01 生成建模与自编码表示

数据、随机生成与重建的边界

学习目标

- 区分数据分布、有限数据集与模型分布，以及概率 / 密度计算、表示和采样的输出。
- 解释确定性网络计算如何与随机输入、随机抽样共同组成生成过程；恢复两变量自回归分解。
- 追踪图像—数值码—重建，解释自编码器的编码、解码和共同训练，指出重建与先验生成的不同输入条件。
- 说明期望与样本均值的关系；在独立深度部分恢复KL到最大似然的推理及条件。
- 按表示、学习目标和采样过程定位生成方法，说明单个清晰样本与小重建误差的证据范围。

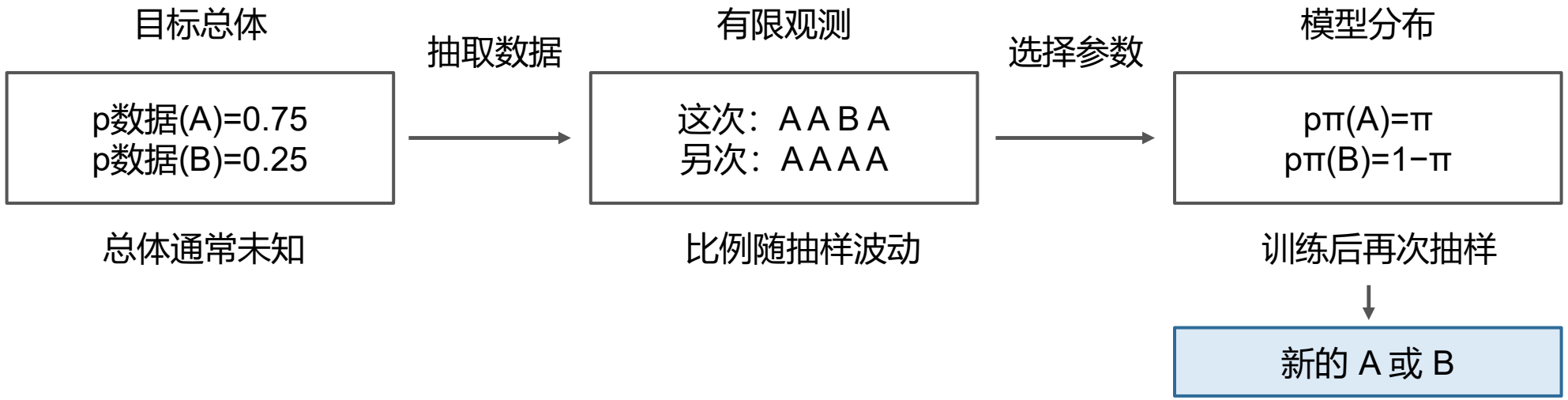
1 数据、生成机制与学习目标

怎样从有限数据学习可生成样本的规律?

有限数据与总体规律

为什么学习分布？保存训练样本可以复现已有观测；生成建模还要用一个可抽样的模型，表达目标总体中各种结果出现的规律。它能否推广到未见情形，需要另行检验。

数据分布 p_{data} 描述目标总体；有限数据集 $\mathcal{D} = \{\boldsymbol{x}^{[n]}\}_{n=1}^{N}$ 是观测，$[n]$ 是样本编号；**模型分布** p_θ 由可学习参数 θ 指定。



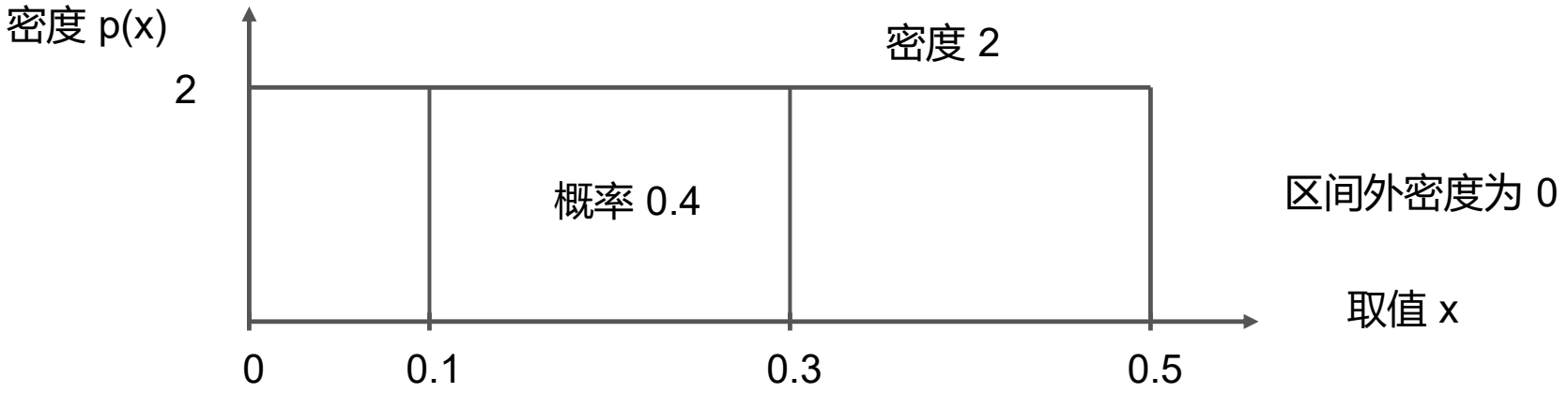
构造例中总体概率由我们给定，真实任务通常未知。有限样本比例会波动；学习目标是逼近总体规律。连续取值怎样用概率描述？

