arXiv:2611.32372v1 [cs.LG] 24 Nov 2026

从 ELBO 到最大似然：用于非线性混合效应建模的变分自编码器的收敛性

Jan Rohleff*
康斯坦茨大学数学与统计系，德国
jan.rohleff@uni-konstanz.de

Johannes Schropp
康斯坦茨大学数学与统计系，德国
johannes.schropp@uni-konstanz.de

摘要

非线性混合效应模型是定量药理学的核心统计模型，在药物开发中支持剂量选择、个体化给药和监管决策。从统计学角度看，非线性混合效应建模是一个边缘化似然问题。变分自编码器通过神经网络编码器对难以处理的后验分布进行摊销近似，从而解决此类问题，并支持将人工智能组件集成到非线性混合效应建模中。本文对 VAE-NLME 算法进行了收敛分析。首先，研究证明，其证据下界（ELBO）版本几乎必然收敛到 ELBO 的驻点集合，且不要求编码器经过精确训练。其次，最终的 VAE-NLME 算法通过基于梯度的优化在 ELBO 上训练编码器，并将其用作重要性加权充分统计量的提议分布。在不对编码器训练作任何假设的情况下，该算法几乎必然收敛到边际对数似然的驻点集合。因此，它将变分自编码器的快速基于梯度的训练与最大似然估计的极限相结合。

关键词： 变分自编码器、最大似然估计、随机逼近、非线性混合效应建模、定量药理学、人工智能

1 引言

变分自编码器（VAEs） [29, 43] 是生成式人工智能（AI）的核心方法。它们通过编码器网络拟合潜变量模型，并利用该网络摊销后验推断。本文分析了变分自编码器在非线性混合效应模型（NLME） [40, 50], 中的收敛性；这类模型是药物开发中用于监管决策的定量药理学（PMX）核心模型 [4, 38]。

非线性混合效应模型是一种分层潜变量模型。要根据所有患者的观测值 x 估计总体参数 θ，需要对个体参数 z进行积分消去；也就是说，非线性混合效应建模是一个边缘化似然问题，

$$\max_{\theta} \log \int p(x|z;\theta)\, p(z;\theta)\, dz.$$

In P在定量药理学中，非线性混合效应模型大多由非线性常微分方程（ODE）给出，因此 t the 积分没有闭式形式。行业黄金标准是一阶条件估计 n

*通讯作者。

与交互作用的（FOCEI）方法，可将模型 [36, 54], 线性化；以及随机近似期望最大化（SAEM）算法, 该算法将期望最大化（EM）算法 [19]与马尔可夫链蒙特卡罗（MCMC）抽样 [18, 32]相结合。

为了自动化模型开发并整合更多患者数据（例如医学图像，如CT扫描），人们越来越有兴趣将包含神经网络组件的人工智能模型纳入非线性混合效应建模 [10, 31, 37, 44], ，而成熟方法 FOCEI 和 SAEM 并不适用于此类建模 [49]。因此，研究人员对变分自编码器进行了改造，使其适用于 NLME模型 [46, 47]。 [2, 3, 35, 52]也探讨了将变分自编码器改造为 NLME模型的相关方法。由此得到的基于证据下界（ELBO）的 VAE-NLME 算法，用一个通过随机梯度训练的神经网络编码器取代了 SAEM 中的马尔可夫链蒙特卡罗步骤。该算法速度快，非常适合人工智能应用，但从未经过数学分析。此外，它仅最大化证据下界，因此其极限（若存在）是证据下界的最大值，而不是最大似然估计量。然而，在 PMX 领域，似然是衡量模型质量的硬通货，例如用于模型比较 [17]。无论速度多快，一个只能近似解决似然问题的算法在实践中都无法使用。监管用途也要求进行数学验证 [22, 53]。首先，尽管一般无法证明神经网络训练过程会收敛，该算法究竟能否收敛？其次，能否在保留基于梯度的训练方式的同时，将其扩展为收敛到最大似然估计量？
核心思想是将证据下界和似然用于不同目的。证据下界通过随机梯度训练编码器，而编码器作为重要性加权充分统计量的提议分布，将极限从证据下界移向边际对数似然。不同于 SAEM [18,32], ，该方法无需从后验分布中采样；不同于一般的有偏随机近似 [25], ，它无需对编码器的训练作任何假设即可控制偏差。因此，该算法在保留基于梯度的优化的计算效率的同时，能够收敛到最大似然估计量 [48]。第4节通过三个定理回答了这两个问题，这些定理基于随机逼近的证明 [7, 34] 在第5节中进行了概述。第6节通过数值实验验证了这些定理。

相关工作。 充分统计量的更新采用在线EM [14], 的方法，而使用有偏MCMC采样器的 SAEM 已在 [24] 中得到分析。已有研究证明，在没有编码器的情况下，黑箱变分推断对于固定模型参数能够收敛 [20, 21, 28], ；编码器导致的精度损失，即摊销差距，则在 [16, 39] 中进行了研究。以编码器作为提议分布的重要性抽样也用于重加权唤醒-睡眠算法 [8] 和重要性加权自编码器 [12]。

2 背景

T本节回顾了 [47]中的 NLME模型以及用于非线性混合效应建模的变分自编码器框架。

2.1 NLME模型

考虑一个包含 N个体的数据集。对于个体 $i \in \{1, \ldots\ldots, N\}$和观测值 $j \in \{1, \ldots\ldots, n_i\}$，在已知变量 $t_{ij} \in \mathbb{R}^p$（例如时间）处的观测值 $x_{ij} \in \mathbb{R}$ 建模为

$$x_{ij} = f(t_{ij}, \phi_i) + g(t_{ij}, \phi_i)\, \sigma\, \varepsilon_{ij}, \quad 1 \le i \le N,\ 1 \le j \le n_i. \tag{2.1}$$

其中， f 为非线性结构模型， g 为残差误差模型， $\sigma > 0$为残差误差参数。

2.1.1 个体参数

个体参数 $\phi_i \in \Phi$（其中 $\Phi \subseteq \mathbb{R}^{n_z}$ 表示其可行取值的开集）由个体模型（先验分布）给出

$$\begin{aligned} h(\phi_i) &= z_i, \\ z_i &= z_{\text{pop}} + \beta c_i + \eta_i, \qquad \eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega), \end{aligned}$$

其中 $h:\ \Phi \to \mathbb{R}^{n_z}$ 是一种光滑可逆变换，用于将先验调整为非高斯分布； $z_{\text{pop}} \in \mathbb{R}^{n_z}$ 表示固定效应， $c_i = (c_{i1}, \ldots, c_{in_c})^\top \in \mathbb{R}^{n_c}$ 表示个体 i的协变量，

$\beta = (\beta_1, \ldots, \beta_{n_c}) \in \mathbb{R}^{n_z \times n_c}$ 表示协变量效应矩阵， η_i 表示具有正定协方差矩阵 $\Omega \in \mathbb{R}^{n_z \times n_z}$ 的随机效应，该矩阵描述个体间变异性。将固定效应和 n_β 估计得到的协变量效应汇总于

$$\hat{z}_{\text{pop}} = (z_{\text{pop}}, \beta) \in \mathbb{R}^{n_z + n_\beta},$$

个体模型表示为

$$z_i = C_i \hat{z}_{\text{pop}} + \eta_i, \quad \eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega), \tag{2.2}$$

其中设计矩阵 C_i 具有满行秩 n_z。

2.1.2 误差模型

残差误差 $\varepsilon_{ij} \sim \mathrm{N}(0, 1)$ 相互独立，且与 η_i 独立。因此， x_{ij} 给定 z_i 时服从正态分布，其均值为 $f(t_{ij}, \phi_i)$，标准差为 $g(t_{ij}, \phi_i)\sigma$。

2.1.3 边际似然与目标函数

非线性混合效应模型（2.1）—（2.2）的总体参数为

$$\theta = (\hat{z}_{\text{pop}}, \Omega, \sigma^2).$$

θ 的开参数空间记为 Θ,

$$\Theta = \mathbb{R}^{n_z + n_\beta} \times \mathbb{S}_{++}^{n_z} \times \mathbb{R}_{>0},$$

其中， $\mathbb{S}_{++}^{n_z}$ 表示对称正定 $n_z \times n_z$ 矩阵锥。观测数据 $x = (x_{ij};\ 1 \le i \le N, 1 \le j \le n_i)$ 的似然函数为

$$p(x) = p(x; \theta) = \int p(x, z)\, dz = \int p(x|z) p(z)\, dz, \tag{2.3}$$

其中

$$p(x|z) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} p(x_{ij}|z_i; \theta) \quad \text{and} \quad p(z) = \prod_{i=1}^{N} p(z_i; \theta).$$

相应的对数似然记为

$$\mathcal{LL}(\theta) = \mathcal{LL}(\theta; x) = \log p(x; \theta).$$

目标函数是关于总体参数 θ 最大化对数似然，即

$$\max_{\theta} \mathcal{LL}(\theta), \tag{2.4}$$

这称为边缘化似然问题。

2.2 用于NLME模型的变分自编码器

积分 (2.3) 还对个体参数的后验分布 [5] 进行归一化。变分自编码器用参数化近似 $q(z|x)$ [29, 43], 替代这一难以处理的后验分布

$$q(z|x) \approx p(z|x) = \frac{p(x|z)p(z)}{p(x)}.$$

编码器提供后验分布的变分近似，而解码器则是非线性混合效应模型（2.1）。

2.2.1 编码器

The 编码器用高斯分布近似后验分布 $p(z_i|x_i)$。它是一个神经网络 k with 参数 $\psi \in \Psi$；对于 NLME 模型，通常采用循环网络 [47]。它读取

$$\begin{aligned}(\mu_i, L_i) &= \text{Encoder}_\psi(\mathcal{D}_i), \quad i = 1, \ldots, N,\\ q_\psi(z_i|x_i) &= \mathcal{N}(z_i; \mu_i, \Sigma_i), \quad \Sigma_i = L_i L_i^\top,\end{aligned} \tag{2.5}$$

其中，$\mathrm{D}_i = \big(x_i,\ (t_{ij})_{j=1}^{n_i},\ c_i\big)$是个体 i的数据，具有 $x_i = (x_{ij})_{j=1}^{n_i}$，而 L_i是一个对角元为正的下三角矩阵，因此 Σ_i 是正定的。μ_i 和 L_i 都取决于 ψ。个体参数通过重参数化技巧 [29]，进行采样

$$z_i = \mu_i + L_i \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I). \tag{2.6}$$

The 重参数化允许在采样步骤中进行反向传播，即编码器可以
进行训
练，并采用基于随机梯度的优化。

2.2.2 解码器

解码器是结构模型 (2.1)。对于从 (2.6) 中抽取的个体参数 z_i，模型预测为

$$\hat{x}_{ij}(z_i) = f(t_{ij}, h^{-1}(z_i)) \quad \text{for } i = 1, \ldots, N \text{ and } j = 1, \ldots, n_i.$$

2.2.3 证据下界（ELBO）

由于边缘化似然问题 (2.4) 中的积分难以计算，变分自编码器改为最大化证据下界

$$\mathcal{L}_\psi(x;\theta) \ = \ \mathbb{E}_{z\sim q_\psi(z|x)}\left[\log \frac{p(x|z)p(z)}{q_\psi(z|x)}\right]$$

。对于每个 ψ，均有 [6, 29]

$$\mathcal{LL}(x;\theta) = \mathcal{L}_\psi(x;\theta) + D_{KL}\big(q_\psi(z|x) \,\|\, p(z|x)\big) \ \geq \ \mathcal{L}_\psi(x;\theta), \tag{2.7}$$

其中，$D_{KL}(q\|p) \ = \mathbb{E}_{z\sim q}\big[\log\ (q(z)/p(z))\big] \geq 0$表示 Kullback-Leibler（KL）散度。因此，证据下界是对数似然的下界，其差值为变分近似与后验分布之间的 KL 散度。与对数似然一样，记号中省略了数据，即 $\mathrm{L}\psi(\theta) = \mathrm{L}\psi(x;\ \theta)$。边缘化似然问题 (2.4) 被替换为

$$\max_{\theta,\psi} \mathcal{L}_\psi(x;\theta). \tag{2.8}$$

将似然项与先验项分开，可得

$$\mathcal{L}_\psi(x;\theta) = \mathcal{L}_\psi(\theta) = \sum_{i=1}^{N}\Big(\mathbb{E}_{z_i\sim q_\psi(z_i|x_i)}[\log p(x_i|z_i)] - D_{KL}(q_\psi(z_i|x_i)\|p(z_i))\Big).$$

非线性混合效应模型的似然项为

$$\log p(x_i|z_i) = -\frac{n_i}{2}\log(2\pi) - \sum_{j=1}^{n_i}\log\big(g(t_{ij}, h^{-1}(z_i))\sigma\big) - \frac{1}{2\sigma^2}\sum_{j=1}^{n_i} r_{ij}(z_i)^2,$$

