模式与协方差坍缩相符）。 式 $(\det\Sigma)^{-1/2}$ 中的因子可能随着 $\det\Sigma \downarrow 0$而发散。在这种情况下，通过这个界最小化 L_2 可能不再意味着最小化 L_1。这正是无约束最大似然训练中观察到的协方差坍缩问题。

A.6. KL 散度的边缘化

接下来，我们陈述 KL 散度在边缘化下的标准单调性，它将 KL($p(z)\|p_\theta(z)$) 与式 (28)–(29) 中出现的联合 KL 散度联系起来。

定理 4（KL 散度在边缘化下的单调性）。 设 $Z \in \mathbb{R}^d$ 为连续变量，$Y \in \{1, \dots, C\}$ 为离散变量。假设 $p(z, y)$ 和 $q(z, y)$ 可以分解为

$$p(z, y) = p(z)\, p(y \mid z), \qquad q(z, y) = q(z)\, q(y \mid z),$$

具有条件概率质量函数。若 $p(z, y) \ll q(z, y)$，则

$$\mathrm{KL}\big(p(z) \,\|\, q(z)\big) \le \mathrm{KL}\big(p(z, y) \,\|\, q(z, y)\big).$$

证明。 根据定义，

$$\mathrm{KL}\big(p(z, y) \,\|\, q(z, y)\big) = \int_{\mathbb{R}^d} \sum_{y=1}^{C} p(z, y) \log \frac{p(z, y)}{q(z, y)}\, dz.$$

利用这些分解式，

$$\log \frac{p(z, y)}{q(z, y)} = \log \frac{p(z)}{q(z)} + \log \frac{p(y \mid z)}{q(y \mid z)}.$$

代入并拆分各项，

$$\mathrm{KL}\big(p(z, y) \,\|\, q(z, y)\big) = \int_{\mathbb{R}^d} \sum_{y} p(z, y) \log \frac{p(z)}{q(z)}\, dz + \int_{\mathbb{R}^d} \sum_{y} p(z, y) \log \frac{p(y \mid z)}{q(y \mid z)}\, dz.$$

第一项可简化，因为 $\log \frac{p(z)}{q(z)}$ 不依赖于 y：

$$\int_{\mathbb{R}^d} \sum_{y} p(z, y) \log \frac{p(z)}{q(z)}\, dz = \int_{\mathbb{R}^d} p(z) \log \frac{p(z)}{q(z)}\, dz = \mathrm{KL}\big(p(z) \,\|\, q(z)\big).$$

第二项等于

$$\mathbb{E}_{Z \sim p}\Big[\mathrm{KL}\big(p(\cdot \mid Z) \,\|\, q(\cdot \mid Z)\big)\Big] \ge 0,$$

因为离散分布之间的 KL散度非负。因此，

$$\mathrm{KL}\big(p(z, y) \,\|\, q(z, y)\big) \ge \mathrm{KL}\big(p(z) \,\|\, q(z)\big).$$

□

A.7. 最小化 L_2 时， L_1也会减小

根据定理 4，对于我们的模型 $q = p_\theta$，

$$\mathrm{KL}\big(p(z) \,\|\, p_\theta(z)\big) \le \mathrm{KL}\big(p(z, y) \,\|\, p_\theta(z, y)\big) \le L_2(\theta).$$

结合命题 3，我们得到

$$\begin{aligned} \int p_\theta(z)^2\, dz &\le \int p(z) p_\theta(z)\, dz + (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \sqrt{2\, \mathrm{KL}\big(p(z) \| p_\theta(z)\big)} \\ &\le \int p(z) p_\theta(z)\, dz + (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \sqrt{2\, L_2(\theta)}. \end{aligned} \tag{33}$$

