02 潜变量模型与变分自编码器

从概率模型到可微训练与先验生成

学习目标

- 区分先验、联合、边缘、真实后验与近似后验，说明概率编码器和解码器输出什么
- 推导ELBO与后验差距的关系，区分两处KL及固定生成参数、联合优化的条件
- 推导高斯KL、重建项与重参数化梯度，说明随机估计和期望目标的区别
- 解释共享推断、训练、重建与先验生成，限定表示和样本质量结论

核心讲授保留机制图、目标与使用条件；完整学习再展开两段推导。各节入口给出课堂保留页，完整推导仍在本讲中。

1 潜变量模型与概率网络

从联合、证据与后验建立编码和生成接口。

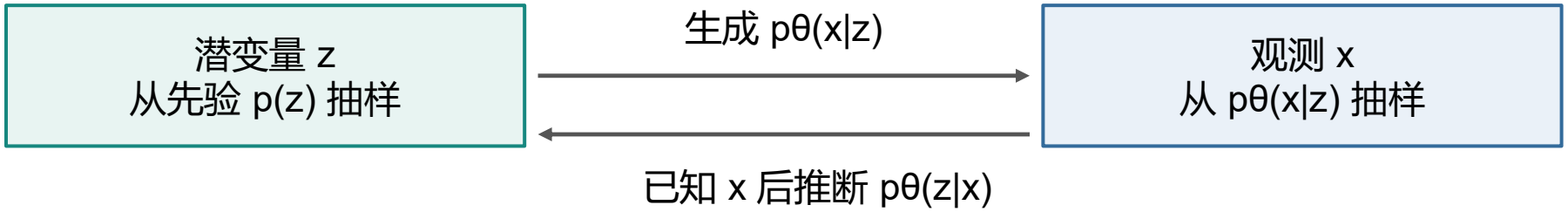
联合分布定义潜变量模型

普通自编码器已经能重建输入。现在希望不提供输入图像，也能生成新图像；重建目标本身不保证从指定先验抽取的潜变量能被合理解码。因此，需要明确新样本的概率来源。

潜变量是参与生成但未被直接观测的随机变量；观测向量记为$\boldsymbol{x} \in \mathbb{R}^d$，潜变量记为$\boldsymbol{z} \in \mathbb{R}^k$。

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z}) p_\theta(\boldsymbol{x} \mid \boldsymbol{z}).$$

- 先验$p(\boldsymbol{z})$规定潜变量怎样取值
- 条件分布$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$规定给定潜变量后怎样生成观测
- θ是模型参数，不是为每个样本重新抽取的随机变量



生成定义联合分布；反向推断由同一模型决定

边缘化与后验回答不同问题

数据只有$\boldsymbol{x}$时，需要把所有可能的潜变量计入其概率。连续潜变量用积分，离散潜变量用求和：

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}.$$

这个量称为观测的**边缘似然或证据**。模型参数固定后，给定观测反推潜变量使用**后验**：

$$p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{x})}, 0 < p_\theta(\boldsymbol{x}) < \infty.$$

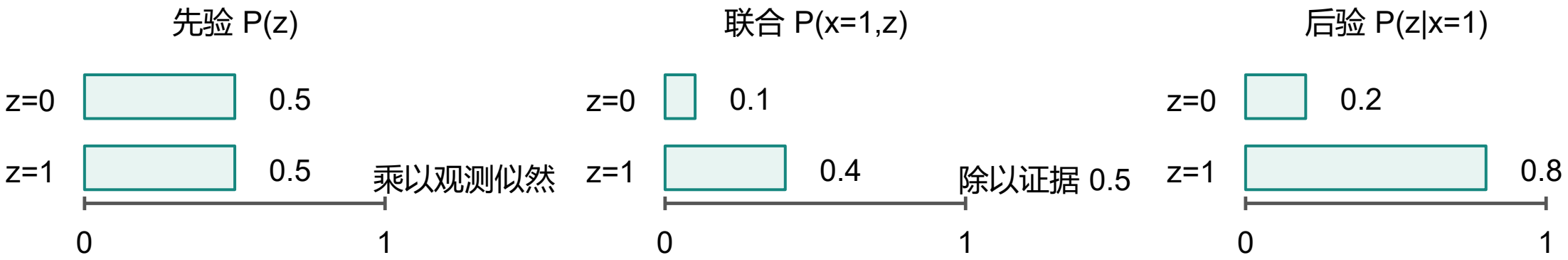
已知量	所求对象	操作
模型参数与观测	潜变量后验分布	推断
训练数据与模型族	生成参数θ	学习

复杂非线性解码分布的联合密度可以逐点计算，但对全部潜变量积分通常困难。因此，真实后验的归一化也可能困难；困难不适用于所有潜变量模型。

一个可精确计算的后验例子

考虑教学构造的离散模型：$z \in \{0,1\}$，两种状态的先验概率均为0.5。

条件概率为$P(x=1 \mid z=0)=0.2$、$P(x=1 \mid z=1)=0.8$。



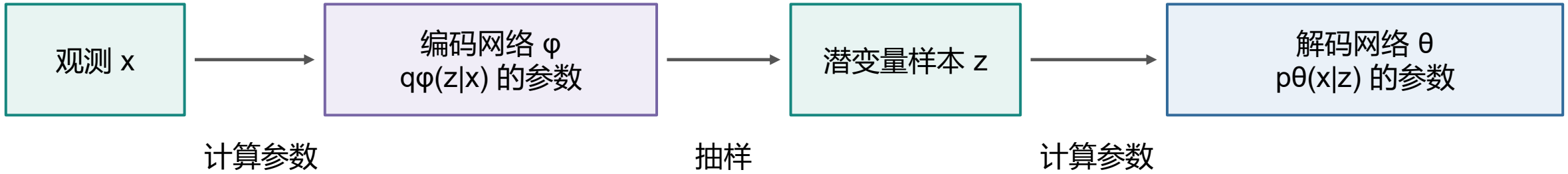
$P(x=1)=0.1+0.4=0.5, P(z=1 \mid x=1)=0.4/0.5=0.8.$

先验的0.5与后验的0.8描述不同的已知信息。本例能精确求和；复杂连续模型可能难以完成同一项边缘化，于是引入可计算的近似后验$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$。

编码器近似后验，解码器定义观测分布

真实后验由生成模型决定。变分自编码器（Variational Autoencoder，**VAE**）另用一个可计算的分布族$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$近似它，并共同学习生成参数$\theta$与编码参数$\phi$。