其中

$$r_{ij}(z_i) \ = \ \frac{x_{ij} - \hat{x}_{ij}(z_i)}{g(t_{ij}, h^{-1}(z_i))}$$

是标准化残差。对于高斯分布变分分布和先验分布，KL 散度具有闭式形式

$$\begin{aligned}D_{KL}(q_\psi(z_i|x_i)\|p(z_i)) &= D_{KL}\big(\mathcal{N}(\mu_i, \Sigma_i) \,\|\, \mathcal{N}(C_i\hat{z}_{\text{pop}}, \Omega)\big)\\ &= \frac{1}{2}\Big[\,\text{tr}(\Omega^{-1}\Sigma_i) + (\mu_i - C_i\hat{z}_{\text{pop}})^\top\Omega^{-1}(\mu_i - C_i\hat{z}_{\text{pop}}) - n_z + \log\frac{|\Omega|}{|\Sigma_i|}\Big].\end{aligned} \tag{2.9}$$

2.2.4 证据下界的梯度与蒙特卡洛估计

由于 KL 散度具有闭式形式，使用重参数化（2.6）和 $M \in \mathbb{N}$ 个蒙特卡洛样本对证据下界进行估计，方法如下

$$\widehat{\mathcal{L}}_{\psi}(\theta;\varepsilon) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\log p\big(x_i|\mu_i + L_i\varepsilon_i^{(m)}\big) - \sum_{i=1}^{N} D_{KL}\big(q_{\psi}(z_i|x_i)\,\|\,p(z_i)\big),$$

其中，对于 $m = 1,\ldots,M$，$\varepsilon_i^{(m)} \sim \mathrm{N}(0,I)$ 是独立同分布样本，这些样本收集在噪声变量 $\varepsilon = (\varepsilon_i^{(m)})_{i=1,\ldots,N,\, m=1,\ldots,M}$ 中。由于噪声不依赖于 ψ，对于具有有界导数的光滑模型，$\nabla_{\psi}\,\widehat{\mathrm{L}}\psi(\theta;\,\varepsilon)$ 是 $\nabla_{\psi}\mathrm{L}\psi(\theta)$ 的无偏估计量。

2.2.5 证据下界的指数族结构与充分统计量

分布 $p(x|z)$、$p(z)$ 和 $q\psi(z|x)$ 均为高斯分布，且属于指数族。因此，$\widehat{\ }$证据下界估计量 $\mathrm{L}\psi(\theta;\varepsilon)$ 可分解为充分统计量和自然参数，

$$\widehat{\mathcal{L}}_{\psi}(\theta;\varepsilon) = \big\langle S(x,\psi;\varepsilon), \zeta(\theta)\big\rangle - \kappa(\theta) + \mathcal{H}(\psi;\varepsilon) + C. \tag{2.10}$$

前两项汇集了所有依赖于 θ 的内容。充分统计量

$$S(x,\psi;\varepsilon) = \big(S_1,\, S_2,\, S_3\big) \tag{2.11}$$

具有三个模块

$$S_1 = \big(\mu_i\big)_{i=1}^{N}, \qquad S_2 = \sum_{i=1}^{N}\big(\mu_i\mu_i^{\top} + L_iL_i^{\top}\big), \qquad S_3 = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{n_i} r_{ij}\big(z_i(\varepsilon_i^{(m)})\big)^2,$$

对于 $z_i(\varepsilon_i^{(m)}) = \mu_i + L_i\varepsilon_i^{(m)}$，自然参数为

$$\zeta(\theta) = \Big((\Omega^{-1}C_i\hat{z}_{\mathsf{pop}})_{i=1}^{N},\, -\tfrac{1}{2}\Omega^{-1},\, -\tfrac{1}{2\sigma^2}\Big), \tag{2.12}$$

and

$$\kappa(\theta) = \frac{N}{2}\log|\Omega| + \frac{1}{2}\sum_{i=1}^{N}(C_i\hat{z}_{\mathsf{pop}})^{\top}\Omega^{-1}(C_i\hat{z}_{\mathsf{pop}}) + \frac{N_{\mathrm{tot}}}{2}\log(\sigma^2), \tag{2.13}$$

其中，$N_{\mathrm{tot}} = \sum_{i=1}^{N} n_i$ 表示观测总数。内积按分量理解，$S,\ \zeta = \sum_{i=1}^{N} S_{1,\,i},\ \zeta_{1,\,i} + \mathrm{tr}(S_2^{\top}\zeta_2) + S_3\zeta_3$。熵项汇集了仅依赖于 ψ 和 ε 的所有项，而 C 表示其余常数，

$$\mathcal{H}(\psi;\varepsilon) = \sum_{i=1}^{N}\log\big|\det L_i\big| - \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\big(g(t_{ij}, h^{-1}(z_i(\varepsilon_i^{(m)})))\big),$$
$$C = -\frac{N_{\mathrm{tot}}}{2}\log(2\pi) + \frac{Nn_z}{2}.$$

充分统计量通过重参数化（2.6）依赖于 ε。对噪声取平均可消除这种依赖。记

$$S(\psi) := \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\big[S(x,\psi;\varepsilon)\big], \qquad \mathcal{H}(\psi) := \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\big[\mathcal{H}(\psi;\varepsilon)\big], \tag{2.14}$$

精确的证据下界为

$$\mathcal{L}_{\psi}(\theta) = J\big(\theta, S(\psi)\big) + \mathcal{H}(\psi) + C, \qquad J(\theta,s) = \langle s,\zeta(\theta)\rangle - \kappa(\theta). \tag{2.15}$$

3 VAE-NLME 算法

本节从证据下界问题（2.8）出发，推导 VAE-NLME 算法。该算法在内循环和外循环之间交替进行：内循环更新编码器，外循环更新总体参数。这一结构遵循 EM算法 [5, 19]及其随机逼近版本 SAEM [18]。随后，重要性加权统计量以边际对数似然而非证据下界为目标，从而得到最终算法。

3.1 通用算法框架

给定初始参数 $\psi^{(0)}$和 $\theta^{(0)}$。ELBO VAE-NLME 算法的第 k次迭代包括三个步骤。

1. **模拟步骤：** 对于固定的总体参数 $\theta^{(k-1)}$，相对于编码器近似最大化证据下界。精确解为

$$\arg\max_{\psi\in\Psi} \mathcal{L}_\psi(\theta^{(k-1)}).$$

通过对编码器参数 ψ执行 L 次随机梯度步骤进行近似，

$$\psi_\ell = \psi_{\ell-1} + \alpha_k \nabla_\psi \widehat{\mathcal{L}}_{\psi_{\ell-1}}\big(\theta^{(k-1)}; \varepsilon_\ell\big), \quad \ell = 1, \dots, L, \tag{3.1}$$

其中 $\varepsilon_\ell \sim \mathrm{N}(0, I)$ 是独立同分布样本，且 $\psi^{(k)} = \psi_L$。

2。**随机逼近：** 抽取一个噪声样本 $\varepsilon^{(k)} \sim \mathrm{N}(0, I)$，并更新 $Q^{(k)}(\theta)$ ed as

$$Q^{(k)}(\theta) = Q^{(k-1)}(\theta) + \gamma_k\big(\widehat{\mathcal{L}}_{\psi^{(k)}}(\theta; \varepsilon^{(k)}) - Q^{(k-1)}(\theta)\big). \tag{3.2}$$

3. **最大化步骤（M步）：** 总体参数更新如下

$$\theta^{(k)} = \arg\max_\theta Q^{(k)}(\theta). \tag{3.3}$$

3.2 利用指数族结构

递推式 (3.2)-(3.3) 更新的是 θ 的一个函数。指数族结构可将其化简为充分统计量的递推，其维数记为 $n_{\mathcal{S}} = Nn_z + n_z(n_z+1)/2 + 1$。若辅助函数 $Q^{(0)}$ 初始化为 $s^{(0)}, \zeta(\theta) - \kappa(\theta) + b^{(0)}$，其中 $s^{(0)} \in \mathbb{R}^{n_S}$ 且 $b^{(0)}$ 为标量，则该形式在递推式 (3.2) 中保持不变，

$$Q^{(k)}(\theta) = \big\langle s^{(k)}, \zeta(\theta)\big\rangle - \kappa(\theta) + b^{(k)}, \tag{3.4}$$

其中，均值充分统计量和标量按如下方式更新

$$\begin{aligned} s^{(k)} &= s^{(k-1)} + \gamma_k\Big(S\big(x, \psi^{(k)}; \varepsilon^{(k)}\big) - s^{(k-1)}\Big), \\ b^{(k)} &= b^{(k-1)} + \gamma_k\Big(\mathcal{H}\big(\psi^{(k)}; \varepsilon^{(k)}\big) + C - b^{(k-1)}\Big). \end{aligned} \tag{3.5}$$

推导见附录 A.1。由于 $b^{(k)}$ 不依赖于 θ，M步只需要 $s^{(k)}$，结合式 (2.15) 中的 J，M 步算子定义如下

$$\mathcal{M} : \mathbb{R}^{n_S} \to \Theta, \quad \mathcal{M}(s) := \arg\max_{\theta\in\Theta} J(\theta, s), \tag{3.6}$$

因此 $\theta^{(k)}$=M $(s^{(k)})$。如 EM算法 [19, 第 2]节所述， $J(\cdot, s)$ 是具有充分统计量 s 的指数族的对数似然。如果该族是正则的，例如对于对角 Ω，则当 s 位于充分统计量的凸支撑[11, 定理 5.5]的内部时，最大化元是唯一的。

此外，统计量空间定义为

$$\mathcal{S} := \overline{\mathrm{conv}}\left\{S(x, \psi; \varepsilon) : \psi \in \Psi,\ \varepsilon \in \mathbb{R}^{NMn_z}\right\}, \tag{3.7}$$

以及通过 M步可达到的相应参数集合，

$$\Theta_{\mathcal{S}} := \mathcal{M}(\mathcal{S}).$$

3.3 ELBOVAE-NLME 算法

算法不再更新函数 $Q^{(k)}(\theta)$，而是更新均值充分统计量。第 3.1 节的步骤 2 和 3 如下所示。

2. **随机逼近：** 抽取一个噪声样本 $\varepsilon^{(k)} \sim \mathrm{N}(0,\ I)$，计算充分统计量$S(x,\ \psi^{(k)};\ \varepsilon^{(k)})$，并按以下方式更新均值充分统计量

$$s^{(k)} = s^{(k-1)} + \gamma_k\Big(S\big(x, \psi^{(k)}; \varepsilon^{(k)}\big) - s^{(k-1)}\Big).$$

3。**最大化步骤：** 使用 M 步算子（3.6）更新总体参数，

$$\theta^{(k)} = \mathcal{M}(s^{(k)}).$$

该算法称为 ELBO VAE-NLME 算法，因为它最大化 ELBO（2.8）。

3.4 重要性加权统计量

根据（2.7），ELBO VAE-NLME 算法最大化的 ELBO 与边际对数似然相差一个取决于 θ 的 KL 散度。因此，其最大值通常并非最大似然估计量。ELBO 仅通过统计量进入 M 步。对于$z = (z_i)_{i=1}^N$，完整数据统计量为

$$S(x,z) \ = \ \Big((z_i)_{i=1}^N,\ \sum_{i=1}^N z_i z_i^\top,\ \sum_{i=1}^N \sum_{j=1}^{n_i} r_{ij}(z_i)^2 \Big), \tag{3.8}$$

即，$\log\ p(x,\ z;\ \theta) = S(x,\ z),\ \zeta(\theta) - \kappa(\theta) + c(x,\ z)$，其中 ζ、κ取自 (2.12)-(2.13)，且 c 独立于 θ。式 (2.14) 中的平均变分统计量 $S(\psi)$ 是其在 q_ψ 下的期望。与 EM算法 和 SAEM [18, 19], 中一样，这一期望由精确后验分布下的期望替代，

$$\bar{S}_p(\theta) \ = \ \mathbb{E}_{z\sim p(\cdot|x;\theta)}\big[S(x,z)\big].$$

¯后验统计量 $S_p(\theta)$ 通过重要性抽样估计，并以编码器作为提议分布。在第 k次迭代中，对每个个体抽取 K 个样本 $z_{i1},\ \ldots\ldots,\ z_{iK}$，它们来自多元 t 分布$r_{(k)_i} \ = t_\nu(\mu_i,\ L_i L_i^\top)$，该分布的自由度为 ν，编码器输出为 $\psi^{(k)}$的 $(\mu_i,\ L_i)$。自归一化权重和加权统计量为

$$\begin{aligned} w_{ij} &= \frac{p(x_i, z_{ij}; \theta^{(k-1)})}{r_i^{(k)}(z_{ij})}, \qquad \tilde{w}_{ij} = \frac{w_{ij}}{\sum_{j'=1}^K w_{ij'}}, \\ \widehat{S}_K &= \Big(\Big(\sum_{j=1}^K \tilde{w}_{ij} z_{ij} \Big)_{i=1}^N,\ \sum_{i,j} \tilde{w}_{ij}\, z_{ij} z_{ij}^\top,\ \sum_{i,j} \tilde{w}_{ij} \sum_{l=1}^{n_i} r_{il}(z_{ij})^2 \Big). \end{aligned} \tag{3.9}$$

该 t 提议分布的尾部比后验分布更重，因此权重有界 [23, 41]。在第3.3节的随机逼近中，加权统计量 S_K 取代了变分统计量 $\widehat{S(x,\ \psi^{(k)};\ \varepsilon^{(k)})}$。模拟步骤和M步保持不变。额外成本为每个个体每次迭代需要进行 K次解码器评估。