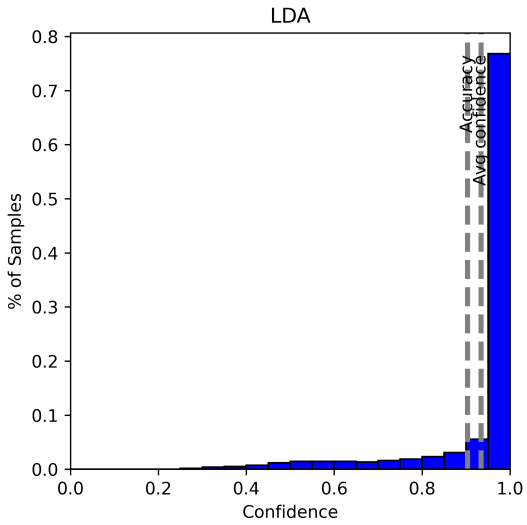
乘以 λ并加上共享的 KL项，即可得到明确的比较

$$\begin{aligned} L_1(\theta) \le L_2(\theta) + \lambda (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \sqrt{2\, \mathrm{KL}\big(p(z) \| p_\theta(z)\big)} \\ \le L_2(\theta) + \lambda (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \sqrt{2\, L_2(\theta)}. \end{aligned} \tag{34}$$

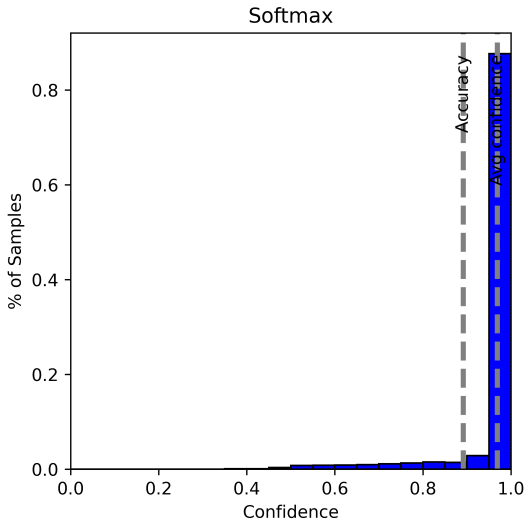
注释 3（同样的注意事项，现在以线性判别分析的形式呈现）。 公式 (34) 表明，随着 $L_2(\theta)$ 减小，内在信息势惩罚目标 $L_1(\theta)$ 也会减小，前提是 $(\det \Sigma)^{-1/2}$ 保持受控。若$\det \Sigma \downarrow 0$，该因子会急剧增大，使界限变得毫无意义；此时，通过这一论证最小化 L_2 可能无法跟踪 L_1。这与无约束似然训练的退化现象一致，也说明有必要明确防止协方差坍缩（例如采用球形参数化，和/或在第*4*节中对σ^2设置下界）。在球形情形下 $\Sigma = \sigma^2 I_d$，该发散因子变为 σ^{-d}。

B. CIFAR-10 和 Fashion-MNIST 上的其他校准结果

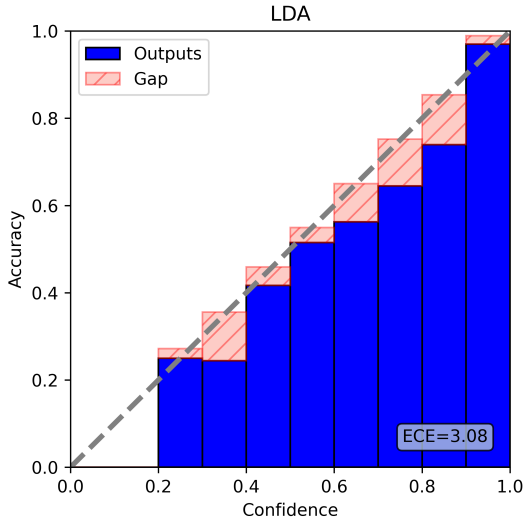
我们按照第 4.1 节的流程，报告了 CIFAR-10 和 Fashion-MNIST 的额外置信度直方图和可靠性图（含期望校准误差）。在这两个数据集上，使用 DNLL 训练的线性判别分析分类头比 Softmax 基线校准得更好（CIFAR-10 上期望校准误差为 3.08% 对 7.73%；Fashion-MNIST 上为 2.98% 对 5.05%）。