对象	输入	输出及身份
编码网络	观测$\boldsymbol{x}$	近似后验的分布参数
从编码分布采样	分布参数与随机性	潜变量样本$\boldsymbol{z}$
解码网络	潜变量$\boldsymbol{z}$	观测条件分布的参数



两个网络并非互逆函数。q_ϕ只近似$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$；解码分布的均值可以展示为重建，但完整概率模型还包含分布形状和不确定性。

本讲采用的高斯设定

本讲选标准高斯先验，便于抽样并提供潜变量的参考分布；用对角高斯近似后验，只需网络输出均值与对数方差，且在此先验下可解析计算KL：

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(\boldsymbol{x}), diag(e^{\ell_\phi(\boldsymbol{x})})\right).$$

编码网络输出均值$\boldsymbol{\mu}_\phi$和对数方差ℓ_ϕ。第j维方差为$e^{\ell_j} > 0$，标准差为$e^{\ell_j/2}$，二者不能互换。

对实值观测，选固定方差的高斯解码分布，使负对数似然成为带确定系数的平方误差加常数：

$$p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(g_\theta(\boldsymbol{z}), \tau^2 \boldsymbol{I}), \tau^2 > 0.$$

g_θ输出观测均值；τ^2描述给定潜变量后的观测波动。编码方差描述潜变量的不确定性，两类方差位于不同空间。这些是本讲便于计算和推导的建模选择；VAE也可采用其他先验、近似后验与观测分布。

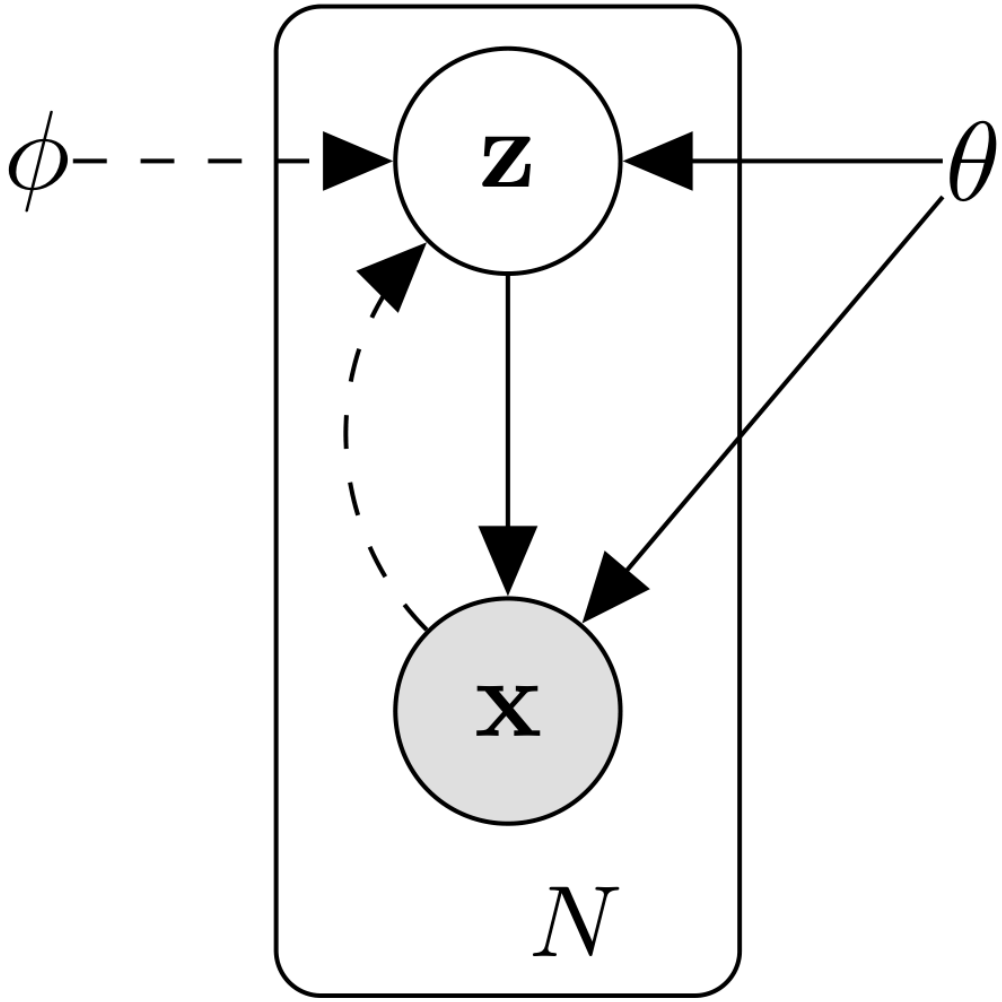
原论文概率图对应生成与近似推断

实线表示生成模型：先产生z，再产生观测x；虚线表示从已知x构造近似后验。方框中的N表示每个观测都有自己的潜变量。

图中θ, ϕ分别对应本讲生成参数与编码参数。原文先验可含θ；本讲固定标准高斯先验，因此只学习解码器中的θ。

这是概率依赖图。前面的可编辑网络图进一步展开“分布参数—抽样—分布参数”；本图没有规定隐藏层数或网络宽度。

来源：Kingma与Welling，AEVB，Fig.1，PDF第2页，arXiv:1312.6114v11（2022修订）。



2 ELBO与变分推断

核心路线：P13机制图、P14两KL、P16目标与条件。

从后验差距寻找可计算目标

在前述复杂模型中，真实后验和证据都受同一积分困难限制。训练仍需要一个可优化目标；下面用可计算的联合密度与近似后验，将难算的证据分离出来。

对分布q, r，**KL散度**定义为 $D_{\mathrm{KL}}(q\|r) = E_q\left[\log\frac{q}{r}\right]$.