3.5 FinalVAE-NLME 算法

最终的 VAE-NLME 算法结合了第 3.1 节的模拟步骤与第 3.4 节的加权统计量。算法 1 对其进行了总结。

算法 1VAE-NLME 算法

1：输入：初始参数 $\psi^{(0)}$**、** θ**(0)、** $s^{(0)}$**，步长 {**α_k**}、{**γ_k**}，内循环长度** L**，样本数量** K **2：对于** $k = 1$**、2、.....，执行 3：设置** $\psi_0 = \psi^{(k-1)}$ **4：模拟步骤： 5：对于 l** $= 1$**、.....、** L**，执行 6：** $\psi_\ell = \psi_{\ell-1} + \alpha_k \nabla_\psi \widehat{\mathbf{L}}_{\psi_{\ell-1}}(\theta^{(k-1)}; \varepsilon_\ell)$**，其中** $\varepsilon_\ell \sim \mathbf{N}(0, I)$ **独立同分布 7：结束循环 8：设置** $\psi^{(k)} = \psi_L$ **9：随机逼近： 10：设置** $r(k)_i = t_\nu(\mu_i, L_i L_i^\top)$**，其中** (μ_i, L_i) **=编码器**$\psi^{(k)}(\mathbf{D}_i)$**，对于** $i = 1$**、.....、** N **11：抽取** z_{i1}**、.....、** $z_{iK} \sim r_i^{(k)}$**，对于** $i = 1$**、.....、** N**，并根据 (3.9) 计算** $\widehat{S}_K$ **12：更新 13：最大化步骤：** $\theta^{(k)} = \mathbf{M}(s^{(k)})$ **14：结束循环**

$$s^{(k)} = s^{(k-1)} + \gamma_k \left(\widehat{S}_K - s^{(k-1)} \right)$$

备注1 (实际实现)。 一次运行会依次经过四个阶段 *(a)–(d)*，并由两个迭代次数 $K_{\mathrm{KL}} \le K_{\mathrm{sm}}$ 将它们分隔开。

(a) 预热$(k \le 0)$。使用较小的 *KL* 项权重对编码器进行预训练，使得最初的充分统计量已包含有关数据的信息。*(b)* KL 退火$(1 \le k < K_{\mathrm{KL}})$ *[9, 51]*。*KL* 项(2.9) 的权重会在K_{KL}增至 *1*，使编码器先拟合数据。与 *(a)* 阶段结合后，这可以降低陷入较差局部最优的风险。*(c)* 探索阶段 $(K_{\mathrm{KL}} \le k \le K_{\mathrm{sm}})$。此处 $\gamma_k = 1$，即 $s^{(k)}$ 等于当前迭代的统计量。这会加快收敛，同时总体参数会波动。*(d)* 平滑阶段 $(k > K_{\mathrm{sm}})$。步长 $\gamma_k = 1/(k - K_{\mathrm{sm}})$ 满足 *Robbins-Monro* 条件 *[45]*，并对 *(c)* 阶段的波动取平均，同时 α_k 和 L 保持不变。*(e)* 优化器。 梯度步(3.1) 可以替换为其他优化器，例如 *Adam*。

计数 K_{KL} 和 K_{sm} 均为有限值，且可自适应地选择。

4 主要结果

本节陈述三个主要定理。

4.1 变分参数与假设

4.1.1 变分参数

定义 1（变分参数）。这些变分参数 为

$$\mathsf{q} = (\mathsf{q}_1, \ldots, \mathsf{q}_N), \qquad \mathsf{q}_i = (\mu_i, L_i) \in \mathbb{R}^{n_z} \times \mathcal{T}_+,$$

其中，$\mathcal{T}_+$ 表示对角元为正的下三角 $n_z \times n_z$ 矩阵集合，而编码器(2.5) 所实现的摊还映射为

$$\mathsf{q}(\cdot) : \Psi \to \mathbb{R}^{N n_z} \times \mathcal{T}_+^N, \qquad \psi \mapsto \mathsf{q}(\psi) = \left(\mathrm{Encoder}_\psi(\mathcal{D}_i) \right)_{i=1}^N.$$

算法的所有量都仅通过编码器的输出$\mathsf{q}(\psi)$依赖于编码器,

$$\mathcal{L}_\psi(\theta) = \mathcal{L}\big(\theta, \mathsf{q}(\psi)\big), \qquad S(x,\psi;\varepsilon) = S\big(x, \mathsf{q}(\psi);\varepsilon\big), \qquad \mathcal{H}(\psi;\varepsilon) = \mathcal{H}\big(\mathsf{q}(\psi);\varepsilon\big).$$

因此，以下假设约束的是编码器输出，而不是网络参数 ψ。由于 (2.15) 中的 H 和 C 与 θ 无关，总体参数只通过代理函数 J 进入证据下界,

$$\nabla_\theta \mathcal{L}(\theta, \mathsf{q}) \;=\; \nabla_\theta J\big(\theta, S(\mathsf{q})\big) \qquad \text{for every } \mathsf{q}. \tag{4.1}$$

全文中，$\|\cdot\|$ 表示欧几里得范数。作为 θ、s 或 q 组成部分的矩阵，均按其元素计量，即 $\|\theta\|^2 = \|\hat{z}_{\mathsf{pop}}\|^2 + \|\Omega\|_F^2 + (\sigma^2)^2$ 对于 $\theta = (\hat{z}_{\mathsf{pop}},\ \Omega,\ \sigma^2)$ 使用弗罗贝尼乌斯范数$\|\cdot\|_F$；对于线性映射，$\|\cdot\|$ 表示诱导算子范数。

4.1.2 假设

Ass假设 1 (正则性)。 *(i)* 有界模型。 对于所有 i 和 j， f 有界，且 $g \geq g_{\min} >$ 0.

(ii) 光滑模型。 对于所有 i 和 j，以下两个映射

$$z \mapsto f\big(t_{ij}, h^{-1}(z)\big) \qquad \textit{and} \qquad z \mapsto g\big(t_{ij}, h^{-1}(z)\big)$$

均为三次连续可微，且最高至三阶的导数有界。

（*iii*）变分参数。 编码器输出取值于紧凸集

$$\mathcal{Q} = \Big\{\mathsf{q} = (\mu_i, L_i)_{i=1}^N :\ \|\mu_i\| \leq \mu_{\max},\ \ell_{\min} \leq (L_i)_{aa} \leq \ell_{\max},\ |(L_i)_{ab}| \leq \ell_{\max}\Big\}$$

其中 $0 < l_{\min} \leq l_{\max} < \infty$ 和 $\mu_{\max} < \infty$ 为常数。特别地，存在 $0 < \tau_{\min} \leq \tau_{\max} < \infty$，使得对于所有 $\mathsf{q} \in \mathcal{Q}$ 以及所有 i，都有 $\tau_{\min} I \preceq \Sigma_i \preceq \tau_{\max} I$。由于 $\mathsf{q}(\Psi) \subseteq \mathcal{Q}$，统计量空间 (3.7) 从此定义为 $\mathcal{S} = \mathrm{conv}\{S(x,\ \mathsf{q};\ \varepsilon):\ \overline{\mathsf{q} \in \mathcal{Q}},\ \ \varepsilon \in \mathbb{R}^{NMn_z}\}$。

（*iv*）非消失残差。 至少有一个个体无法被精确拟合，即

$$s_{\min} \;:=\; \sum_{i=1}^{N} \inf_{z \in \mathbb{R}^{n_z}} \sum_{j=1}^{n_i} r_{ij}(z)^2 \;>\; 0.$$

假设 2 （M步局部极大值点）。 存在一个连续可微映射 $\mathcal{M}:\ \mathcal{S} \to \Theta$，使得对于每个 $s \in \mathcal{S}$,

$$\nabla_\theta J\big(\mathcal{M}(s), s\big) \;=\; 0, \tag{4.2}$$

$$\nabla_\theta^2 J\big(\mathcal{M}(s), s\big) \;\prec\; 0. \tag{4.3}$$

假设3 （变分局部极大值）。 存在常数 $\mathcal{K}_{\min} > 0$、$r \in (0,\ l_{\min})$、一个开集 $\mathcal{O} \supseteq \Theta_{\mathcal{S}}$ 以及一个连续可微映射

$$\mathsf{q}^\star :\ \mathcal{O} \;\longrightarrow\; \mathcal{Q}, \qquad \theta \;\longmapsto\; \mathsf{q}^\star(\theta) = \big(\mu_i^\star(\theta), L_i^\star(\theta)\big)_{i=1}^N,$$

使得对于每个 $\theta \in \Theta_{\mathcal{S}}$

$$\nabla_{\mathsf{q}} \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big) \;=\; 0, \tag{4.4}$$

$$\nabla_{\mathsf{q}}^2 \mathcal{L}\big(\theta, \mathsf{q}\big) \;\preceq\; -\mathcal{K}_{\min} I \qquad \textit{for all } \mathsf{q} \in \mathcal{U}(\theta), \tag{4.5}$$

其中 $\mathcal{U}(\theta) :\ =\ \big\{\mathsf{q} :\ \big\|\mathsf{q} - \mathsf{q}^\star(\theta)\big\| \leq r\big\}$。定义

$$\mathcal{L}^\star(\theta) \;:=\; \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big), \qquad \bar{S}(\theta) \;:=\; S\big(\mathsf{q}^\star(\theta)\big).$$

假设 4 （步长）。外层步长满足 $\gamma_k \in (0,1]$, γ_k 单调不增，且 $\gamma_{k-1}/\gamma_k \to 1$，并满足 *Robbins-Monro* 条件 *[45]*

$$\sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty.$$

假设5 （内循环）。*(i)* 压缩。 存在 $\rho \in [0,1)$ 和 $c_0 < \infty$，使得对于每个 $k \geq 1$，几乎必然有：模拟步骤的编码器输出 $\mathsf{q}^{(k)} = \mathsf{q}(\psi^{(k)})$ 的内循环误差

$$\delta_k \;:=\; \left\|\mathsf{q}^{(k)} - \mathsf{q}^\star(\theta^{(k-1)})\right\| \;\leq\; \rho \left\|\mathsf{q}^{(k-1)} - \mathsf{q}^\star(\theta^{(k-1)})\right\| \;+\; c_0\,\alpha_k. \tag{4.6}$$

(ii) 耦合步长。 内层步长满足 $\alpha_k \leq c_\alpha \gamma_k$，其中 $c_\alpha < \infty$。

假设 6 （加权统计量的紧迭代序列）。 *VAE-NLME*（算法 *1*）的迭代序列 $s^{(k)}$，即使用加权统计量 (3.9)时，几乎必然保持在一个紧凸集$\mathcal{S}_p \subset \mathbb{R}^{n_S}$ 内，在该集合上 $s_3 > 0$且

$$\Sigma^{\mathsf{tot}}(s) \;:=\; s_2 - \sum_{i=1}^{N} s_{1,i} s_{1,i}^\top \;\succ\; 0 \qquad \textit{for } s = (s_1, s_2, s_3),\; s_1 = (s_{1,i})_{i=1}^N.$$

令 $\Theta_{\mathcal{S}_p} :=\mathcal{M}(\mathcal{S}_p)$。

4.2 收敛到证据下界的驻点

前两个定理讨论第3.3节中的 ELBO VAE-NLME 算法。

定理1 （ELBO VAE-NLME 算法的收敛性）。 设假设 *1-5* 成立，并令$\{(s^{(k)}, \theta_{(k)})\}$ 由 *ELBO VAE-NLME* 算法生成。那么集合

$$\Theta^\star \;:=\; \left\{\theta \in \Theta_{\mathcal{S}} :\; \nabla_\theta \mathcal{L}^\star(\theta) = 0\right\}$$

非空且紧，并且，几乎必然，

$$\lim_{k\to\infty} \left\|\nabla_\theta \mathcal{L}^\star\left(\theta^{(k)}\right)\right\| = 0 \qquad \textit{and} \qquad \lim_{k\to\infty} \min_{\theta^\star \in \Theta^\star} \left\|\theta^{(k)} - \theta^\star\right\| = 0.$$

对于$\{\theta^{(k)}\}$的每个极限点 $\theta^\star$，数对 $\left(\theta^\star、\mathsf{q}^\star(\theta^\star)\right)$都是证据下界(2.8)的驻点。此外，如果 $\Theta^\star$ 中的元素是孤立的，那么 $\theta_{(k)}$几乎必然收敛到单点 $\theta^\star \in \Theta^\star$，且 $s^{(k)} \to \bar{S}(\theta^\star)$。

如果内循环不收缩，内循环误差不一定会消失。在这种情况下，可以得到一个定量结论。

定理 2 （持续的内循环不准确性）。 考虑 *ELBO VAE-NLME* 算法，设假设 *1-4* 成立，并假设不采用假设*5*，而是要求由(4.6)给出的内循环误差 δ_k 满足

$$\limsup_{k\to\infty} \delta_k \;\leq\; \delta$$

对于某个 $\delta \geq 0$。那么存在一个与 δ 无关的常数 $C < \infty$，使得以下结论几乎必然成立。

(i) 下限。

$$\liminf_{k\to\infty} \left\|\nabla_\theta \mathcal{L}^\star\left(\theta^{(k)}\right)\right\| \;\leq\; C\,\delta. \tag{4.7}$$