密度通过面积给出概率

离散变量的概率质量 $p(x)$ 可直接表示取值事件的概率，满足 $\sum_x p(x) = 1$。

有密度的连续变量满足 $p(x) \geq 0, \int p(x)\mathrm{d}x = 1, \mathrm{P}(a \leq X \leq b) = \int_a^b p(x)\,\mathrm{d}x$. 这里 X 是随机变量，x 是其可能取值。密度在某一点的高度与区间概率有不同含义。

教学构造：在 $[0, 0.5]$ 上均匀分布时，密度为2，区间外为0。总面积 $2\times0.5 = 1$；落在 $[0.1, 0.3]$ 的概率是 $2\times0.2 = 0.4$。任一单点的概率为0。



计算概率、学习表示与采样

使用任务	给定对象	返回对象
概率或密度计算	模型和观测 $\boldsymbol{x}$	标量 $p_\theta(\boldsymbol{x})$
表示计算	编码函数和观测 $\boldsymbol{x}$	编码向量 $\boldsymbol{z} = f_\phi(\boldsymbol{x})$
生成采样	模型、抽样规则及随机源	新样本 $\tilde{\boldsymbol{x}}$

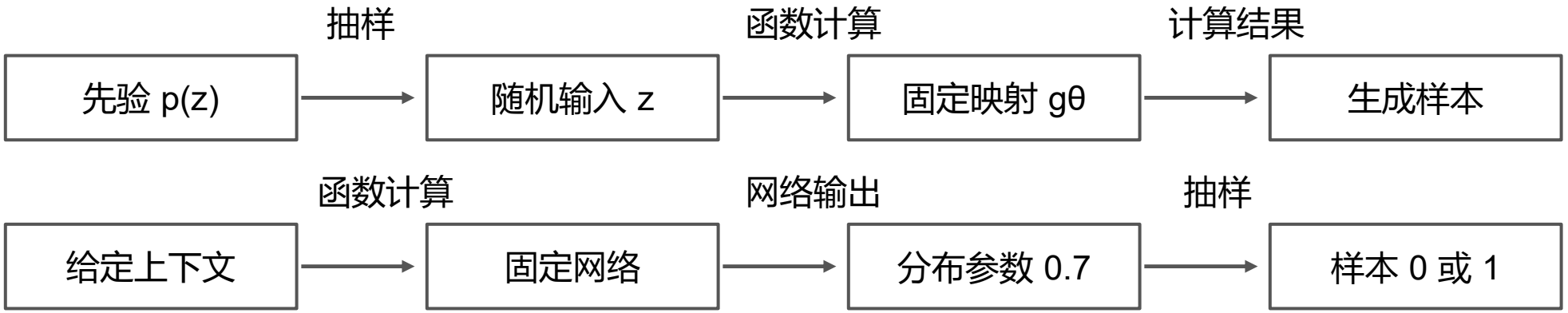
这些能力可以组合，但一种能力并不自动提供其他能力。例如，一个映射能够输出图像，并不说明其图像密度容易计算。

条件生成研究 $p_\theta(\boldsymbol{x} \mid c)$：给定类别、文本等条件 c，生成满足该条件的样本。给定条件后仍可保留多种可能结果；生成过程的随机源需要另外明确。

确定性映射处理随机输入

固定参数与输入后，确定性网络的前向计算返回固定结果。随机生成可以把随机性放在函数之前：$\boldsymbol{z}\sim p(\boldsymbol{z}), \tilde{\boldsymbol{x}} = g_\theta(\boldsymbol{z})$. $\boldsymbol{z}$ 是本次抽到的随机输入；θ 是训练后的参数；$g_\theta(\boldsymbol{z})$ 是计算结果。输入分布与映射共同决定输出分布。

另一种方式是网络先输出分布参数，再从该分布抽样。例如输出概率0.7，再按概率0.7取值1、按概率0.3取值0。数值0.7与抽到的0或1具有不同身份。多个相互依赖的变量怎样依次生成?



θ 是训练后的参数；z 与 0/1 是本次抽到的值

条件分解保留变量之间的依赖

对两变量 X_1, X_2，当 $p(x_1) > 0$ 时，$p(x_2 \mid x_1) = \frac{p(x_1, x_2)}{p(x_1)} \Longrightarrow p(x_1, x_2) = p(x_1)p(x_2 \mid x_1)$. 条件概率在固定 x_1 的范围内归一化。该分解允许第二变量依赖第一变量。