期望$E_q[f]$表示按q加权平均f：连续情形为$\int q(\boldsymbol{z})f(\boldsymbol{z})d\boldsymbol{z}$，离散情形改为求和。KL的方向有意义，交换两分布通常改变数值。

本段固定观测$\boldsymbol{x}$，简写$q = q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$。假设证据正且有限、$q$的支持包含于真实后验支持，且以下KL和相关期望有限，避免无穷量相减。

后验差距为$D_{\mathrm{post}} = D_{\mathrm{KL}}(q\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))$。将Bayes关系代入其分母，可以把证据与可计算项分开。

证据等于ELBO加后验差距

代入$p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) = p_\theta(\boldsymbol{x}, \boldsymbol{z})/p_\theta(\boldsymbol{x})$：

$$\begin{aligned} D_{\text{post}} &= E_q[\log q - \log p_\theta(\boldsymbol{z} \mid \boldsymbol{x})] \\ &= E_q[\log q - \log p_\theta(\boldsymbol{x}, \boldsymbol{z}) + \log p_\theta(\boldsymbol{x})] \\ &= \log p_\theta(\boldsymbol{x}) - E_q[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q]. \end{aligned}$$

第二行使用Bayes关系和对数商法则；第三行中观测已固定，$\log p_\theta(\boldsymbol{x})$与$\boldsymbol{z}$无关，且$E_q[1] = 1$。定义**证据下界**（Evidence Lower Bound，ELBO）：

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = E_q[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q], \log p_\theta(\boldsymbol{x}) = \mathcal{L} + D_{\text{post}}.$$

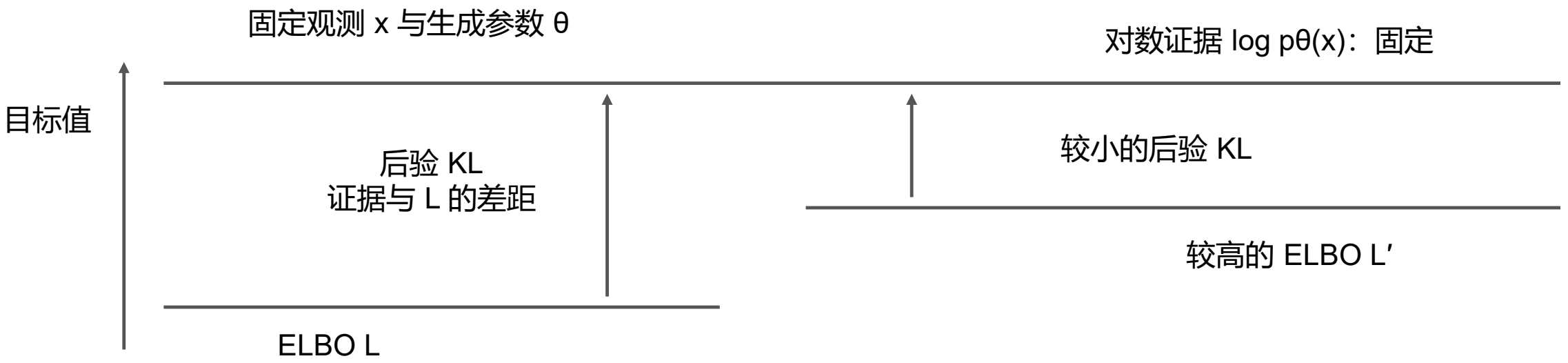
KL非负，所以$\mathcal{L} \leq \log p_\theta(\boldsymbol{x})$。非负性可由$\log u \leq u - 1$得到：$-E_q \log(r/q) \geq 1 - \int_{q>0} r \geq 0$。两分布几乎处处相同时KL为零，ELBO等于对数证据。

ELBO与后验差距的几何关系

给定观测$\boldsymbol{x}$，记$q = q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$。ELBO按该编码分布取平均：

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = E_q[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q].$$

它由可计算的联合密度与编码密度构造，是可优化的训练目标；固定生成参数后，证据与ELBO之差是后验KL。



高度为示意；竖直距离表示非负后验差距，不表示先验 KL

后验匹配时差距为零。先验KL比较编码分布与先验，属于ELBO内部的另一项，不是图中的后验差距。

ELBO中的KL比较编码分布与先验

将联合分解$p_\theta(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{z})p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$代入：

$$\begin{aligned}\mathcal{L} &= E_q[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) + \log p(\boldsymbol{z}) - \log q] \\ &= E_q[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{\mathrm{KL}}(q \| p(\boldsymbol{z})).\end{aligned}$$

第一项按编码分布平均观测的对数概率，通常称为重建项；第二项是编码分布与先验之间的KL。

KL	第二个分布	在恒等关系中的职责
D_{post}	真实后验$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$	对数证据与ELBO的差距
D_{prior}	先验$p(\boldsymbol{z})$	ELBO中被减去的一项

固定θ时，对数证据不随ϕ变化，因此最大化**完整ELBO**等价于最小化后验差距。联合更新θ, ϕ时，证据和真实后验也随θ改变；不能仅用“逼近一个固定后验”解释整个训练。

零先验KL仍可能有后验误差

沿用离散例：给定$x = 1$，先验为$(0.5, 0.5)$，真实后验为$(0.2, 0.8)$，$\log P(x = 1) = \log 0.5 \approx -0.693147$。取自然对数。

编码分布q	重建期望	先验KL	ELBO	后验KL
$(0.5, 0.5)$	$0.5\log 0.2 + 0.5\log 0.8 = -0.916291$	0	−0.916291	0.223144
$(0.2, 0.8)$	$0.2\log 0.2 + 0.8\log 0.8 = -0.500402$	0.192745	−0.693147	0

第一行中q等于先验，因此先验KL为零，却没有匹配真实后验。第二行后验匹配，ELBO达到证据；先验KL反而大于零。两行都满足“ELBO + 后验KL = 对数证据”。

ELBO的结论接口

给定观测$\boldsymbol{x}$，生成模型为$p_\theta(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{z})p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$。近似后验记为$q = q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$。

期望E_q按该近似后验平均；KL定义为$D_{\mathrm{KL}}(q\|r) = E_q \log(q/r)$。

$$\begin{aligned} \mathcal{L} &= E_q[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q] \\ &= E_q \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) - D_{\mathrm{KL}}(q\|p(\boldsymbol{z})), \\ \log p_\theta(\boldsymbol{x}) &= \mathcal{L} + D_{\mathrm{KL}}(q\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})). \end{aligned}$$

证据正且有限、支持相容且相关期望与KL有限时，后验KL非负，$\mathcal{L}$是对数证据下界；后验匹配时取等。

- 固定θ时，提高完整ELBO等价于减小后验差距
- 联合更新θ,ϕ时，生成模型、证据与后验共同变化
- 先验KL和后验KL比较不同对象；采样估计只近似期望目标，不保证逐次为下界

3 重参数化与训练梯度

核心路线：P21梯度路径、P26目标、P27依赖图。

从高斯对数密度得到先验KL

ELBO给出了概率目标。对角高斯设定使其中的先验KL可以直接计算，不需要对这一项抽样。

固定一个观测，记$q = \mathcal{N}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}^2))$，$p = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$，每个$\sigma_j > 0$。独立分量的对数密度相加：

$$\log q - \log p = \frac{1}{2}\sum_{j=1}^{k}\left[-\log\sigma_j^2 - \frac{(z_j - \mu_j)^2}{\sigma_j^2} + z_j^2\right].$$