(ii) 局部收敛。 如果 $\theta^\star \in \Theta^\star$ 是$\mathcal{L}^\star$ 的局部极大值点，且 $m = \lambda_{\min}\big(-\nabla^2_\theta \mathcal{L}^\star(\theta^\star)\big) > 0$，那么存在 $\theta^\star$ 的一个邻域 $\mathcal{U}$，使得对于充分小的 δ，在 $\theta^{(k)} \in \mathcal{U}$ 对无穷多个 k 成立的事件上，

$$\limsup_{k\to\infty} \big\|\theta^{(k)} - \theta^\star\big\| \;\le\; C_\delta, \qquad C_\delta \;:=\; \frac{2C}{m}\,\delta. \tag{4.8}$$

备注 2（固定内循环）。 定理 *1* 未涵盖该算法的常见实现方式。对于长度固定为 L 的随机内循环（$c_0 > 0$），假设*5*的两个部分互不相容。当 $\alpha_k \to 0$ 时，收缩因子 $(1-\alpha_k \mathcal{K}_{\min})^L$ 趋近于 *1*；而对于固定的 $\alpha_k = \alpha$，内循环误差稳定在 $c_0\alpha/(1-\rho) > 0$。两个部分同时成立要求 $L \propto 1/\gamma_k$。定理 *2* 涵盖了实践中使用的固定内循环，且不需要满足其中任何一个部分。

4.3 收敛至对数似然的驻点

第三个定理针对 VAE-NLME（算法 1）。定理 1 和定理 2 针对证据下界，而我们关注的目标函数是边际对数似然（2.4）。由于（2.7）对每个q都成立，将其应用于$\mathsf{q}^\star(\theta)$可得

$$\text{对数似然}\,(\theta) \;=\; \mathrm{L}^\star(\theta) \;+\; \mathrm{G}_{\mathrm{var}}(\theta), \qquad \mathrm{G}_{\mathrm{var}}(\theta) :\; = \sum_{i=1}^{N} D_{KL}\Big(q_{\mathsf{q}_i^\star(\,\theta)}\,(\cdot\;|x_i)\;\Big\|\;p(\cdot\;|x_i;\,\theta)\Big) \;\ge\; 0, \tag{4.9}$$

因此，在定理1的任意极限点 $\theta^\star$ 处，

$$\nabla_\theta \mathcal{LL}(\theta^\star) \;=\; \nabla_\theta \mathcal{G}_{\mathrm{var}}(\theta^\star). \tag{4.10}$$

如果高斯族包含真实后验分布，则 $\mathrm{G}_{\mathrm{var}} \equiv 0$。加权统计量 (3.9) 消除了 (4.10) 的右侧，而不改变编码器或 M步。其准确性取决于编码器对后验分布 [1, 15]的近似程度。令 θ_{ML} 表示一个最大似然估计量，即 (2.4) 的解。

定理3（VAE-NLME 算法的收敛性）。 设假设 *1(i)* 和 *(iii)*、以 $\mathcal{S}^p$ 代替 $\mathcal{S}$ 的假设 *2*，以及假设*4*和假设*6*均成立，并设 $\{(s^{(k)}, \theta_{(k)})\}$ 由 *VAE-NLME*（算法 *1*）生成。那么存在一个常数 $C < \infty$，与 K 无关，且仅通过迭代序列中的权重矩依赖于编码器，使得以下结论几乎必然成立。

(i) 下限。

$$\liminf_{k\to\infty} \big\|\nabla_\theta \mathcal{LL}\big(\theta^{(k)}\big)\big\| \;\le\; \frac{C}{K}. \tag{4.11}$$

(ii) 增长的样本量。 若 $K = K_k \to \infty$ 满足 $\sum_k \gamma_k/K_k < \infty$，则 $\nabla_\theta \mathcal{LL}(\theta_{(k)}) \to 0$，且 $\theta_{(k)}$ 收敛到 $\Theta_{\mathcal{S}_p}$ 中 LL 的驻点集合。若 θ_{ML} 是其中唯一的驻点，则 $\theta^{(k)} \to \theta_{\mathrm{ML}}$。

(iii) 局部收敛。 若 $\theta_{\mathrm{ML}} \in \Theta_{\mathcal{S}_p}$ 满足 $m = \lambda_{\min}\big(-\nabla^2_\theta \mathcal{LL}(\theta_{\mathrm{ML}})\big) > 0$，则存在 θ_{ML} 的一个邻域$\mathcal{U}$，使得在 $\theta^{(k)} \in \mathcal{U}$ 对无穷多个 k 成立的事件上，

$$\limsup_{k\to\infty} \big\|\theta^{(k)} - \theta_{\mathrm{ML}}\big\| \;\le\; C_K, \qquad C_K \;:=\; \frac{2C}{m\,K}, \tag{4.12}$$

如果 C_K 足够小，且当 $K = K_k \to \infty$ 时 $\theta^{(k)} \to \theta_{\mathrm{ML}}$。

与定理 1 和定理 2 不同，定理3无需对编码器训练作出任何假设，既不需要假设3，也不需要假设5；同时也不要求模型具备光滑性，即不需要假设 1(ii)。编码器仅作为提议分布，并影响常数 C。

4.4 讨论

4.4.1 假设

这些假设是随机逼近和 EM 类算法收敛分析中通常采用的条件 [18, 34], , 并扩展了关于编码器的条件。其中每一项对于典型的 NLME模型都成立，或可通过实现加以保证。

• 模型。 对于通过右端平滑的常微分方程定义的结构模型，假设 1(i)-(ii)通常自然成立，前提是将个体参数限制在有界且生理上合理的范围内，这也是实际非线性混合效应建模中的标准做法。对于含噪数据，条件 (iv)通常自然成立，因为没有任何个体会被精确拟合。

• M步。 假设 2通常自然成立，因为 $J(\cdot, s)$对应于指数族的对数似然（第 3.2 节）。对于正则指数族（例如使用对角 Ω时），最大化元存在且唯一，并且在凸支撑 [11, 的内部随 s 平滑变化（定理 5.5]）。

• E步。 假设3对E步施加了类似的正则性条件。它要求关于q 的证据下界存在一个海森矩阵负定的局部极大值点。该条件与个体参数的局部可辨识性自然相关。在PMX中，结构模型的开发通常旨在确保能够根据现有数据辨识个体参数。通常会移除无法可靠辨识的参数，或者在模型开发过程中以其他方式简化模型。因此，假设3体现了设定良好且可辨识的NLME模型所具有的一种理想自然属性。

• 步长。 假设 4 包含随机逼近的经典 Robbins-Monro 条件 [45], , 所选的步长序列满足这些条件。假设5是唯一专门针对编码器训练的假设。根据假设3，对变分参数进行梯度更新即可满足第(i)部分。第(ii)部分确保内循环能够跟踪不断变化的目标 $\mathsf{q}_\star(\theta_{(k-1)})$，这是双时间尺度随机逼近 [7]所要求的。重要的是，即使不满足假设5，定理 2仍然适用。

• 稳定性。 假设 6 仅适用于加权统计量。变分统计量$\widehat{}$是有界统计量的凸组合，因此按构造始终位于 S 中。相比之下， S_K无界；如果某个权重占据主导地位，即加权统计量集中到单个样本上， $\Sigma_{\text{tot}}(\widehat{S}_K)$就会变得奇异。与 SAEM [18, 定理 5], [32], 类似，因此假定迭代序列具有稳定性，并且可以通过重新初始化 [32]来确保这一点。

4.4.2 编码器的作用

该分析是针对变分参数 q 而非编码器参数ψ来表述的。事实上，根据链式法则，

$$\nabla_\psi \mathcal{L}\big(\theta, \mathsf{q}(\psi)\big) = D\mathsf{q}(\psi)^\top \nabla_{\mathsf{q}} \mathcal{L}\big(\theta, \mathsf{q}(\psi)\big).$$

对于过参数化神经网络， $D\mathsf{q}(\psi)$ 通常是秩亏的。因此，关于 ψ 的平稳性并不一定意味着关于q的平稳性。因此，编码器参数仅通过近似误差 δ_k进入收敛分析。

为刻画这一误差，令 $\mathrm{L}_\Psi^\star(\theta)$ 表示在满足$\mathsf{q}(\psi) \in \mathrm{U}(\theta)$的条件下，编码器所能实现的最佳证据下界。其差值

$$\mathcal{G}_{\text{am}} = \mathcal{L}^\star - \mathcal{L}_\Psi^\star \geq 0$$

即摊销差距，它量化了将变分参数限制为编码器可表示的参数所导致的证据下界损失。在 $\theta = \theta_{(k-1)}$ 处，推断间隙可分解为三个部分 [16],

$$\underbrace{\mathcal{LL}(\theta) - \mathcal{L}\big(\theta, \mathsf{q}^{(k)}\big)}_{\text{inference gap}} = \underbrace{\mathcal{LL}(\theta) - \mathcal{L}^\star(\theta)}_{\text{(a) variational family gap}} + \underbrace{\mathcal{G}_{\text{am}}(\theta)}_{\text{(b) amortization gap}} + \underbrace{\mathcal{L}_\Psi^\star(\theta) - \mathcal{L}\big(\theta, \mathsf{q}^{(k)}\big)}_{\text{(c) optimization gap}}$$
$$= \mathcal{G}_{\text{var}}(\theta) + \mathcal{G}_{\text{am}}(\theta) + \mathcal{G}_{\text{opt}}(\theta).$$

部分 (a) 是式 (4.9) 的变分间隙，即高斯族的近似误差，不影响定理1。部分 (b) 和 (c) 由 δ_k 衡量。一旦 $\mathsf{q}^{(k)} \in \mathrm{U}(\theta_{(k-1)})$，在 $\mathsf{q}^\star(\theta_{(k-1)})$ 附近进行泰勒展开，可得，其中 $\mathrm{K}_{\min}$ 来自假设3，$\mathrm{K}_{\max} < \infty$ 是 $\|\nabla^2_{\mathsf{q}} \mathrm{L}\|$ 在 Q 上的上界，

$$\frac{\mathcal{K}_{\min}}{2}\,\delta_k^2 \;\le\; \mathcal{L}^\star\big(\theta^{(k-1)}\big) - \mathcal{L}\big(\theta^{(k-1)}, \mathsf{q}^{(k)}\big) \;\le\; \frac{\mathcal{K}_{\max}}{2}\,\delta_k^2. \tag{4.13}$$

因此，持续的摊销间隙以$\sqrt{\mathrm{G}_{\mathrm{am}}}$ +$\mathrm{G}^{\mathrm{opt}}$的阶对定理 2 中的残差作出贡献，其中 $\mathrm{G}^{\mathrm{opt}}$ 表示 (c) 中优化差距的上极限。若内层优化问题关于 ψ 被精确求解，则残差减小为 $\sqrt{\mathrm{G}_{\mathrm{am}}}$。

4.4.3 与 SAEM 的关系

VAE-NLME 算法是一种 SAEM 算法 [18]，具有相同的统计量、M步和李雅普诺夫函数 $-\mathrm{LL} \circ \mathrm{M}$。只有模拟步骤有所不同。SAEM 通过马尔可夫链蒙特卡罗 [32, 33] 从精确后验分布中抽取个体参数，而 VAE-NLME 算法则从编码器中抽取这些参数并对其加权。因此，模拟步骤是可微的，模型的神经网络组件也在同一次迭代中进行训练。