教学构造：两个变量均取0或1，$p(X_1 = 0) = 0.6$，$p(X_1 = 1) = 0.4$。

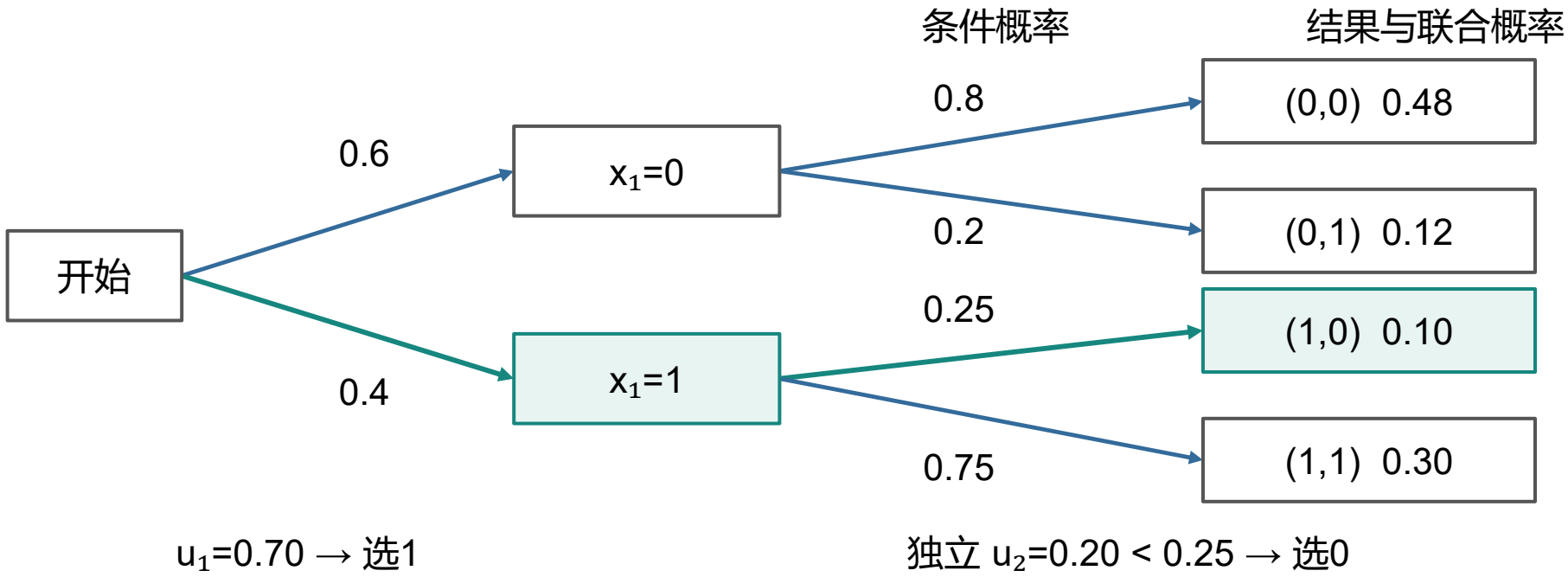
给定 x_1	$p(X_2 = 0 \mid x_1)$	$p(X_2 = 1 \mid x_1)$
0	0.8	0.2
1	0.25	0.75

于是四个联合概率依次为 $p(0,0) = 0.48, p(0,1) = 0.12, p(1,0) = 0.10, p(1,1) = 0.30$. 总和为1；每个条件行分别求和为1。若前项概率为0，相应联合事件也为0，该位置的条件值不由上述比值确定。

同一分解支持概率计算与逐步采样

已知观测：沿路径相乘。 对 $(1,0)$，读取上一页条件表可得 $p(1,0) = 0.4\times0.25 = 0.10$。

生成样本：先抽前项，再选条件分支。



自回归模型把这种先后依赖写为 $p_\theta(\boldsymbol{x}) = \prod_{i=1}^{d} p_\theta\left(x_i \mid x_1, \ldots, x_{i-1}\right)$. 网络可输出条件分布参数；分解允许变量依赖，不限定架构。下一步用数据学习这些参数。

期望与有限样本平均

期望是对指定分布取平均。离散变量下，$\mathbb{E}_{X\sim q}[h(X)] = \sum_x q(x)h(x)$. q 决定各种取值的权重，h 决定平均什么量。连续情形在相关积分存在时改用 $\int q(x)h(x)\mathrm{d}x$。

构造 $q(A) = 0.75, q(B) = 0.25$，取 $h(A) = 0, h(B) = 4$。

期望为 $0.75\times0 + 0.25\times4 = 1$。样本序列AABA的平均也是1；样本序列AAAA的平均为0。

如果观测独立同分布地来自 q，样本均值 $\frac{1}{N}\sum_{n=1}^{N} h(x^{[n]})$ 是总体期望的估计。估计值依赖抽到的数据，有限样本下不要求与总体完全相等。

似然目标与它的适用条件

把上一页的平均函数取为 $h_\theta(\boldsymbol{x}) = \ln p_\theta(\boldsymbol{x})$。**似然**固定观测数据，比较不同参数为它赋予的概率或密度；训练调整的是 θ：

$$max_\theta \frac{1}{N}\sum_{n=1}^{N} \ln p_\theta(\boldsymbol{x}^{[n]}).$$

总体解释。 KL散度（Kullback–Leibler divergence）是有方向的非负分布差异量。记 $q = p_{\text{data}}$，其数据到模型方向为

$$D_{\text{KL}}(q\|p_\theta) = \mathbb{E}_q\left[\ln \frac{q(X)}{p_\theta(X)}\right].$$

连续情形使用密度。KL为0表示两分布在相应测度意义下相同。

KL通常不对称；模型漏掉正概率事件时，数据到模型的KL可为无穷。

可直接使用的结论。 数据分布固定、拆出的期望均存在且为有限量时，最小化该KL等价于最大化总体期望对数似然；iid观测的均值估计这个期望。有限数据与所选模型类仍限制总体恢复。

训练准则约束模型给数据的概率；重建准则则约束输入信息能否恢复。两者需要分别检查。

2 KL、期望似然与样本目标

可选深度：完整推导KL到似然；不讲可转第3节。

数据到模型的KL

本节独立展开上一节结论。先以离散分布说明，记固定数据分布为 $q = p_{\text{data}}$： $D_{\text{KL}}(q\|p_\theta) = \sum_{x:q(x)>0} q(x)\ln\frac{q(x)}{p_\theta(x)}$. 平均权重来自 q，分母是待学习的模型。

交换两者会同时改变平均分布和比值，数值通常不同。

构造两个离散分布：$q = (0.75,0.25)$ 与 $p = (0.5,0.5)$。对应的数据到模型KL为 $D_{\text{KL}}(q\|p) = 0.75\ln 1.5 + 0.25\ln 0.5 \approx 0.130812$. 反向为 $0.5\ln(2/3) + 0.5\ln 2 \approx 0.143841$。自然对数下单位为nat。

若某项 $q(x) > 0$ 而 $p_\theta(x) = 0$，该项发散；以下有限量分解需排除此情况并要求相应期望存在。对于有密度的连续分布，以共同参考测度下的密度和积分写对应表达。

固定数据分布后消去参数无关项

1．展开对数比。 在各项均存在且为有限量时，$D_{\mathrm{KL}}(q\|p_\theta) = \mathbb{E}_q[\ln q(X)] - \mathbb{E}_q[\ln p_\theta(X)]$.