两边对q求期望。由方差定义，$E_q[(z_j - \mu_j)^2] = \sigma_j^2$；再由$z_j = (z_j - \mu_j) + \mu_j$，中心项均值为零，得到$E_q[z_j^2] = \sigma_j^2 + \mu_j^2$。因此

$$D_{\text{prior}} = \frac{1}{2}\sum_{j=1}^{k}(\mu_j^2 + \sigma_j^2 - 1 - \log\sigma_j^2).$$

该结果比较的是编码分布和标准高斯先验，未计算真实后验KL。

观测方差决定平方误差系数

对d维观测，取$p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(g_\theta(\boldsymbol{z}), \tau^2 \boldsymbol{I})$。从高斯密度取负对数：

$$-\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \frac{d}{2}\log(2\pi\tau^2) + \frac{1}{2\tau^2}\|\boldsymbol{x} - g_\theta(\boldsymbol{z})\|_2^2.$$

最小化负ELBO对应 $-\mathcal{L} = E_q\left[\frac{\|\boldsymbol{x}-g_\theta(\boldsymbol{z})\|_2^2}{2\tau^2}\right] + D_{\text{prior}} + C, C = \frac{d}{2}\log(2\pi\tau^2).$

τ^2固定时，C与待优化的θ, ϕ无关，可以从训练损失删除；若学习观测方差，该项就影响优化。编码方差σ_j^2仍属于KL和潜变量采样，不能删去。

平方范数是d个维度的**求和**。只把它改为维度平均而保持KL系数为1，会改变两项权重；若把整个损失除以d，KL也应除以d。若观测采用Bernoulli分布，负对数概率产生二元交叉熵，平方误差不再是该模型的重建项。

二维数值核对KL与损失尺度

取教学构造的$\boldsymbol{\mu} = (1,0)$、$\boldsymbol{\sigma}^2 = (0.25,1)$，则

$$D_{\text{prior}} = \frac{1}{2}(1 + 0.25 - 1 - \log 0.25) + \frac{1}{2}(0 + 1 - 1 - \log 1) \approx 0.818147.$$

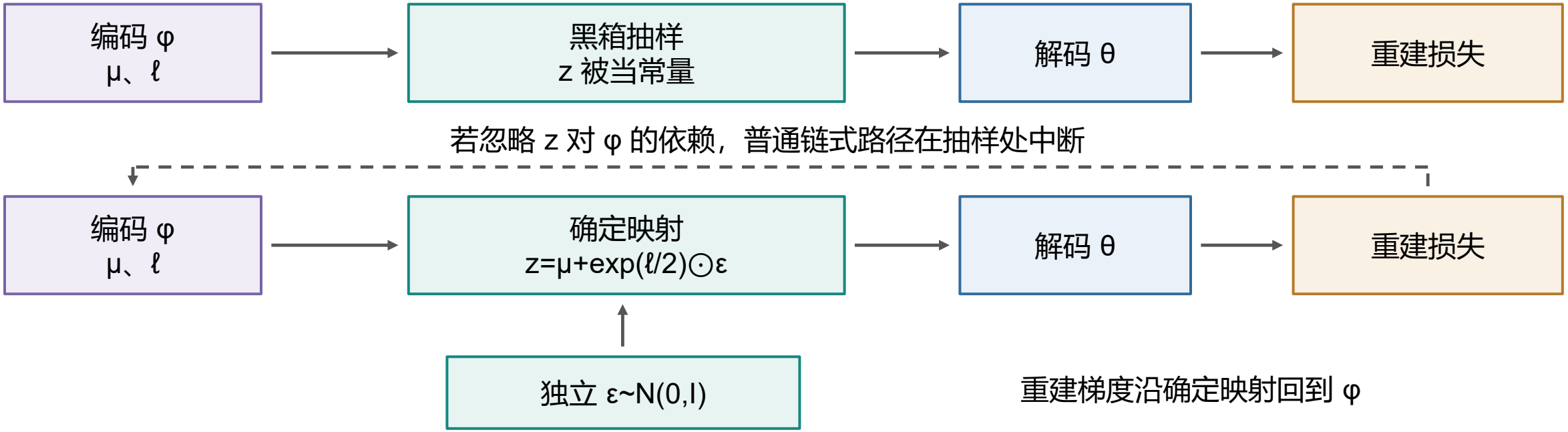
第二维编码分布与标准高斯相同，所以贡献为零。全部维度均为$\mu_j = 0, \sigma_j^2 = 1$时，总KL为零。

对某次潜变量样本，设观测残差$\boldsymbol{x} - g_\theta(\boldsymbol{z}) = (1,-1)$、固定$\tau^2 = 0.5$：$\|\boldsymbol{x} - g_\theta(\boldsymbol{z})\|^2 = 2, \hat{J} = \frac{2}{2\times 0.5} + 0.818147 = 2.818147.$

$\hat{J}$省略固定常数，并以一次抽样近似重建期望。若只把平方和2替换为维度平均1，会得到1.818147，改变重建项与KL的相对比例；将整个损失除以2则为1.409074，二者不同。

随机采样与重建梯度的路径

沿用前面的高斯设定，训练既要保留潜变量的随机性，也要让重建误差反馈给编码参数。两行使用相同的编码分布与解码器，只改变采样的实现方式。



基础噪声分布独立于编码参数；⊙表示逐元素乘法。固定噪声后，样本仍依赖均值和尺度。同分布核对见P22；核心路线接P26。

把参数依赖写进确定映射

高斯目标已经可计算，但编码参数ϕ还决定被抽样的分布。若把抽到的$\boldsymbol{z}$当成与ϕ无关的常数，就会丢掉重建项对编码器的梯度路径。

对高斯编码分布，可改写采样过程： $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),\boldsymbol{z}=\boldsymbol{\mu}_{\phi}(\boldsymbol{x})+\boldsymbol{\sigma}_{\phi}(\boldsymbol{x})\odot\boldsymbol{\epsilon}.$

$\odot$表示逐元素乘法。基础噪声的分布不依赖ϕ；给定一次噪声后，样本值通过均值和标准差依赖ϕ。

由$E\epsilon_j=0$、$Var(\epsilon_j)=1$及高斯变量的仿射变换性质，z_j的均值为μ_j，方差为σ_j^2，不同维仍独立。因此改写前后得到同一对角高斯分布，随机性仍然保留。

固定噪声分布后的期望与梯度

记$h_\phi(\boldsymbol{x},\boldsymbol{\epsilon}) = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}$。改写保持分布，因而对可积函数$f$： $E_{q_\phi}[f(\theta,\phi,\boldsymbol{z})] = E_{\boldsymbol{\epsilon}\sim N(0,I)}[f(\theta,\phi,h_\phi(\boldsymbol{x},\boldsymbol{\epsilon}))]$.