5 证明策略

定理 1 和定理 2 的证明遵循算法的两个部分：内循环和外部更新。图 1 概述了这四个步骤。

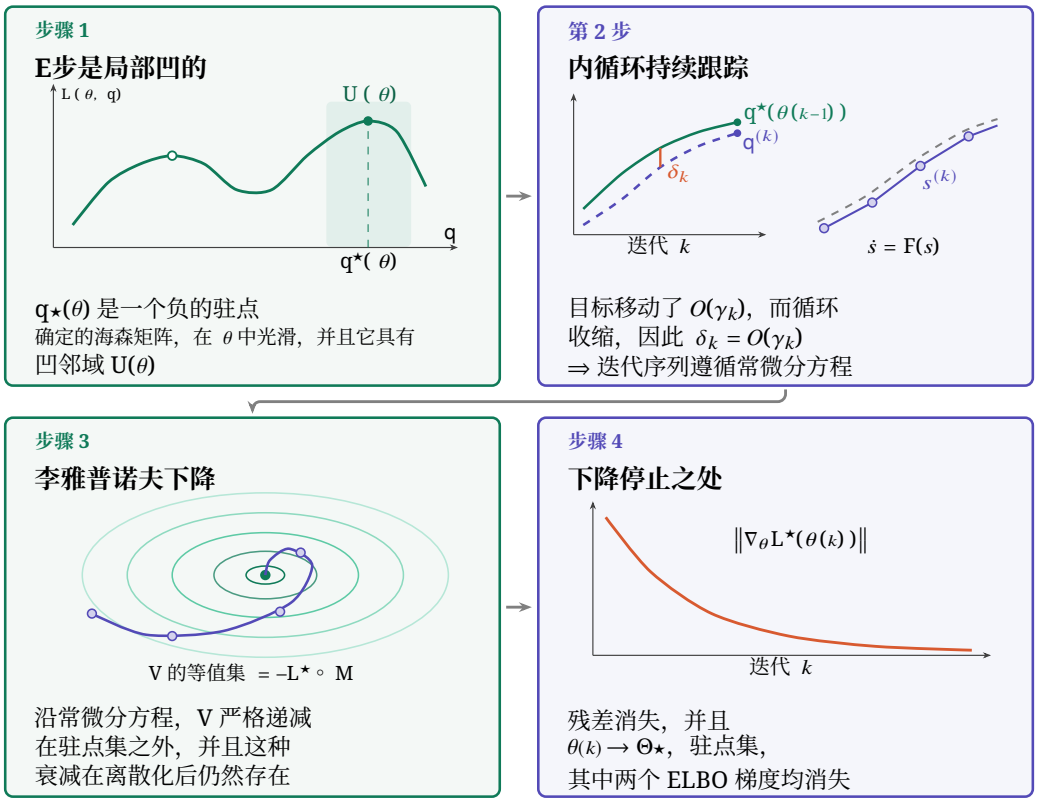

图 1：收敛性证明的四个步骤。

定理3可通过相同步骤得到，只需将内循环误差替换为重要性采样偏差。所有证明均见附录，其中三个定理的证明见附录 A.7。

5.1 预备知识：统计量空间与 M 步算子

预备知识

思想。 该算法仅存储经过平均的充分统计量 $s^{(k)}$，总体参数则通过 M步由这些统计量得到， $\theta^{(k)}$=M ($s^{(k)}$)。因此，算法是在统计量空间中进行研究的。这里需要两个性质。

1. 统计量始终处于有界集 S 中。
2. M步会对统计量的变化作出平滑响应。

用途：一个正则模型和一个具有局部极大值点的 M 步（假设1和假设2）。提供：无需额外假设即可保证迭代序列有界，因为每次更新都是凸组合；此外，Lipschitz连续的M步具有显式导数。四个步骤均会用到这两者。

引理1 （统计量空间）。设假设 *1* 成立。则$\mathcal{S}$ 是紧且凸的， $s_3 \geq s_{\min} > 0$在$\mathcal{S}$上，并且

$$\Sigma^{\text{tot}}(s) \succeq N\tau_{\min} I \succ 0 \qquad \textit{for every } s \in \mathcal{S}.$$

此外，若对所有 k 和 $s^{(0)} \in \mathcal{S}$， $\gamma_k \in [0,1]$ ，则 *ELBO VAE-NLME* 算法的迭代序列满足 $s^{(k)} \in \mathcal{S}$ ，对所有 k 几乎必然成立。

最后这条结论很重要，因为在随机逼近中，最难验证的条件就是迭代序列始终处于统计量空间的某个紧子集中。这里这一条件由构造保证，因为$\gamma_k \in [0,1]$ 使得 (3.5) 成为一个凸组合。

引理2 （M步算子的性质）。 设假设*1*和假设*2*成立，并记 $\theta_s = \mathcal{M}(s)$。那么

(i) 对每个 $s \in \mathcal{S}$，矩阵 $H_s := -\nabla^2_\theta J(\theta_s, s)$均为正定，且 $\sup_{s\in\mathcal{S}} \|H_s\| < \infty$ 和 $\sup_{s\in\mathcal{S}} \|H_s^{-1}\| < \infty$,

(ii) $\Theta_{\mathcal{S}} = \mathcal{M}(\mathcal{S})$ 是 Θ 的紧子集，且$\mathcal{M}$ 在$\mathcal{S}$上是 *Lipschitz*连续的，

(iii) *M*步通过以下方式响应统计量的变化

$$\nabla_s \mathcal{M}(s) = H_s^{-1} D\zeta(\theta_s)^\top, \tag{5.1}$$

其中， $D\zeta$ 是自然参数映射 ζ 的雅可比矩阵，来自(2.12)。

5.2 步骤 1：变分 E 步

步骤 1

思路。 对于固定的总体参数 θ，编码器执行 E步，即最大化
关于其输出q的证据下界。该问题存在一个局部解q⋆(θ)，它随
θ平滑变化。因此，具有最佳编码器输出的证据下界 L⋆(θ) =L(θ，q⋆(θ)) 是一个平滑函数，它仅取决于 θ，且其梯度具有与 M步目标函数相同的简单形式。
适用条件：正则模型和局部凹E步（假设 1-3）。
提供。 内循环（第 2 步）的目标q⋆(θ)，以及算法所增加（步骤 3 和 4），以及包络恒等式。

对于变分参数，E步按个体拆分，L (θ，q) = $\sum_{i=1}^{N}$ L$_i$(θ，q$_i$) wit h

$$\mathcal{L}_i(\theta, \mathsf{q}_i) = \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\big[\log p(x_i, \mu_i + L_i\varepsilon; \theta)\big] + \log\big|\det L_i\big| + \tfrac{n_z}{2}\log(2\pi e), \tag{5.2}$$

且重参数化 (2.6) 关于 q$_i$ = (μ_i， L_i) 是仿射的。因此，假设3是关于模型的条件。

引理 3（关于 $\mathcal{U}(\theta)$ 的光滑性与最大化元）。 设假设 *1-3* 成立。则 $\mathsf{q} \mapsto \mathcal{L}(\theta、\mathsf{q})$ 在 $\mathcal{K}_{\max}$ 上是光滑的，即

$$\left\|\nabla_{\mathsf{q}}\mathcal{L}(\theta,\mathsf{q}_1) - \nabla_{\mathsf{q}}\mathcal{L}(\theta,\mathsf{q}_2)\right\| \;\le\; \mathcal{K}_{\max}\left\|\mathsf{q}_1 - \mathsf{q}_2\right\| \qquad \textit{for all } \mathsf{q}_1,\mathsf{q}_2 \in \mathcal{Q},$$

其中 $\mathcal{K}_{\max} < \infty$ 仅取决于假设 *1* 的界和紧集 $\Theta_{\mathcal{S}}$，并且对于每个 $\theta \in \Theta_{\mathcal{S}}$，点 $\mathsf{q}^\star(\theta)$都是 $\mathcal{L}(\theta,\cdot)$在 $\mathcal{U}(\theta)$上的唯一最大化元。

引理 4（变分包络恒等式）。 设假设 *1-3* 成立。则 $\mathsf{q}^\star$ 和 $\bar{S} = S \circ \mathsf{q}^\star$ 在 $\Theta_{\mathcal{S}}$上是利普希茨的，$\bar{S}$ 取值于 $\mathcal{S}$，并且 $\mathcal{L}^\star$ 连续可微，具有 *Lipschitz* 梯度，且

$$\nabla_\theta \mathcal{L}^\star(\theta) \;=\; \nabla_\theta \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big) \;=\; \nabla_\theta J\big(\theta, \bar{S}(\theta)\big) \;=\; D\zeta(\theta)^\top \bar{S}(\theta) - \nabla_\theta \kappa(\theta). \tag{5.3}$$

5.3 第 2 步：Robbins-Monro 形式、内循环误差与极限常微分方程

第 2 步

思路。 一次迭代仅以大小为 γ_k 的小步长改变充分统计量。因此，迭代序列的行为类似于常微分方程 $\dot{s} = \mathrm{F}(s)$ 的解，但会受到蒙特卡罗噪声和内循环误差的扰动。这两种扰动的平均值都会趋于零，因此，当 k 较大时，迭代序列会遵循该常微分方程。

用途。 目标$\mathsf{q}^\star(\theta)$以及第 1 步中的统计量 $\bar{S}$、预备知识中的Lipschitz连续的M步、步长（假设 4）和具有收缩性的内循环（假设5）。
提供。 平均场 F 以及迭代序列遵循其常微分方程这一结论。因此，只需研究第 3 步中的常微分方程。

5.3.1 Robbins-Monro 分解

令 $\theta(k-1)=\mathrm{M}(s^{(k-1)})$，并定义平均场

$$\mathcal{F} : \mathcal{S} \to \mathbb{R}^{n_S}, \qquad \mathcal{F}(s) \;=\; \bar{S}\big(\mathcal{M}(s)\big) - s.$$

By 根据引理 2 和引理 4，平均场在 S 上是利普希茨连续且有界的。更新（3.5）采用确切的 t Robbins-Monro 形式

$$s^{(k)} \;=\; s^{(k-1)} + \gamma_k\Big(\mathcal{F}\big(s^{(k-1)}\big) + \Delta_k^{\mathrm{mc}} + \Delta_k^{\mathrm{in}}\Big), \tag{5.4}$$

with

$$\Delta_k^{\mathrm{mc}} := S\big(x, \mathsf{q}^{(k)}; \varepsilon^{(k)}\big) - S\big(\mathsf{q}^{(k)}\big), \qquad \Delta_k^{\mathrm{in}} := S\big(\mathsf{q}^{(k)}\big) - \bar{S}\big(\theta^{(k-1)}\big),$$

其中， Δ_k^{mc}是蒙特卡罗噪声， Δ_k^{in}是内循环误差。下一个引理汇总了收敛性证明所需的这两项的性质。

引理 5（噪声条件）。 设假设 *1-4* 成立，并令 $\mathcal{G}_k$ 为由截至第 k 次内循环（含该次）为止的所有随机变量生成的 σ-代数。则 $\mathbb{E}[\Delta_k^{\mathrm{mc}} \mid \mathcal{G}_k] = 0$，蒙特卡罗项有界于 $2C_S$，其中 $C_S = \sup_{\mathsf{q}\in\mathcal{Q},\ \varepsilon} \|S（x，\mathsf{q}；\varepsilon）\| < \infty$，级数 $\sum_k \gamma_k \Delta_k^{\mathrm{mc}}$几乎必然收敛，并且

$$\left\|\Delta_k^{\mathrm{in}}\right\| \;\le\; L_S\,\delta_k \tag{5.5}$$

对于 Q 上 $\mathsf{q} \mapsto S（\mathsf{q}）$ 的 *Lipschitz* 常数 L_S。

5.3.2 内循环误差趋于零

命题 1(消失的内循环误差**)**。 设假设 *1-5* 成立。那么，几乎必然地，

$$\delta_k \;=\; O(\gamma_k) \xrightarrow[k\to\infty]{} 0 \qquad \textit{and} \qquad \sum_{k\ge 1} \gamma_k\,\delta_k \;<\; \infty.$$

5.3.3 极限常微分方程

定义插值时间

$$t_0 = 0, \quad t_k = \sum_{j=1}^{k} \gamma_j,$$

以及插值过程

$$\bar{s}(t_k) = s^{(k)}, \quad \bar{s}(t) = s^{(k)} + \frac{t - t_k}{t_{k+1} - t_k}\left(s^{(k+1)} - s^{(k)}\right) \text{ for } t \in [t_k, t_{k+1}].$$

并令 $\Phi^{\mathcal{F}}$ 表示该自治常微分方程的半流

$$\dot{s}(t) \;=\; \mathcal{F}(s(t)) \;=\; \bar{S}\big(\mathcal{M}(s(t))\big) - s(t). \tag{5.6}$$

由于 F 在 S 上是利普希茨的，(5.6) 对每个初始值 [27] 都有唯一解。由于 S 是凸集，且 $\bar{S}(\Theta_{\mathcal{S}}) \subseteq \mathrm{S}$，该向量场在 ∂S 上指向 S 内部，因此 S 是正向不变的，并且对于所有$v \geq 0$以及所有 $s \in \mathrm{S}$， $\Phi_v^{\mathcal{F}}(s)$ 都有定义。

命题 2 （插值过程是渐近伪轨道）。 设假设 *1-5* 成立。则，几乎必然地，对于每个 $T > 0$

$$\lim_{t\to\infty} \sup_{0\le v\le T} \left\|\bar{s}(t+v) - \Phi_v^{\mathcal{F}}\left(\bar{s}(t)\right)\right\| \;=\; 0, \tag{5.7}$$

并且 $\{s^{(k)}\}$ 的极限点集合是 S 的一个非空、紧、连通且不变的子集，该不变性是指相对于 (5.6) 的流而言。

5.4 第3步：李雅普诺夫稳定性

第 3 步

思路。 沿着第 2 步的常微分方程，证据下界 L⋆ 不断增加，即 V = –L⋆ ◦ M 不断减小。
除非 L⋆ 的梯度消失，否则这种减小是严格的。由于迭代序列遵循该常微分方程，
算法也表现出相同的下降趋势，只是存在消失扰动。
用途。 第 1 步中的 L⋆ 梯度、预备知识中 M步的导数，以及
带有来自第 2 步的扰动的常微分方程。
提供。 L⋆ 的梯度沿迭代序列趋于零。

李雅普诺夫函数定义为

$$\mathcal{V} : \mathcal{S} \to \mathbb{R}, \qquad \mathcal{V}(s) \;=\; -\mathcal{L}^{\star}\big(\mathcal{M}(s)\big), \qquad \theta_s = \mathcal{M}(s),$$