2．固定数据分布。 q 不随模型参数改变，因此 $C = \mathbb{E}_q[\ln q(X)] = \sum_x q(x)\ln q(x)$ 相对于 θ 为常数，KL写成 $C - \mathbb{E}_q[\ln p_\theta(X)]$。

3．只优化随参数改变的项。 减去的项越大，KL越小：$argmin_\theta D_{\mathrm{KL}}(q\|p_\theta) = argmax_\theta \mathbb{E}_q[\ln p_\theta(X)]$. 这是模型类内的最优参数集合关系；最优值未取得时讨论逼近。总体权重仍未知，需要用数据估计期望。

从总体目标到有限样本似然

1．估计总体目标。 对独立同分布观测 $x^{[n]} \sim q$，用数据平均估计期望：$\mathbb{E}_q[\ln p_\theta(X)] \approx \frac{1}{N}\sum_{n=1}^{N} \ln p_\theta(x^{[n]})$. 这里是**统计估计**，有限样本不要求两者相等。

2．写出有限数据的似然。 模型也按iid表示这些观测时，$L(\theta;\mathcal{D}) = \prod_{n=1}^{N} p_\theta\,(x^{[n]}), \frac{1}{N}\ln L(\theta;\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N} \ln p_\theta(x^{[n]})$.

3．得到可优化准则。 对数单调递增，正因子 $1/N$ 不改变最优参数。因此最大化乘积似然与平均对数似然**代数等价**；取负号便得到负对数似然损失。

开头的3个A、1个B提供一个可直接求解的模型参数例。

样本计数怎样确定简单模型

对A、B两类模型，令 $p_\pi(A)=\pi, p_\pi(B)=1-\pi$，其中 $0<\pi<1$。

1．从样本写目标。 3个A、1个B给出

$$J(\pi)=\frac{3}{4}\ln\pi+\frac{1}{4}\ln(1-\pi).$$

2．求驻点。 令导数为0：

$$\begin{aligned} J'(\pi) &= \frac{3}{4\pi}-\frac{1}{4(1-\pi)}=0 \\ &\Longrightarrow 3(1-\pi)=\pi \\ &\Longrightarrow \hat{\pi}=0.75. \end{aligned}$$

求得驻点还不够；还须检验最大值，并区分样本最优与总体规律。

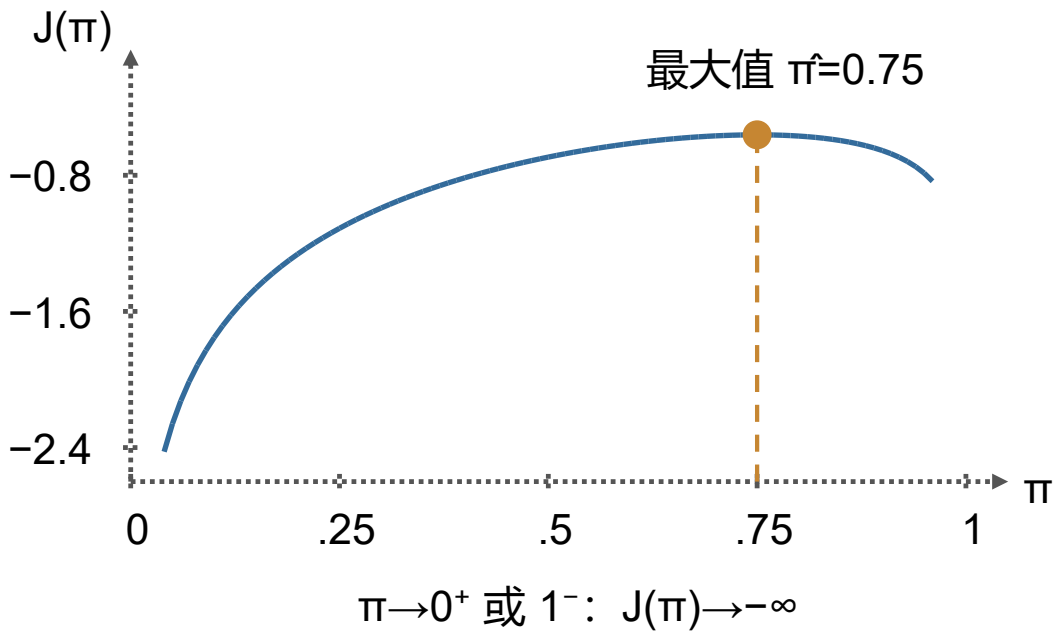
检验最大值与有限样本边界

上一页从3个A、1个B得到内部驻点 $\hat{\pi} = 0.75$。现在判断它代表什么。

先确认驻点是最大值。

$$J''(\pi) = -\frac{3}{4\pi^2} - \frac{1}{4(1-\pi)^2} < 0.$$

目标严格凹，内部驻点唯一最大；两端目标趋于负无穷。



本样本下，$J(0.75) \approx -0.562335$，优于 $J(0.5) \approx -0.693147$。真实比例在构造中恰好为0.75；换样本，估计值可能改变。