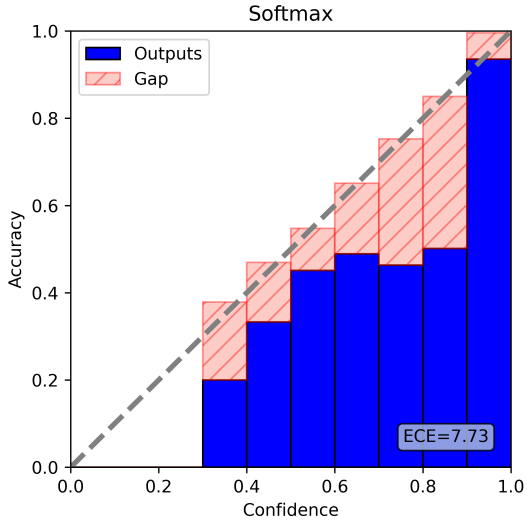
(a) CIFAR-10：置信度直方图



(b) CIFAR-10：置信度直方图



(c) CIFAR-10：可靠性图

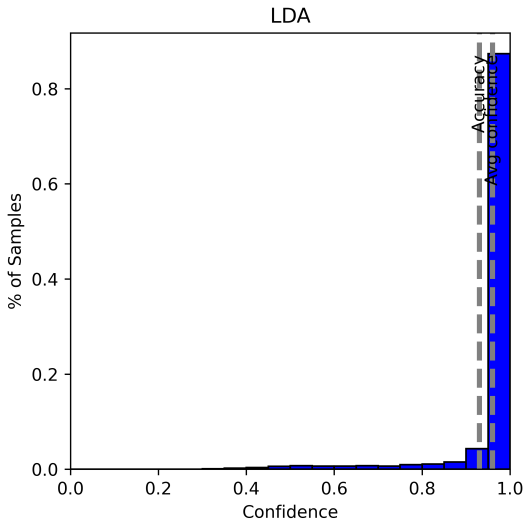


(d) CIFAR-10：可靠性图

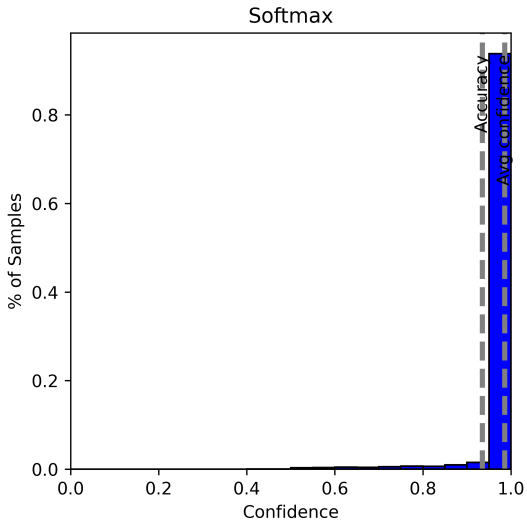
图 11：CIFAR-10 上的校准和置信度行为（格式与图 8 相同）。

C。关于对 DNLL 权重 λ敏感性的补充分析

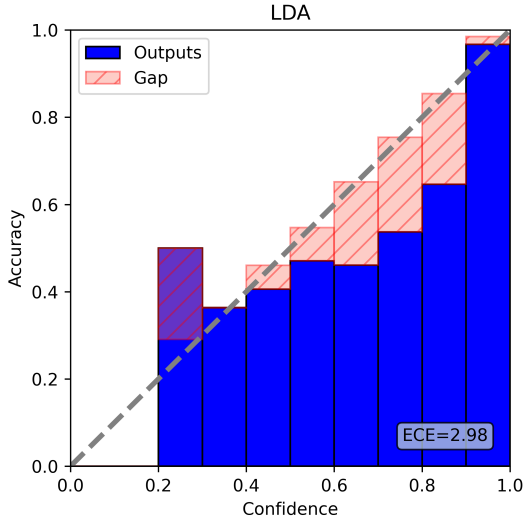
我们在 CIFAR-10 和 Fashion-MNIST 上重复了第4节的 λ 扫描。图 13 显示，在 λ跨越多个数量级时，测试准确率和期望校准误差都保持稳定，而学习得到的球面尺度 σ 会随 λ单调变化，这支持了将 λ 解释为尺度/温度调节旋钮的观点。