根据引理 2 和引理 4，该函数在 S 上有界且连续可微，并具有 Lipschitz 梯度（附录 A.5）。

引理 6 （李雅普诺夫衰减）。 设假设 *1-3* 成立。那么，对于每个 $s \in \mathcal{S}$，

$$\left\langle \nabla_s \mathcal{V}(s), \mathcal{F}(s) \right\rangle \;=\; -\left\|\nabla_\theta \mathcal{L}^{\star}(\theta_s)\right\|^2_{H_s^{-1}} \;\le\; 0, \tag{5.8}$$

并且存在常数 c_{dec}、 $c_{\mathrm{grad}} \in$ （0、 ∞），使得

$$\left\langle \nabla_s \mathcal{V}(s), \mathcal{F}(s) \right\rangle \;\le\; -c_{\mathrm{dec}} \left\|\nabla_\theta \mathcal{L}^{\star}(\theta_s)\right\|^2, \qquad \left\|\nabla_s \mathcal{V}(s)\right\| \;\le\; c_{\mathrm{grad}} \left\|\nabla_\theta \mathcal{L}^{\star}(\theta_s)\right\|. \tag{5.9}$$

特别地，在驻点集之外，$\mathcal{V}$ 沿着(5.6) 严格递减

$$\mathcal{S}^{\star} \;:=\; \left\{s \in \mathcal{S} :\; \nabla_\theta \mathcal{L}^{\star}(\mathcal{M}(s)) = 0\right\}.$$

命题 3 (离散李雅普诺夫衰减)。 设假设 *1-5*成立。那么，几乎必然地，$\mathcal{V}(s^{(k)})$收敛到一个有限极限，并且

$$\sum_{k\geq 1}\gamma_k \left\|\nabla_\theta \mathcal{L}^\star\left(\theta^{(k-1)}\right)\right\|^2 \;<\; \infty. \tag{5.10}$$

命题 4 (消失的平稳性残差)。 在命题 *3* 的假设下，以下结论几乎必然成立

$$\lim_{k\to\infty}\left\|\nabla_\theta \mathcal{L}^\star\left(\theta^{(k)}\right)\right\| \;=\; 0.$$

5.5 步骤 4：平稳最优性条件

步骤 3 是关于 $\mathrm{L}^\star$ 的陈述，也就是说，仅涉及 θ。最后一步将其转换回证据下界，它取决于 θ 和编码器输出 q。

引理7 (极限集刻画)。 设假设 *1-3* 成立，且令 $\theta^\star \in \Theta_S$ 满足$\nabla_\theta \mathcal{L}^\star(\theta^\star) = 0$。令 $\mathsf{q}^\star = \mathsf{q}^\star(\theta^\star)$，并令 $s^\star = \bar{S}(\theta^\star)$。则

$$\nabla_\theta \mathcal{L}(\theta^\star, \mathsf{q}^\star) = 0, \qquad \nabla_{\mathsf{q}} \mathcal{L}(\theta^\star, \mathsf{q}^\star) = 0, \qquad \nabla_\theta J\left(\theta^\star, s^\star\right) = 0. \tag{5.11}$$

前两个恒等式表明，$(\theta^\star,\mathsf{q}^\star)$ 是证据下界 (2.8) 的一个驻点；第三个恒等式表明，$\theta^\star$ 在 $s^\star$ 处满足 M步条件，因此算法不会离开 $(s^\star, \theta^\star)$。反过来，每一对满足这三个性质的参数，且其中的 $\mathsf{q}^\star = \mathsf{q}^\star(\theta^\star)$ 来自假设*3*，都满足 $\nabla_\theta \mathcal{L}^\star(\theta^\star) = 0$。

命题 4 和引理7给出定理1。对于定理 2，内循环误差不会消失。根据（5.5），它以大小为 $L_S\delta_k$ 的扰动形式影响步骤 3 的衰减，使下界（4.7）带有 $C = c_{\mathsf{grad}} L_S / c_{\mathsf{dec}}$。

5.6 扩展到重要性加权统计量

定理3的证明沿用相同的四个步骤，只是以精确后验分布取代变分分布。这就是 SAEM [18, 定理 5], 的收敛性证明，只不过噪声有偏且无界。

The 由于费舍尔恒等式 [18, 19], 加权统计量消除了（4.10）的右端项

$$\nabla_\theta \mathcal{L}\mathcal{L}(\theta) \;=\; D\zeta(\theta)^\top \bar{S}_p(\theta) - \nabla_\theta \kappa(\theta) \;=\; \nabla_\theta J\left(\theta, \bar{S}_p(\theta)\right). \tag{5.12}$$

这是包络恒等式（5.3），其中以 LL 和 $\bar{S}_p$ 代替 $\mathrm{L}^\star$ 和 $\bar{S}$。因此，若统计量由 S_p 驱动，则不动点 $s = S_p(\mathrm{M}(s))$ 满足

$$\nabla_\theta \mathcal{L}\mathcal{L}(\mathcal{M}(s)) = \nabla_\theta J(\mathcal{M}(s), s) = 0$$

根据 (4.2)，也就是说，该算法的不动点是对数似然的驻点。

用 S_p替代 S后，步骤 3 的平均场和李雅普诺夫函数分别为 $\mathrm{F}_P(s) = S_p(\mathrm{M}(s)) - s$和 $\mathrm{V}_P(s) = -\mathrm{LL}(\mathrm{M}(s))$。更新仍采用 Robbins-Monro 形式 (5.4),

$$s^{(k)} = s^{(k-1)} + \gamma_k\left(\mathcal{F}_p\left(s^{(k-1)}\right) + \Delta_k^{\mathrm{mc}} + \Delta_k^{\mathrm{is}}\right), \qquad \Delta_k^{\mathrm{is}} := \mathbb{E}\left[\widehat{S}_K \,\middle|\, \mathcal{G}_k\right] - \bar{S}_p\left(\theta^{(k-1)}\right),$$

其中蒙特卡罗噪声为 $\Delta_k^{\mathrm{mc}} = \widehat{S}_K - \mathbb{E}[\widehat{S}_K|\mathsf{G}_k]$。重要性采样偏差 Δ_k^{is}取代了内循环误差。其大小取决于权重的矩，尤其是二阶矩[1, 15],

$$\rho(\theta, \mathsf{q}) \;=\; \max_{i\leq N} \mathbb{E}_{z\sim r_i}\left[\left(\frac{p(z|x_i;\theta)}{r_i(z)}\right)^2\right] \;=\; 1 + \max_{i\leq N} \chi^2\left(p(\cdot|x_i;\theta)\,\middle\|\, r_i\right),$$

与卡方散度 $\chi^2(p\|r) = \int \left(p(z) - r(z)\right)^2 / r(z)\; dz$。因此， $\rho \geq 1$衡量编码器对后验分布的近似程度；如果二者一致，则 $\rho = 1$ 。下面的引理控制偏差。

引理 8（加权统计量的一致性）。 设假设 *1(i)*成立，设 $\theta \in \Theta$、$\mathsf{q} \in Q$，并设 $\widehat{S}_K$ 由带有 $\nu > 4$ 的（3.9）给出。则 $S_K \to S_p(\theta)$ 几乎必然地成立，当 $K \to \infty$ 时；并且对于每个有限的 K，

$$\left\|\mathbb{E}[\widehat{S}_K] - \bar{S}_p(\theta)\right\| \;\le\; \frac{c_{\mathrm{is}}(\theta,\mathsf{q})}{K}, \qquad \mathbb{E}\Big[\left\|\widehat{S}_K - \bar{S}_p(\theta)\right\|^2\Big] \;\le\; \frac{c_{\mathrm{is}}(\theta,\mathsf{q})}{K}, \tag{5.13}$$

其中 $c_{\mathrm{is}}(\theta,\mathsf{q}) < \infty$ 与 K相互独立，在 $\Theta \times Q$ 的紧子集上有界，并且仅通过权重的矩依赖于编码器。对于 S_K 的有界第三个分块，它与 $\rho(\theta,\mathsf{q})$ 成正比。

因此，

$$\|\Delta_k^{\mathrm{is}}\| \le \frac{\bar{c}_{\mathrm{is}}}{K} \quad \text{with} \quad \bar{c}_{\mathrm{is}} = \sup\{c_{\mathrm{is}}(\theta,\mathsf{q}) : \theta \in \Theta_{S_p},\, \mathsf{q} \in Q\} < \infty.$$

因此，将 $L_S\delta$ 替换为 $\bar{c}_{\mathrm{is}}/K$ 后，定理 2 的证明仍然适用；也就是说，使用 $C = c_{\mathsf{grad}}\bar{c}_{\mathrm{is}}/c_{\mathsf{dec}}$ 以及 (5.9) 中 V_P 在 S_P 上对应的常数。对于满足 $\sum_k \gamma_k/K_k < \infty$ 的 $K_k \to \infty$，偏差消失，且 $\sum_k \gamma_k\|\Delta_k^{\mathrm{is}}\| < \infty$；这些正是步骤 2–4（命题 1）中用到的内循环误差性质。

假设 6 替代引理1，因为 S_K 不一定属于 S。此外，S_K 无界。因此，引理 5 中有界的蒙特卡罗噪声被条件二阶矩有界的噪声所取代。

6 数值实验

This 本节展示 VAE-NLME（算法 1）和证据下界
VAE-NLME 算法的理论结果如何体现在数值实验中。
我们在一项独立的基准研究中，将 VAE-NLME 算法与 SAEM 和 FOCEI 进行比较，开展了全面的实证评估，共拟合了 16 个药代动力学和药效动力学模型 1,224 次 [48]。该基准测试评估了算法的准确性、稳健性和计算效率。因此，本节仅聚焦于使用一个更简单的示例——真实的茶碱数据集 [42], ——来验证理论结果；该数据集是非线性混合效应建模中的标准基准。其低维特性使得 $\mathsf{q}^{(\star)}(\theta)$ 和 $\nabla_\theta \mathrm{L}^\star$ 等关键量能够以较高精度计算。

首先介绍模型和总体参数的收敛。其次，在理想化设置下验证定理，然后使用应用中采用的实际实现进行验证。最后，对比内循环误差与第 4 节中引入的差距。

备注1中所述的实际实现版本 VAE-NLME 算法，可在基准测试软件 VAE Studio 中使用。VAE Studio 是一款适用于 Windows、macOS 和 Linux 的桌面应用程序，提供图形用户界面，可通过高度优化的实现来试用该算法。截至2026年11月，VAE Studio 仍处于测试阶段，可通过 https://www.math.uni-konstanz.de/vae-studio 申请免费的限时试用许可证。

6.1 模型

茶碱数据包含 $N = 12$ 个体，每个个体在接受单次口服剂量后有 $n_i = 10$ 个浓度测量值。结构模型为带有一级吸收的一室模型。对于剂量 D，模型可写为

$$\begin{aligned} \frac{dAbs(t)}{dt} &= -k_a\, Abs(t), & Abs(0) &= D, \\ \frac{dA(t)}{dt} &= k_a\, Abs(t) - k_e\, A(t), & A(0) &= 0, \end{aligned}$$

以个体参数 $\phi = (k_a、\ k_e、\ V)$ 和浓度 $f(t、\ \phi) = A(t)/V$ 作为式（2.1）中的结构模型。解码器使用解析解

$$f(t,\phi) = \frac{D\,k_a}{V(k_a - k_e)}\left(e^{-k_e t} - e^{-k_a t}\right). \tag{6.1}$$

对于个体 i，剂量为 D_i，个体参数为 $\phi_i = (k_{a,i}, k_{e,i}, V_i)$。它们均为正数，即 $\Phi = \mathbb{R}^3_{>0}$。因此，h = 取对数，并且每个参数均服从对数正态分布，其中体重作为体积的协变量，

$$\begin{aligned}
\log(k_{a,i}) &= \log(k_{a,\text{pop}}) + \eta_{a,i}, & \eta_{a,i} &\sim \mathcal{N}(0, \omega^2_{k_a}),\\
\log(k_{e,i}) &= \log(k_{e,\text{pop}}) + \eta_{e,i}, & \eta_{e,i} &\sim \mathcal{N}(0, \omega^2_{k_e}),\\
\log(V_i) &= \log(V_{\text{pop}}) + \beta_V\, c_i + \eta_{V,i}, & \eta_{V,i} &\sim \mathcal{N}(0, \omega^2_V), \qquad c_i = \log\Big(\frac{W_i}{W_{\text{pop}}}\Big),
\end{aligned}$$

其中 W_i 是个体 i 的体重，W_{pop} 是平均体重。按照 (2.2) 中的记号，$z_i = C_i \hat{z}_{\text{pop}} + \eta_i$，其中

$$C_i = \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & c_i \end{pmatrix}, \qquad \Omega = \operatorname{diag}\big(\omega^2_{k_a}, \omega^2_{k_e}, \omega^2_V\big),$$

$$\hat{z}_{\text{pop}} = \big(\log(k_{a,\text{pop}}), \log(k_{e,\text{pop}}), \log(V_{\text{pop}}), \beta_V\big).$$