总体目标等价、有限数据估计、模型内优化是三个不同环节。推导解释了目标来源，未保证任何有限数据或优化算法都恢复真实分布。

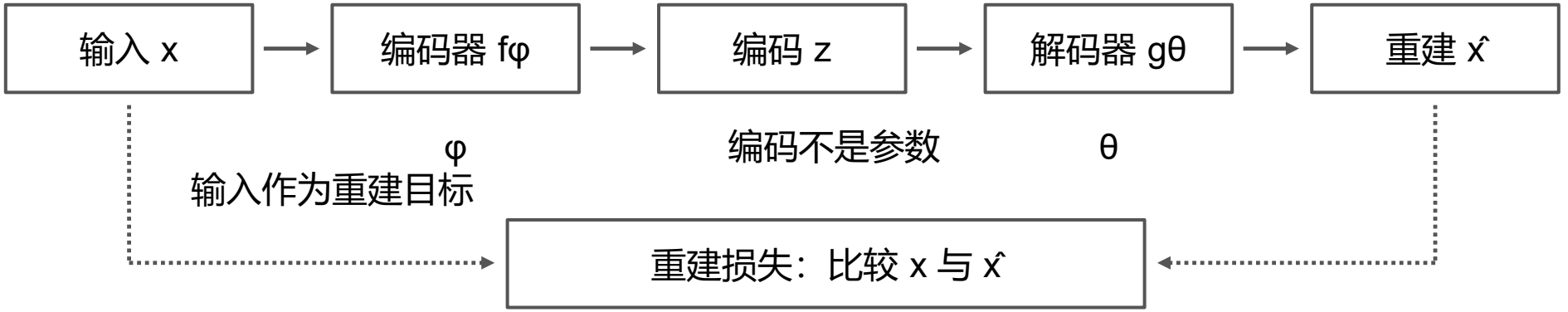
3 自编码表示与先验生成

换成重建任务：保留了什么，生成还缺什么？

编码与解码构成重建路径

前面用似然学习概率分布；现在换成重建任务：不要求直接计算样本概率，先编码再恢复输入，观察编码保留了什么。

自编码器（Autoencoder，AE）以输入自身作为重建目标。本讲使用确定性编码器与解码器：$\boldsymbol{z} = f_\phi(\boldsymbol{x}), \hat{\boldsymbol{x}} = g_\theta(\boldsymbol{z}) = g_\theta(f_\phi(\boldsymbol{x}))$. 其中 $\boldsymbol{x} \in \mathbb{R}^d$，$\boldsymbol{z} \in \mathbb{R}^m$，$\hat{\boldsymbol{x}} \in \mathbb{R}^d$；$\phi, \theta$ 是两组参数，$\boldsymbol{z}$ 是随输入变化的编码。



重建目标与前面的分布似然目标不能直接等同；重建得到的表示是否适合其他任务，还需相应数据与任务检验。

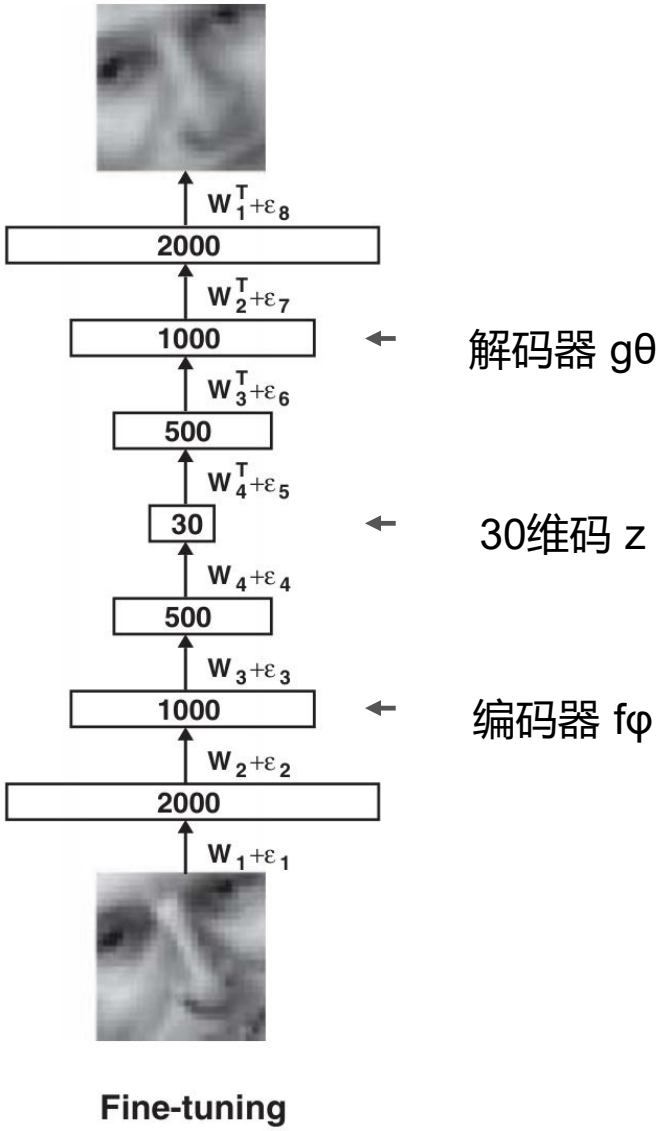
原论文网络怎样实现编码与解码

原图由下向上：输入、编码器、30维码、解码器、重建。中文标签为教学添加。

- 下半部编码，上半部解码，图中权重对应参数 ϕ, θ。
  - 2000、1000、500是隐藏层单元数。
- 中央码对应给定图像算出的 z，不是参数或小图像。
  - 中间30表示30个实数。