若映射和被积函数几乎处处可微，且参数邻域中的导数有可积函数控制，便可交换求导与期望。链式法则给出 $\nabla_\phi E_q[f] = E_\epsilon\left[\partial_\phi f|_{\boldsymbol{z}} + (\partial_\phi h_\phi)^\top \nabla_{\boldsymbol{z}} f\right]$.

第一项是固定潜变量时的显式参数依赖；第二项是经过潜变量的路径依赖。对本讲的重建项$f = \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$，固定$\boldsymbol{z}$后没有显式$\phi$依赖，但先验KL仍有自己的$\phi$导数。

抽取S个独立基础噪声，用S项平均估计期望与梯度；合法换序及可积条件下这是相应目标的无偏估计，单次值仍会波动。离散抽样通常不能直接套用这个连续可微映射；得分函数估计是另一类方法。

对数方差带来一半的链式因子

编码网络输出$\ell_j = \log\sigma_j^2$，故$\sigma_j = e^{\ell_j/2}$：$\frac{\partial z_j}{\partial \mu_j} = 1, \frac{\partial z_j}{\partial \ell_j} = \frac{1}{2} e^{\ell_j/2}\epsilon_j = \frac{1}{2}\sigma_j\epsilon_j$.

省略固定常数，单个噪声样本的损失为 $\hat{J} = \frac{\|g_\theta(\boldsymbol{z})-\boldsymbol{x}\|^2}{2\tau^2} + \frac{1}{2}\sum_j(\mu_j^2 + e^{\ell_j} - 1 - \ell_j)$.

设$\boldsymbol{a} = \nabla_{\boldsymbol{z}}\|g_\theta(\boldsymbol{z}) - \boldsymbol{x}\|^2/(2\tau^2) = (\partial_{\boldsymbol{z}} g_\theta)^\top(g_\theta(\boldsymbol{z}) - \boldsymbol{x})/\tau^2$。于是 $\frac{\partial \hat{J}}{\partial \mu_j} = a_j + \mu_j, \frac{\partial \hat{J}}{\partial \ell_j} = a_j\frac{\sigma_j\epsilon_j}{2} + \frac{e^{\ell_j}-1}{2}$.

这两组导数再经过编码网络对ϕ使用链式法则。解码参数θ由重建项更新；本讲先验固定，所以解析KL不含θ。

一次噪声样本的梯度核对

采用一维教学模型：$x = 2$、$g(z) = z$、$\mu = 1$、$\sigma = 0.5$、$\tau^2 = 0.5$。固定一次噪声$\epsilon = -1$，得到$z = 1 + 0.5(-1) = 0.5$。

$$\hat{J} = (0.5 - 2)^2 + 0.818147 = 3.068147, a = (z - x)/\tau^2 = -3.$$

$$\frac{\partial \hat{J}}{\partial \mu} = -3 + 1 = -2, \frac{\partial \hat{J}}{\partial \ell} = (-3)(-0.25) + \frac{0.25 - 1}{2} = 0.375.$$

本例还可精确求期望，因为$E[(z - x)^2] = (\mu - x)^2 + \sigma^2$。因此期望损失为$J = 1.25 + 0.818147 = 2.068147$，其导数为 $\partial_\mu J = 2(\mu - x) + \mu = -1, \partial_\ell J = \sigma^2 + \frac{1}{2}(\sigma^2 - 1) = -0.125.$

一次采样的梯度与期望梯度可以不同，甚至分量符号不同。增加适当采样用于降低估计波动，不会把每个样本值变成精确期望。

高斯训练目标与梯度的结论接口

本讲采用$q_\phi = \mathcal{N}(\boldsymbol{\mu}_\phi, diag(e^{\ell_\phi}))$、标准高斯先验，以及固定$\tau^2 > 0$的观测分布$\mathcal{N}(g_\theta(\boldsymbol{z}), \tau^2 I)$。$\ell_j$是对数方差。

$$\boldsymbol{z}^{(s)} = \boldsymbol{\mu}_\phi(\boldsymbol{x}) + e^{\ell_\phi(\boldsymbol{x})/2} \odot \boldsymbol{\epsilon}^{(s)}, \boldsymbol{\epsilon}^{(s)} \sim \mathcal{N}(0, I),$$

$$\hat{J} = \frac{1}{S} \sum_{s=1}^{S} \frac{\left\| \boldsymbol{x} - g_\theta(\boldsymbol{z}^{(s)}) \right\|^2}{2\tau^2} + \frac{1}{2} \sum_{j=1}^{k} (\mu_j^2 + e^{\ell_j} - 1 - \ell_j).$$

解析KL来自对角高斯与标准高斯的计算，完整推导见P18。

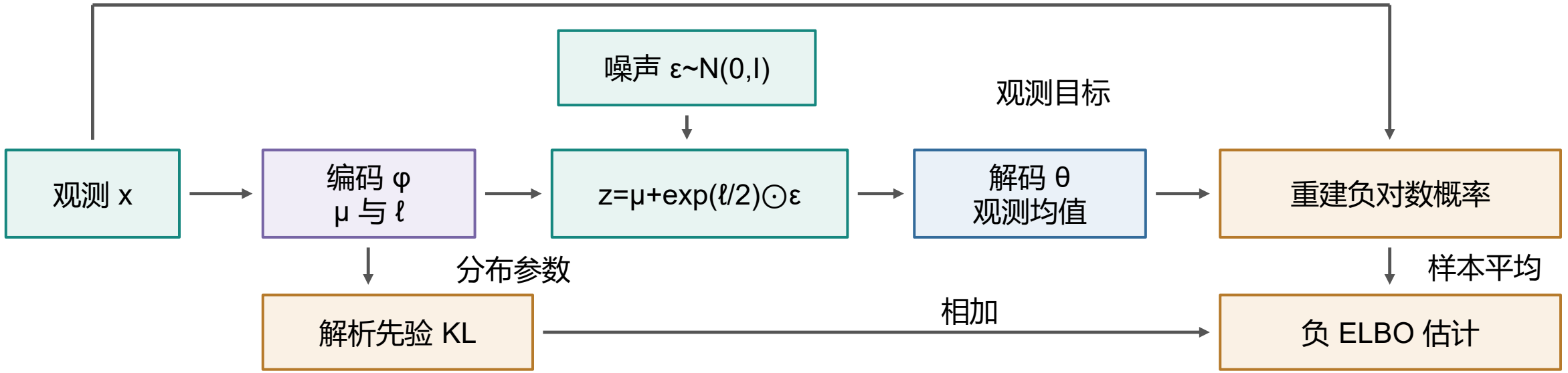
平方误差来自固定方差高斯观测的负对数似然，完整推导见P19。

$\hat{J}$估计省略固定常数后的负ELBO；S为每个观测的噪声样本数。平方误差按观测维求和；换成平均时须相应调整KL权重。

重建项经解码器和$\boldsymbol{z}$回传至编码器，KL直接提供编码梯度；$\partial z_j / \partial \ell_j = e^{\ell_j/2} \epsilon_j / 2$。该路径梯度要求连续可微变换及允许交换求导与期望的可积条件。