连同残差误差参数 σ，共估计八个总体参数 $\theta = (\hat{z}_{\text{pop}}, \Omega, \sigma^2)$。参数 $k_{a,\text{pop}}$, $k_{e,\text{pop}}$ 和 V_{pop} 以原始尺度报告，即对数正态先验的中位数；β_V 和 ω 则以 z的尺度报告。

6.2 总体参数的收敛性

本节的所有运行均遵循备注1，采用自适应选择的 K_{KL} 和 K_{sm}，并使用表1中的设置。

L	8	每次迭代的 Adam 步数
α	$5 \cdot 10^{-3}$	Adam 的学习率
M	1	ELBO梯度的样本
K	10	加权统计量的样本
ν	5	提议分布的自由度

表 1：茶碱数据上所有运行的设置。

本小节的运行均从初始猜测开始

$$\phi^{\text{init}} = (k_a, k_e, V) = (1,\ 1,\ 15), \qquad \omega^{\text{init}} = (\omega_{k_a}, \omega_{k_e}, \omega_V) = (0.5,\ 0.5,\ 0.5), \qquad \sigma^{\text{init}} = 1.$$

协变量效应 β_V 无需初始值，并从零开始。图2比较了 VAE-NLME（算法 1）与 ELBO VAE-NLME 算法。

VAE-NLME 算法的波动更大，因为加权统计量的方差更大。平滑阶段会平均掉这些波动，而 θ 的所有分量都趋于收敛。正如式（ 4.10 ）所预期的，VAE-NLME 达到更大的对数似然。

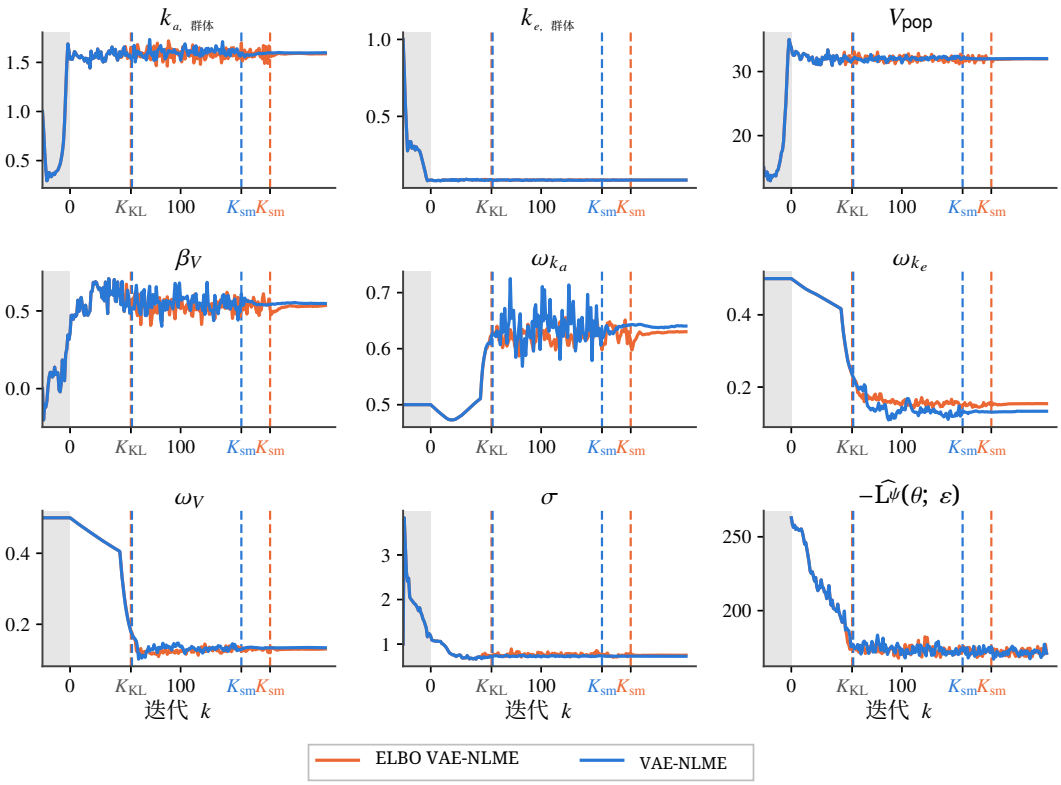

图2：VAE-NLME（算法 1，−2LL = 335.7）和 ELBO VAE-NLME 算法（−2LL = 335.9）在茶碱数据上的总体参数 $\theta^{(k)}$，以及最后一个面板中的负 ELBO 蒙特卡罗估计。VAE-NLME 的运行在 205 次迭代后停止，并继续运行至相同的迭代次数。阴影部分表示预热阶段。

6.3 收敛到平稳点

两个参考点 $\theta^{\star}_{\mathrm{ELBO}}$（对于 $\mathrm{L}^{\star}$ 是平稳点）和 θ_{ML}（对于 LL 是平稳点）根据式（4.10）相互确定对方的残差下限，

$$\left|\nabla_\theta \mathcal{LL}(\theta^{\star}_{\mathrm{ELBO}})\right| = \left|\nabla_\theta \mathcal{G}_{\mathrm{var}}(\theta^{\star}_{\mathrm{ELBO}})\right| = 4.8\cdot 10^{-2}, \tag{6.2}$$

$$\left|\nabla_\theta \mathcal{L}^{\star}(\theta_{\mathrm{ML}})\right| = \left|\nabla_\theta \mathcal{G}_{\mathrm{var}}(\theta_{\mathrm{ML}})\right| = 4.8\cdot 10^{-2}. \tag{6.3}$$

这两个点在三位数字上相符，并且 $\mathrm{G}_{\mathrm{var}}(\theta^{\star}_{\mathrm{ELBO}}) = 0.06$，即高斯变分族几乎是精确的，但其梯度并不为零。

6.3.1 理论验证

首先，考虑一种理想化情形，其中采用精确 E 步，即$\mathrm{q}^{(k)} = \mathrm{q}^{\star}(\theta_{(k-1)})$，且内循环误差为 $\delta_k = 0$。由于只有 $n_z = 3$ 个体参数，因此每次迭代都可以通过高斯-埃尔米特求积和确定性优化器（附录 A.8），以接近机器精度的精度计算最大化元$\mathrm{q}^{\star}(\theta)$和残差 $\|\nabla_\theta \mathrm{L}^{\star}(\theta_{(k)})\|$及 $\|\nabla_\theta \mathrm{LL}(\theta_{(k)})\|$。在变分充分统计量中，$S_3$ 每次迭代使用 $M = 2^{20}$ 个新样本 ε，并通过求积计算加权统计量，这对应于 $K \to \infty$。

根据定理1，沿着 ELBO VAE-NLME 算法，$\mathrm{L}^{\star}$ 的梯度趋于零；根据定理3，沿着 VAE-NLME 算法（算法 1），LL 的梯度趋于零。另一个目标函数的梯度则稳定在其下限（6.2）。

图4展示了引理 6 中的李雅普诺夫函数 $\mathrm{V}(s^{(k)})$ $\mathrm{L}^{\star}(\theta_{(k)})$ 以及第 5节中 $\mathrm{V}(s^{(k)})$ $\mathrm{LL}(\theta_{()})$。除 KL 退火阶段外，V_P 单调下降，直至达到数值精度。由于 (5.4) 中的蒙特卡罗噪声 Δ_k^{mc}，函数 V 并非单调，但其最大增幅仅为 $6.5\cdot 10^{-7}$。

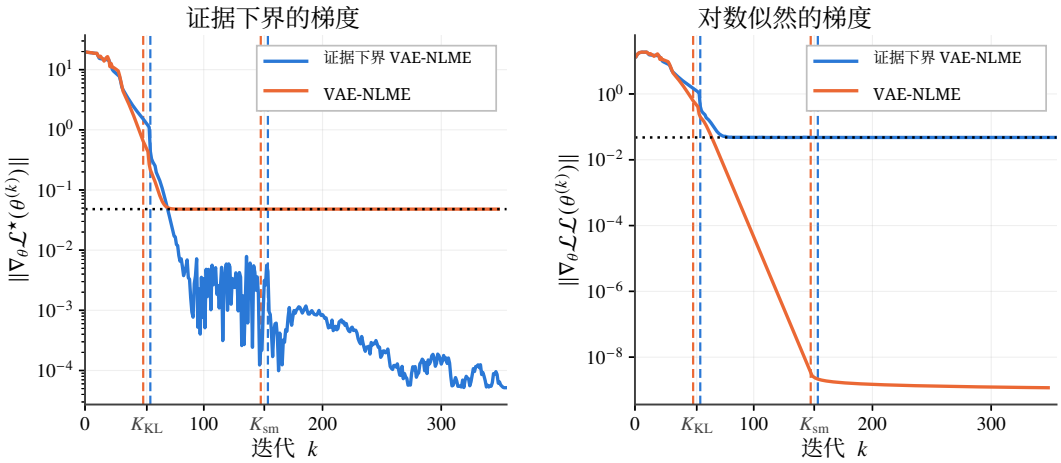

图3：精确 E 步和 $K \to \infty$ 下证据下界及对数似然的梯度。左：定理1。右：定理3。两个面板均显示两种算法的结果。点线表示下限（6.2），对应于变分间隙的梯度。

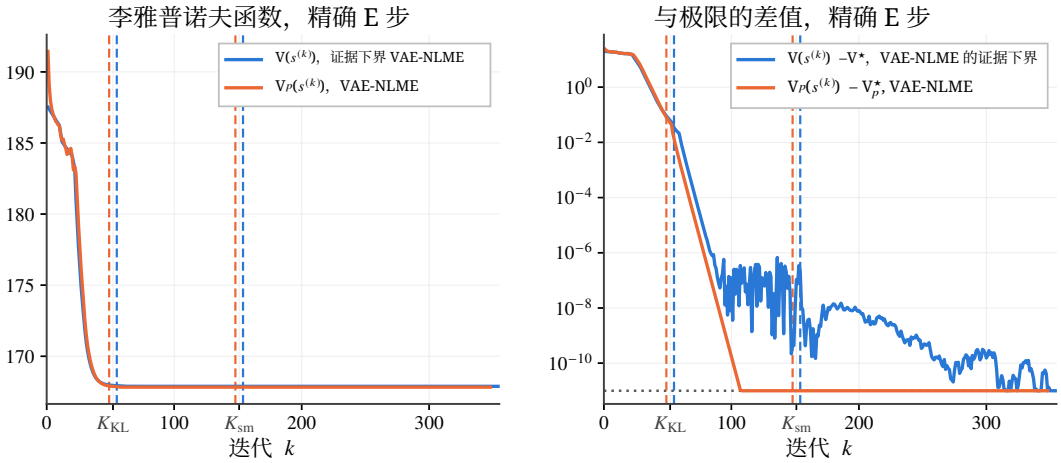

图4：采用精确 E 步时，ELBO VAE-NLME 算法和 VAE-NLME（算法 1）的李雅普诺夫函数。右图显示运行极限处与值 $V^\star$ 和 $V_p^\star$的差值。低于 10^{-11} 数值精度的差值以点线表示。

6.3.2 实际验证

图5展示了带编码器的两种算法，即带有 $\delta_k > 0$ 和有限 K 的算法。对于 ELBO VAE-NLME 算法，$L^\star$ 的梯度停滞了。这就是定理 2 的下限（4.7）。对于 VAE-NLME（算法 1），平滑阶段延长至 2000 次迭代，因为剩余噪声的平均化过程较为缓慢。剩余的 LL 梯度随 K 减小，这与定理 3 的下限（4.11）一致，尽管编码器未达到 $q^\star(\theta)$。

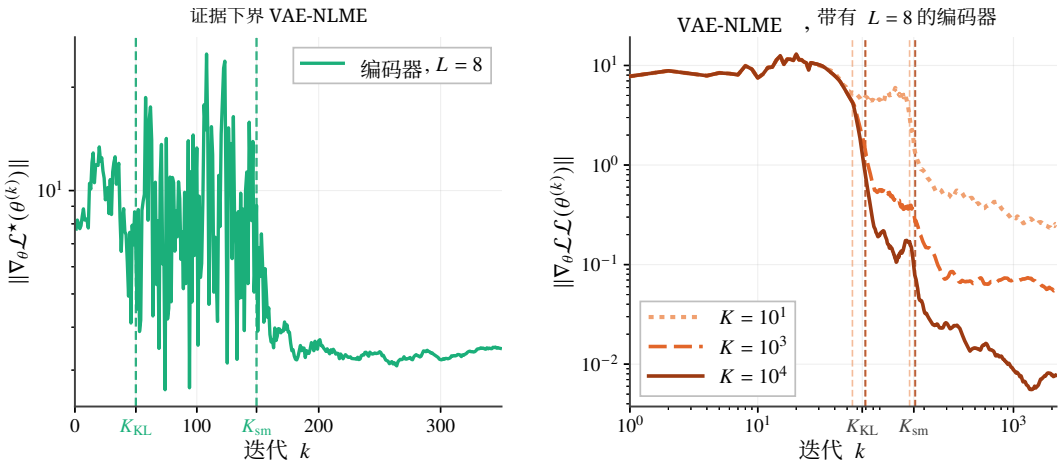

图5：采用编码器的两种算法。左：ELBO VAE-NLME 算法中 L⋆ 的梯度，定理 2。右：采用有限 K的 VAE-NLME（算法 1）中对数似然的梯度，定理 3 (i)。