这是历史AE实例：论文先用RBM预训练再共同微调；这里只摘Fig.1右侧微调网络。预训练不是所有AE的必要步骤。

来源：Hinton等，2006，Fig.1右图



Fine-tuning

原始数字经过编码后怎样重建

逐列比较上行测试输入与下行AE重建；图中未含PCA行。

- 输入：28×28 数字图像，784像素
- 编码：30个实数 $\boldsymbol{z}=(z_1,\ldots,z_{30})^{\mathrm{T}}$。
- 解码：恢复784个像素值，排列为下行图像。

本页为MNIST数字实验，编码层宽784–1000–500–250–30，与前页网络示意配置不同。中间码为**教学机制示意**，原图未给码值；上下行是输入与重建，不是先验生成。

来源：Hinton等（2006），Fig.2B前两行，PDF p.2。





编码过程为教学示意；原图未展示码值

同一损失共同约束两组参数

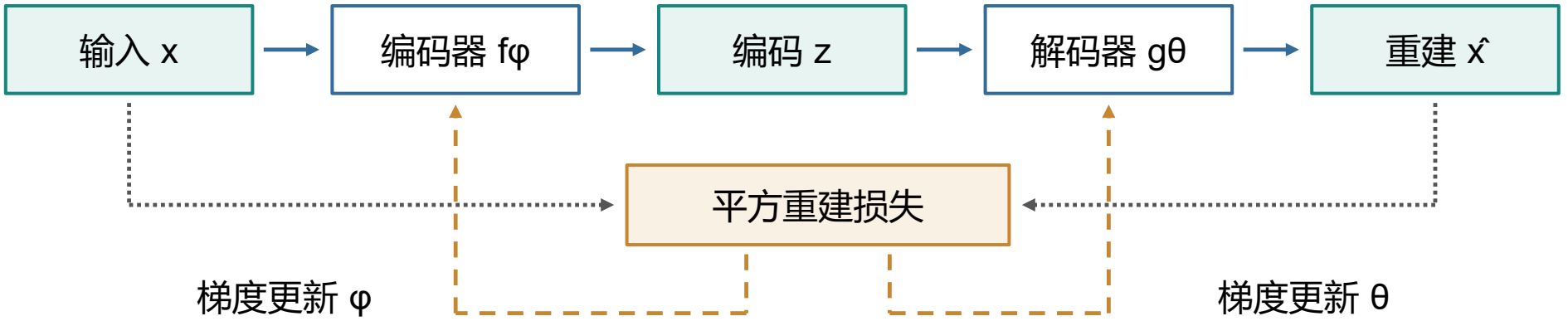
把输入与重建的差异作为训练信号。

下面用平方误差说明训练机制，不是前两页论文实验损失的复现：

$$\hat{\boldsymbol{x}}^{[n]} = g_\theta(f_\phi(\boldsymbol{x}^{[n]})),$$

$$min_{\phi,\theta} \ \frac{1}{N}\sum_{n=1}^{N}\left\|\boldsymbol{x}^{[n]} - \hat{\boldsymbol{x}}^{[n]}\right\|_2^2.$$

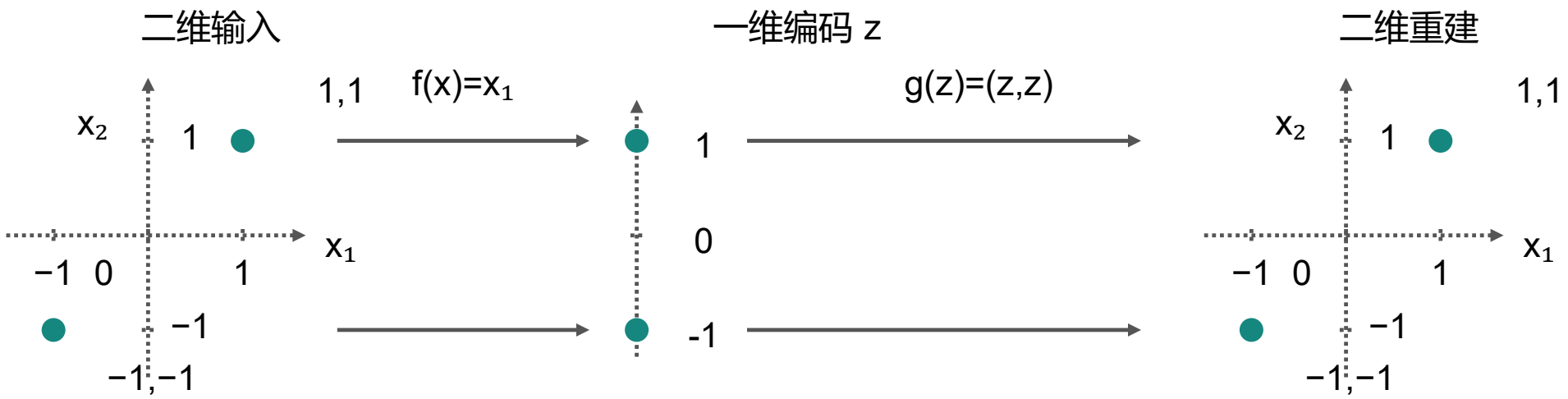
$\|\boldsymbol{a}\|_2^2 = \sum_j a_j^2$，各分量误差相加。



两组参数影响同一损失，可通过复合映射梯度共同调整。限制维数、容量或加入正则可促使编码选择信息；并非所有AE都要求 $m < d$。固定一对简单映射，可以直接检查零误差保留了什么。

一个可直接核对的重建构造

教学构造：目标只可能出现图中的两个对角点。 固定 $f(\boldsymbol{x}) = x_1, g(z) = (z, z)^{\mathsf{T}}$.