两项损失共同连接编码器与解码器

沿用前页的高斯设定、固定先验与观测方差。观测既是编码器输入，也是重建项比较的目标；两项损失在负ELBO估计中相加。



梯度沿依赖反向传播；噪声分布不更新

- 重建项经解码器和潜变量回传至编码器
- 解析KL由编码分布参数直接计算

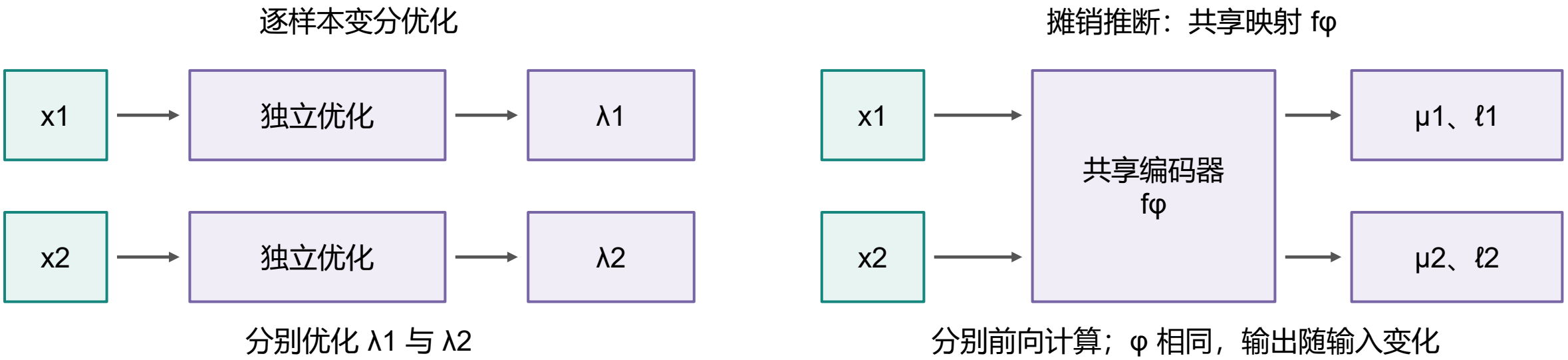
梯度沿计算依赖反传，噪声分布不更新。路径梯度仍要求连续可微变换及可积条件。

4 共享推断与完整训练

连接数据、噪声、批损失及参数更新。

共享网络把推断成本分配到整个数据集

已有可计算目标和梯度后，还需说明不同输入怎样得到各自的近似后验。**摊销推断**（amortized inference）用同一组编码参数ϕ处理多个观测。



不同输入通常产生不同的均值和方差。共享的是计算这些分布参数的映射，不是要求所有输入拥有同一近似后验。

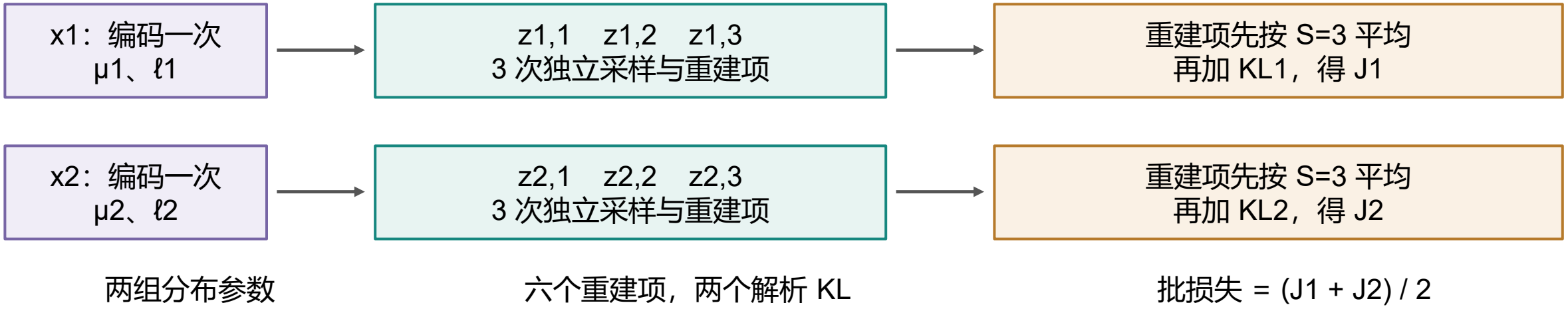
EM中若后验可精确计算，E步使用当前模型的真实后验；VAE在受限分布族中通过共享网络近似推断，并联合更新生成与编码参数。共享推断可能带来近似和优化误差，不保证成为精确E步，也不继承精确EM的似然单调保证。