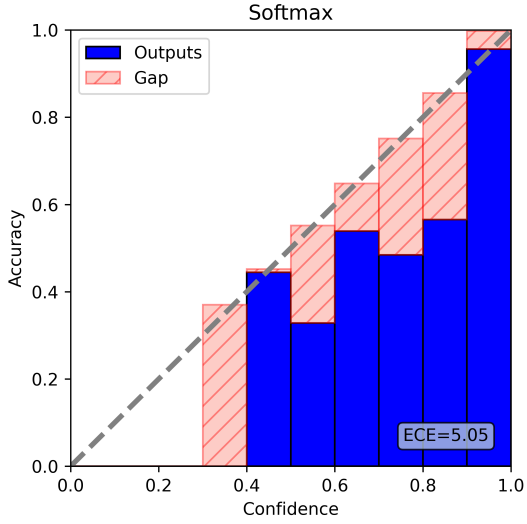
(a) Fashion-MNIST：置信度直方图



(b) Fashion-MNIST：置信度直方图

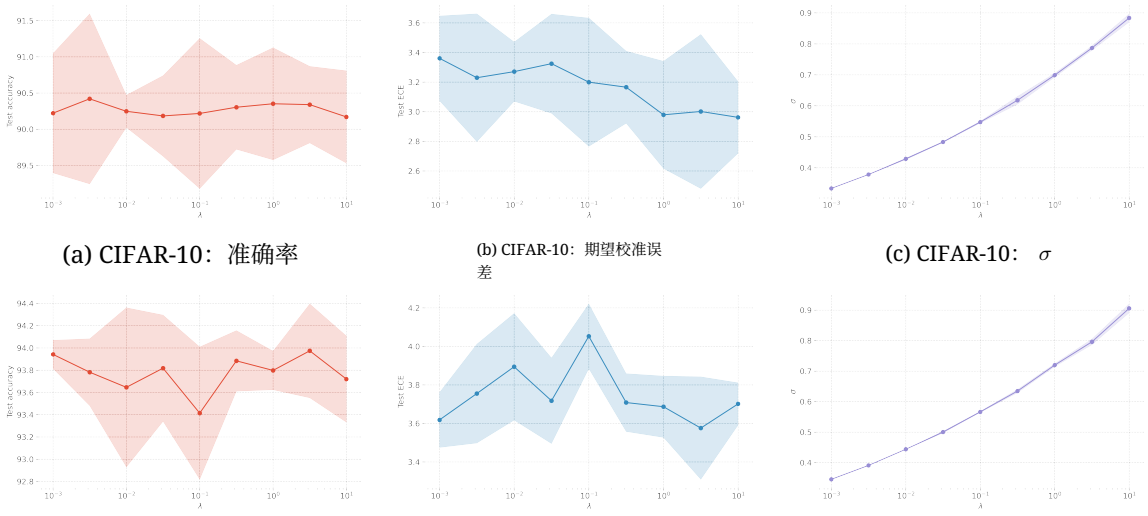


(c) Fashion-MNIST：可靠性图



(d) Fashion-MNIST：可靠性图

图 12：Fashion-MNIST 上的校准和置信度行为（格式与图 8 相同）。



(a) CIFAR-10：准确率

(b) CIFAR-10：期望校准误差

(c) CIFAR-10： σ

(d) Fashion-MNIST：准确率

(e) Fashion-MNIST：期望校准误差

(f) Fashion-MNIST： σ

图 13：深度线性判别分析对 CIFAR-10（上）和 Fashion-MNIST（下）中正则化权重 λ 的敏感性。左：测试准确率。中：测试期望校准误差。右：学习到的球形协方差尺度σ。标记表示多次运行的均值；阴影区域表示不同随机种子间的变化。准确率和期望校准误差在跨越数个数量级的范围内变化不大，而 σ 随 λ系统性变化，这与 λ主要充当生成式头的尺度/温度调节旋钮相一致。