目标支持集外的输入：(1,−1) → 1 → (1,1)，平方重建误差 4

对两个对角点，保留 x_1 即可恢复 $x_2 = x_1$，重建误差为0。

目标分布支持集外的 $(1,-1)^{\mathsf{T}}$ 被重建为 $(1,1)^{\mathsf{T}}$，平方误差为4。

这些映射是人工给定，非训练结果。**重建准确需说明所针对的输入集合；从先验抽取新的编码，还能得到相同分布吗?**

生成时由先验决定解码器的输入分布

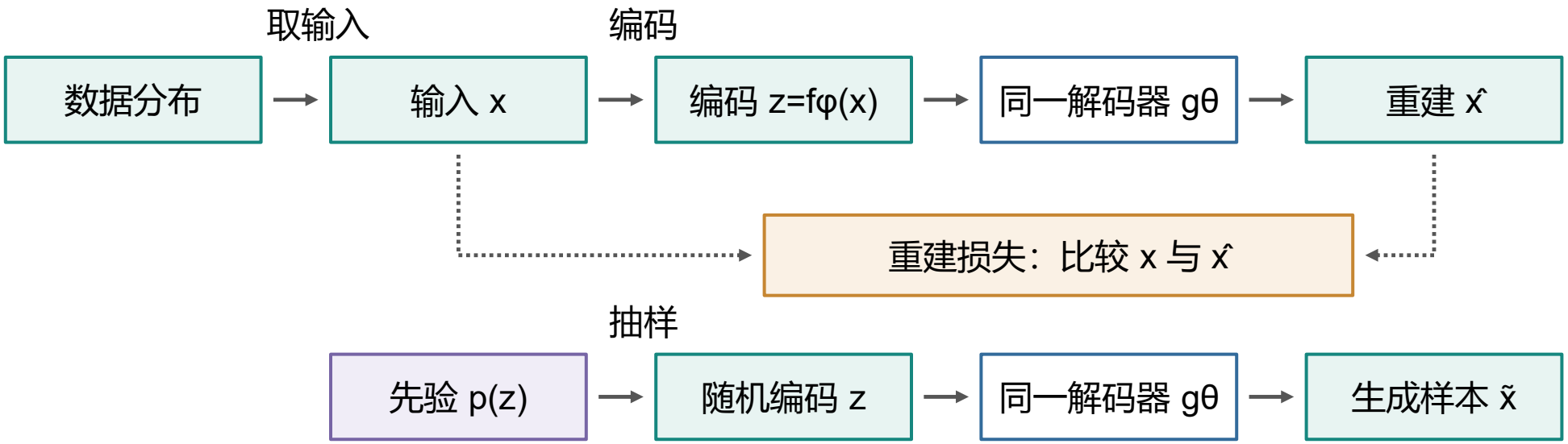
重建时先有真实输入，编码由它决定：$\begin{aligned} \boldsymbol{x} &\sim p_{\text{data}}, \boldsymbol{z} = f_\phi(\boldsymbol{x}), \\ \hat{\boldsymbol{x}} &= g_\theta(\boldsymbol{z}); \end{aligned}$

从先验生成时，编码由规定的先验抽样得到：$\boldsymbol{z} \sim p(\boldsymbol{z}), \tilde{\boldsymbol{x}} = g_\theta(\boldsymbol{z}).$

重建目标检查第一条路径上的输入与重建差异。第二条路径还需规定 $p(\boldsymbol{z})$，并使它与解码器学得的映射相配合。两条路径的编码分布可以相同，也可以不同；匹配关系需另外保证。

重建与先验生成的两条路径

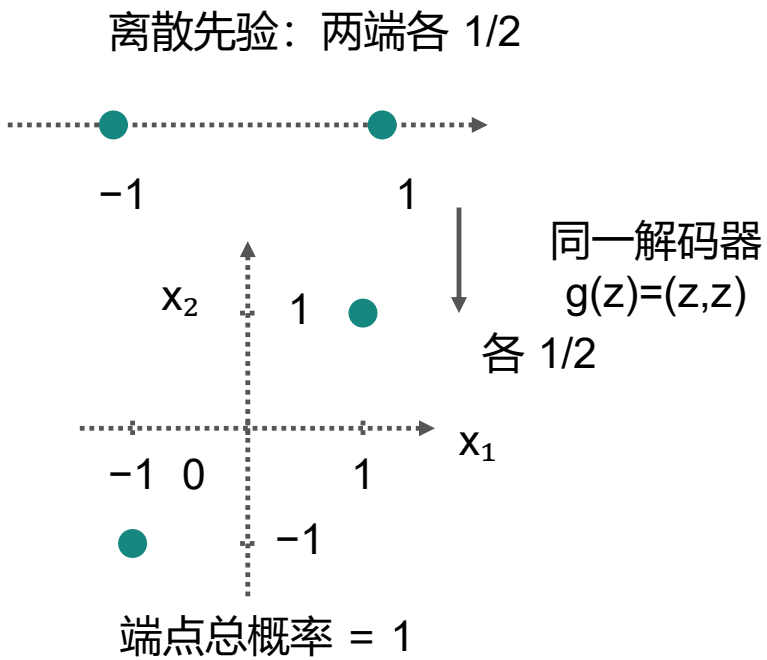
两条路径使用同一解码器。上路的编码由数据和编码器决定，下路的编码由规定的先验决定；重建损失只比较上路的输入与重建。