数据批平均与潜变量采样平均

给定N个观测，训练最大化数据平均ELBO，等价于最小化平均负ELBO。一次训练取B个观测，每个观测取S个独立基础噪声：

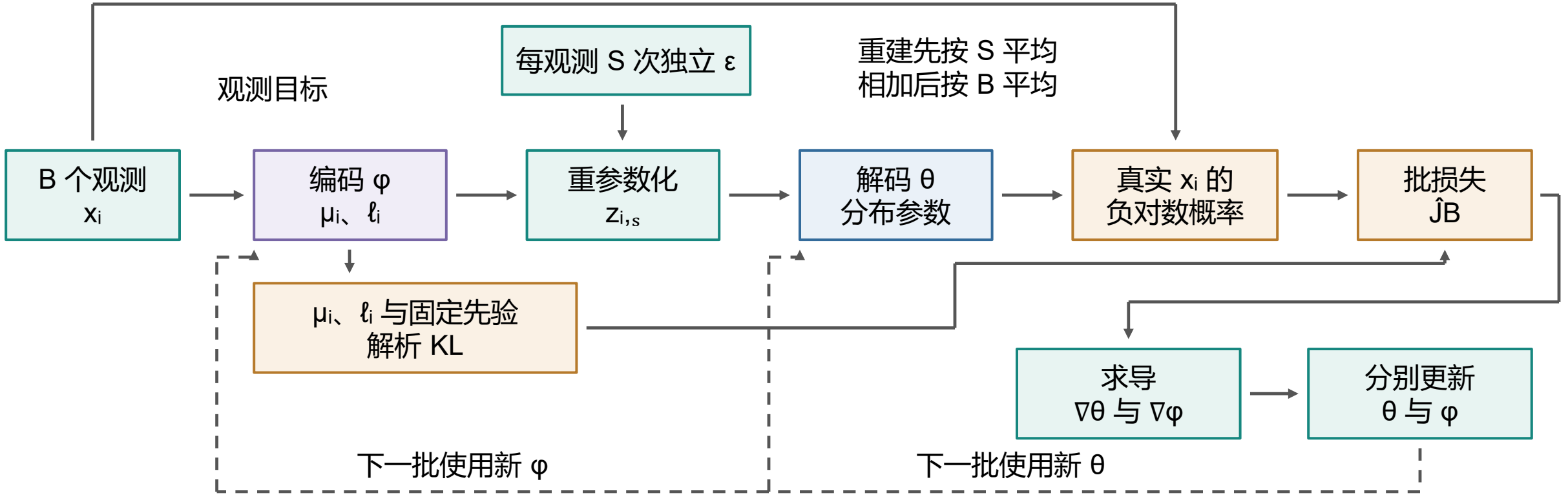
$$\hat{J}_B = \frac{1}{B}\sum_{i=1}^{B}\left[-\frac{1}{S}\sum_{s=1}^{S}\log p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}_{i,s}) + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)\|p(\boldsymbol{z}))\right].$$

数据批平均近似训练数据上的平均；潜变量采样平均近似每个观测内部的期望。若$B = 2, S = 3$，编码网络产生两组分布参数，抽出六个潜变量样本，计算六个重建项和两个解析KL。



一批数据的前向计算与参数更新

每次更新先抽取数据批和基础噪声，再计算重建项、解析KL与批损失。损失对两组网络参数求导，更新后的参数用于下一批数据。



更新两组网络参数，保留固定量

前页批损失对两组网络参数求导，以下更新用于下一批数据：

$$\theta \leftarrow \theta - \eta \nabla_\theta \hat{J}_B, \phi \leftarrow \phi - \eta \nabla_\phi \hat{J}_B, \eta > 0.$$

- 重建项同时影响θ和ϕ，解析KL在本讲只影响ϕ
- 基础噪声的分布、先验和观测方差固定
- 单次求导固定已抽到的噪声值，不对噪声或观测数据做参数更新
  - 下一批重新抽样时噪声值可以改变

有限批和噪声带来估计波动，每次更新不保证提高真实数据似然或精确ELBO。

AEVB算法把目标与梯度组成训练循环

输入为数据集X、批大小B、每个观测的噪声数S和初始化参数(θ, ϕ)。据原论文Algorithm 1中文译写，采用本讲可解析KL的估计器：

1. 从X随机取一批X_B，为每个观测独立抽取S个基础噪声。
2. 编码得到$\boldsymbol{\mu}_i, \ell_i$，重参数化得到$\boldsymbol{z}_{i,s}$；解码并计算前述批损失$\hat{J}_B$。
3. 固定本次数据与噪声，求$\nabla_{\theta,\phi}\hat{J}_B$，按前页下降式更新两组参数。
4. 重复抽批、计算与更新，满足所选停止条件后返回(θ, ϕ)。

原文M, L对应本讲B, S。原文最大化全数据ELBO估计；本讲最小化平均负ELBO，目标符号相反，整体尺度相差N。学习率需与尺度匹配。

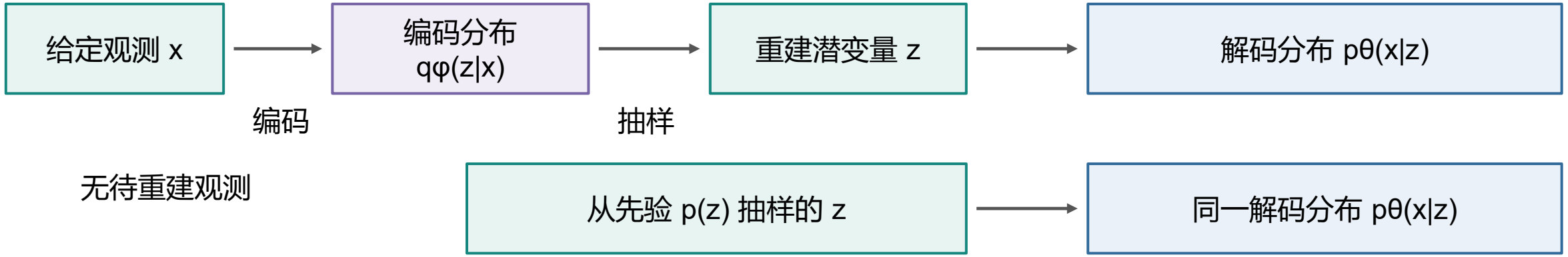
来源：Kingma与Welling，AEVB，Algorithm 1及式(7)—(8)，PDF第4页，arXiv:1312.6114v11。停止条件不等于保证全局最优。

5 重建与先验生成

依据潜变量来源区分重建与先验生成。

训练后两种使用路径

训练得到θ, ϕ后，使用模型时先明确是否给定一个待重建输入。



两行共享同一 θ；可从观测分布抽样，也可展示条件均值

若用编码均值代替抽样，再展示解码均值，可得到一种确定性重建方式；这一般不等于对整个编码分布求解码输出的期望。对非线性g_θ，通常$g_\theta(E_q \boldsymbol{z}) \neq E_q[g_\theta(\boldsymbol{z})]$。