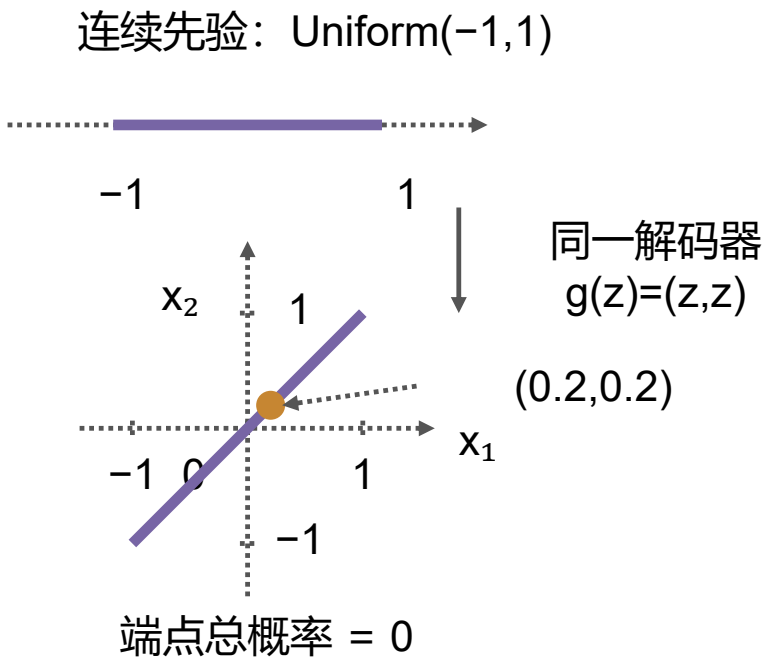


上路编码分布由数据与编码器决定；下路输入分布由先验决定

零重建误差仍不能决定生成分布

继续固定 $f(\boldsymbol{x}) = x_1$、$g(z) = (z, z)^{\mathsf{T}}$。目标总体只在 $(-1,-1)^{\mathsf{T}}$、$(1,1)^{\mathsf{T}}$ 各占 $1/2$；编码只取 $-1,1$，重建误差为0。





相同解码器、相同重建误差，生成分布却不同。**重建目标没有单独确定应该从什么分布抽取编码。** 连续先验给两个端点的概率为0，与目标全部质量在端点不同。

完整生成模型需要补充的设计

从AE重建路径扩展到先验生成，至少需要回答：

- 生成时从什么先验取得编码，这个先验是否依赖给定条件？
- 解码器如何使用编码，是直接输出样本，还是输出条件分布的参数再抽样？
- 学习目标如何约束这条生成路径，使其输出分布与数据规律相配合？

AE可为生成系统提供编码器或解码器。下一讲的变分自编码器（VAE）将明确先验、条件生成分布与近似推断的关系；本讲只确立需要补充的对象，不预设它已解决所有生成质量问题。不同生成方法可按“分布怎样表示、目标怎样学习、样本怎样生成”比较。

4 方法地图与证据范围

按表示、目标和采样比较路线与证据。

同一任务的不同建模路线

同一生成任务，可以用不同的分布表示和生成过程实现。下图只比较训练后怎样取得一个样本。



典型学习目标：自回归使用对数似然；VAE用变分下界联系似然与推断；GAN使用判别器提供的对抗学习信号。

“能生成样本”与“能方便地计算密度”是两个问题；生成过程与训练目标也要分别说明。

六类路线的共同比较维度

路线	分布表示	典型学习目标	生成采样
自回归	条件分布的链式乘积	数据对数似然	首项边缘，后续条件抽样
VAE	先验与条件生成分布	变分下界联系似然与推断	先抽潜变量，再条件抽样
GAN	随机输入经映射诱导分布	判别器提供对抗信号	抽随机输入，运行生成器
正规化流	可逆变换与变量变换密度	变量变换计算的似然	基础样本经可逆变换
扩散模型	学习加噪对应的逆向机制	常用去噪或噪声预测	由噪声逐步逆向生成
Flow Matching	沿概率路径传输的速度场	匹配构造的目标速度	基础样本积分速度场

正规化流侧重可逆变换与密度表示；Flow Matching侧重速度场的训练，也可用于连续正规化流，两者存在交叉。

这不是互斥网络分类；条件控制、潜在表示和多步采样可以组合。具体目标与采样器仍须在各方法的假设下讨论。

不同证据回答不同问题

观察到的证据	能支持的判断	还不能独立回答的问题
一张清晰的生成图	该输出的局部可辨识性	是否覆盖目标中的多种情况
给定数据上的小重建误差	编码—解码路径重建这些输入较准确	任意先验是否匹配生成路径
较高训练对数似然	对当前训练数据，所用目标值较高	新数据表现与真实总体恢复
采样运行较快	指定配置下的生成耗时较低	生成质量、覆盖与成本的整体权衡

比较方法应同时说明任务、数据和使用条件，再分别讨论质量、覆盖及生成成本。本讲的分布语言用于解释证据的范围，评价指标与具体评价设计留到后续专题。

本讲总结

- 数据集提供有限观测，数据分布描述目标总体，模型分布由建模形式及参数确定；连续密度通过区域积分给出概率。
- 概率计算、表示计算与采样返回不同对象。固定网络可处理随机输入，也可输出供抽样使用的分布参数。
- 自回归用条件分解表示联合分布；概率计算读取完整观测，生成依次抽取未知分量。
- 固定数据分布且相关量定义良好时，数据到模型KL与期望对数似然给出等价的总体优化方向；iid样本均值提供经验目标，有限样本结果仍有局限。
- AE把输入映射为数值码再重建，共同训练编码器与解码器以降低重建误差；先验生成还需规定编码的输入分布并使其与解码机制匹配。
- 方法比较要分开表示、学习目标和采样；单个样本、小重建误差或单一成本值各自只回答部分问题。