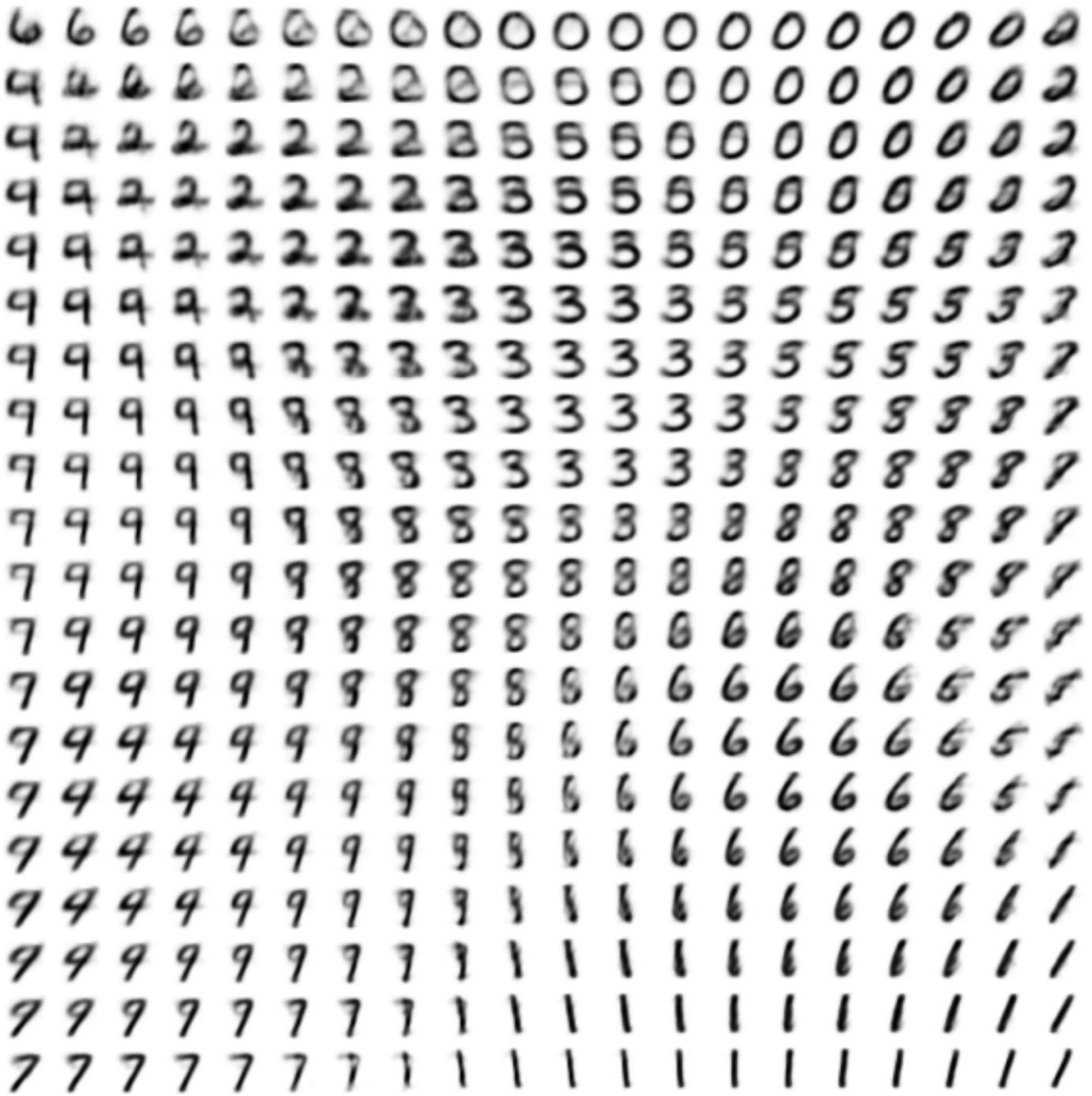
原论文在二维潜空间逐点展示解码结果

固定训练好的θ，在二维潜空间选取网格。每个z独立解码成完整图像。相邻图像的笔画与形状随位置变化。

原文将单位方形上等距坐标经高斯逆累积分布函数变换后得到z。因此图上格点等距不表示高斯潜坐标等距。

这是原文MNIST观测分布的可视化，不代表本讲固定方差高斯模型的实测效果，也不预设每个潜坐标有类别含义。

来源：Kingma与Welling，AEVB，Fig.4(b)，PDF第10页，arXiv:1312.6114v11；原文报告结果，非本课实验。



同一行三个潜变量位置的独立解码

从前图第10行取第1、10、20列，按原顺序放大。三个面板分别标为第1、10、20列；变化的是潜变量坐标以及解码后的笔画形状。

固定同一个θ，保持该行的一个坐标不变，依次改变另一个坐标。机制为：每个位置$\boldsymbol{z}^{(a)}$各自输入同一解码器，得到对应的$p_\theta(\boldsymbol{x} \mid \boldsymbol{z}^{(a)})$可视化。

中间潜变量是数值向量；图中间面板是它的完整解码图像。前一幅图像不输入下一位置，也没有逐步去噪。此处未恢复原作者的精确坐标数值。

来源：AEVB Fig.4(b)，PDF第10页，同一第10行第1/10/20列摘录；仅放大与排版，未合成实验结果。



第 1 列　　第 10 列　　第 20 列

原论文从生成模型随机抽取MNIST样本

与前面的网格选点不同，这里展示原论文从训练好的生成模型随机抽取的样本，不提供待重建输入。先从先验抽取潜变量，再由解码分布生成观测。

图为Fig.5(a)，潜变量维度为2。它补充了“先验随机生成”的实际证据；前面的Fig.4(b)用于观察指定潜变量位置上的输出变化。

样本外观仅展示这一批结果；整体分布匹配和生成质量还需独立评价。

这是原文MNIST实验结果，采用的观测模型与本讲设定不同。

来源：Kingma与Welling，AEVB，Fig.5(a)，arXiv:1312.6114v11。



重建、先验匹配与生成质量分别判断

观察或目标	能说明什么	尚不能单独说明什么
输入重建误差较小	对这些输入的重建表现	从先验生成的整体样本质量
编码分布与先验KL较小	两个指定分布较接近	编码后验一定更准确、潜变量一定保留信息
完整ELBO提高	当前模型与近似推断目标改善	每个潜在点都合理、每维都对应可解释语义

先验生成使用的潜变量来源与重建不同，必须分别检查。ELBO通过概率模型约束学习，没有直接写入“人能解释每个坐标”或“所有样本都符合语义”的保证。

本讲总结

本讲围绕“没有输入时怎样生成新样本”形成四个相互连接的回答。

1. 新样本先从先验抽取潜变量，再经解码分布生成观测。两者定义联合模型，边缘化得到观测的证据。
2. 给定观测，编码器输出近似后验的参数，用来近似由生成模型和Bayes关系确定的真实后验。
3. 训练通过ELBO与重参数化连接概率目标和参数更新。
   - ELBO等于重建对数概率期望减先验KL；后验KL为证据与ELBO的差距
   - 重参数化保留随机性和参数梯度，数据与噪声平均形成随机估计
4. 同一解码器用于重建、遍历和先验生成；潜变量分别来自编码分布、人工选点和先验抽样。

共享推断、较好重建或较小先验KL，均不单独保证精确后验与生成语义质量。