6.4 编码器与内循环误差

图 6考察了图5中 ELBO VAE-NLME 算法的模拟步骤。它将编码器与对每个个体的变分参数 $\mathsf{q}_i = (\mu_i, L_i)$ 进行自由优化的情况进行比较；后者如同 [52],的非摊销变分 EM 算法一样，在每次迭代中执行相同数量 L 的 Adam 步骤。

备注1中的实际算法将 Adam 的步长 α 固定。固定步长会在 $\mathsf{q}^\star(\theta)$ 周围留下半径为 $O(\alpha)$ 的噪声球，因此在实践中，模拟步骤永远无法到达 $\mathsf{q}^\star(\theta)$。因此，即使内循环很长，δ_k 也不会消失。这解释了为什么图5中 L⋆ 的梯度会停滞在定理 2 的下界（4.7）处。

编码器比自由优化收敛得更快。由于共享权重的每一步都会同时更新所有个体，因此它在大约 70 次迭代后便达到平台期，而自由优化则需要约 160 次。相应地，它的平台期更高。借助共享权重，编码器同时包含第 4.4.2 节所述的摊销差距和优化差距，而自由参数只包含优化差距。根据（4.13），编码器更大的 δ_k 也意味着更大的差距 $\mathrm{L}^\star - \mathrm{L} = \mathrm{G}_{\mathrm{am}} + \mathrm{G}^{\mathrm{opt}}$。

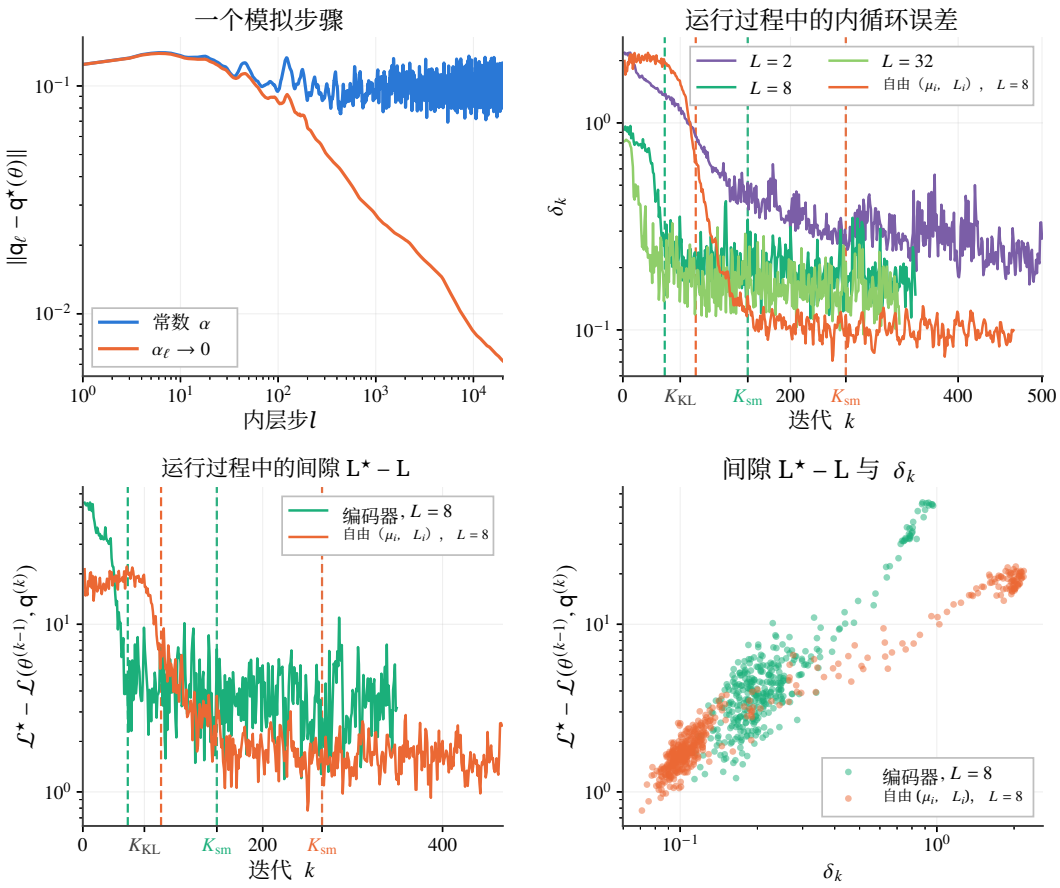

图 6：使用编码器和自由 $(\mu_i,\ L_i)$ 时，证据下界 VAE-NLME 算法的内循环误差；定理 2。(a) 在固定 θ 下，对自由 $(\mu_i,\ L_i)$ 进行一次模拟步骤，从距离 $\mathsf{q}^\star(\theta)$ 0.13 处开始，并使用新样本。(b) 运行过程中的 δ_k。(c) 间隙 $\mathrm{L}^\star - \mathrm{L}$。(d) 间隙与 δ_k 的关系。

6.5 证据下界、似然与间隙

表 2 总结了第 4.3 节和第 4.4.2 节中的差距，这些差距是在图 6 中 ELBO VAE-NLME 算法的两次运行以及 VAE-NLME（算法 1）的一次运行达到极限时测得的。在这三次运行中，变分间隙 $\mathrm{G}_{\mathrm{var}} = \mathrm{LL} - \mathrm{L}^\star$ 均为 0.06。VAE-NLME 的差距 $\mathrm{L}^\star - \mathrm{L}$ 与 ELBO VAE-NLME 算法处于同一数量级，但其 LL 的梯度小一个数量级。

	$\mathcal{L}$	$\mathcal{L}^\star$	LL	$\mathrm{L}^\star - \mathrm{L}$	$\mathcal{G}_{\mathrm{var}}$	$\|\nabla \mathrm{L}^\star\|$	$\|\nabla \mathrm{LL}\|$
证据下界 VAE-NLME，编码器	−171.19	−168。02	−167。97	3.16	0.06	3.5	3.5
证据下界 VAE-NLME，自由 q	−170。42	−167。93	−167。88	2.49	0.06	2.1	2.1
VAE-NLME，编码器	−169。94	−167。89	−167。83	2.05	0.06	$2。4 \cdot 10^{-1}$	$2。5 \cdot 10^{-1}$
$\theta^\star_{\mathrm{ELBO}}$	.	−167。89	−167.83	·	0.06	$4.9 \cdot 10^{-6}$	$4.8 \cdot 10^{-2}$
θ_{ML}	.	−167。89	−167.83	·	0.06	$4。8 \cdot 10^{-2}$	$5.9 \cdot 10^{-7}$

表 2：三次运行收敛时及两个参考点处的证据下界、边际对数似然及相应差值。$\mathrm{L} = \mathrm{L}(\theta, \mathsf{q})$ 是最后一次迭代时变分参数 q 所达到的证据下界，$\mathrm{L}^\star = \mathrm{L}(\theta, \mathsf{q}^\star(\theta))$ 是变分最大化解处的证据下界，所有梯度均相对于 θ 计算。

7 结论

VAE-NLME 算法收敛于最大似然估计量，尽管变分自编码器只最大化证据下界。更准确地说，借助其重要性加权充分统计量，该算法几乎必然收敛到边际对数似然的驻点，并在局部收敛于最大似然估计量；对于 K 个样本，其误差不超过 $O(1/K)$（定理3），且无需对编码器训练作任何假设。编码器通过随机梯度在证据下界上训练，并作为提议分布。因此，证据下界兼具基于梯度训练的速度，而重要性加权则使极限从证据下界的最大值转移至最大似然估计量。编码器对后验分布的近似越准确，所需样本就越少 [15]。额外开销为每个个体、每次迭代需要进行 K 次解码器评估。

即使不假设其神经网络编码器的训练会收敛，ELBO VAE-NLME 算法本身仍然收敛。原因在于网络嵌入 VAE 和算法的方式。分析使用编码器的输出而非其权重，并且 M步只使用充分统计量。因此，训练不精确只会扰动统计量。对于收缩内循环，这种扰动会消失（定理1）；否则，会留下一个至多与内循环误差成正比的残余梯度（定理 2），且与网络架构无关。

数值实验验证了全部三个定理，另一个独立的基准测试则证实了该算法在实践中的准确性、稳健性和速度 [48]。在 PMX 中，目前已知变分自编码器的解不够准确。先前工作 [46, 47] 中的 ELBO VAE-NLME 算法速度很快，但只能收敛到证据下界的驻点（定理1）。重要性加权弥合了其与最大似然估计量之间的差距。由于它只替换统计量，先前工作的扩展也同样适用。具体而言， [47] 中的协变量选择过程完全沿用原来的方式，只是现在使用最大似然估计量，而不是证据下界的最大值。由于该算法通过梯度进行训练，因此也可以纳入神经网络组件；鉴于其自由度众多，若没有梯度，这些组件几乎无法训练 [49]。这使 VAE-NLME 算法成为 PMX 下一代 NLME模型的一种有前景的方法。除了 PMX 之外，定理3的证明仅用到了完整数据似然的指数族形式。因此，对于具有这种形式的一般边缘化似然问题，该算法最大化的也是对数似然，而不是证据下界。

致谢

作者感谢 Claude（Anthropic）在准备本文过程中提供的协助，尤其是在讨论想法、编写证明概要以及准备数值实验图表方面。ChatGPT（OpenAI）和 Claude 还被用于提升文本的可读性和清晰度。作者对本文内容承担全部责任。

参考文献

[1] S. Agapiou、O. Papaspiliopoulos、D. Sanz-Alonso 和 A. M. Stuart。“重要性抽样：内在维数与计算成本”。载于：《统计科学》32.3（2017），第405–431页。doi：10.1214/17-STS611。[2] J. Arruda、Y. Schälte、C. Peiter、O. Teplytska、U. Jaehde 和 J. Hasenauer。“基于神经后验估计的非线性混合效应建模的摊销方法”。载于：第*41*届国际机器学习会议论文集。第235卷。PMLR，2024年，第1865–1901页。[3] M. Baaz、A. Sjöberg 和 M. Jirstrand。“经验贝叶斯变分自编码器——一种用于药理学群体建模的神经常微分方程方法”。载于：*CPT*：群体药代动力学与系统药理学15.7（2026年）。doi：10.1002/psp4.70280。[4] V. A. Bhattaram、B. P. Booth、R. P. Ramchandani、B. N. Beasley、Y. Wang、V. Tandon、J. Z. Duan、R. K. Baweja、P. J. Marroum、R. S. Uppoor、N. A. Rahman、C. G. Sahajwalla、J. R. Powell、M. U. Mehta 和 J. V. S. Gobburu。“定量药理学对药物审批和标签决策的影响：对42项新药申请的调查”。载于：《*AAPS*期刊》7.3（2005年），E503–E512页。doi：10.1208/aapsj070351。[5] C. M. Bishop。《模式识别与机器学习》。信息科学与统计学。纽约：Springer，2006年。[6] D. M. Blei、A. Kucukelbir 和 J. D. McAuliffe。“变分推断：面向统计学家的综述”。载于：《美国统计学会杂志》112.518（2017年），第859–877页。doi：10.1080/01621459.2017.1285773。[7] V. S. Borkar。《随机逼近：动力系统视角》。第2版。第48卷。数学文本与读本系列。新加坡：Springer，2023年。doi：10.1007/978-981-99- 8277-6。[8] J. Bornschein 和 Y. Bengio。“重加权唤醒-睡眠算法”。载于：第*3*届国际学习表征会议（*ICLR*）。2015年。[9] S. R. Bowman、L. Vilnis、O. Vinyals、A. M. Dai、R. Józefowicz 和 S. Bengio。“从连续空间生成句子”。载于：第*20*届*SIGNLL*计算自然语言学习会议论文集。2016年，第10–21页。doi：10.18653/v1/K16-1002。[10] D. S. Bräm、B. Steiert、M. Pfister、B.Steffens 和 G. Koch。“在 Monolix 和 NONMEM 中实现的考虑个体间变异性的低维神经常微分方程”。载于：*CPT*：群体药代动力学与系统药理学14.1（2025年），第5–16页。doi：10.1002/psp4.13265。[11] L. D. Brown。《统计指数族基础：及其在统计决策理论中的应用》。第9卷。讲义—专著系列。加利福尼亚州海沃德：数理统计学会，1986年。[12] Y. Burda、R. Grosse 和 R. Salakhutdinov。“重要性加权自编码器”。载于：第*4*届国际学习表征会议（*ICLR*）。2016年。[13] J. Burroni、J. Domke 和 D. Sheldon。“具有确定性目标函数的黑箱变分推断：更快、更准确，且更加黑箱化”。载于：《机器学习研究杂志》25.122（2024年），第1–39页。[14] O. Cappé 和 E. Moulines。“潜在数据模型的在线期望最大化算法”。载于：《皇家统计学会会刊：*B*辑（统计方法学）》71.3（2009年），第593–613页。doi：10.1111/j.1467-9868.2009.00698.x。[15] S. Chatterjee 和 P. Diaconis。“重要性抽样所需的样本量”。载于：《应用概率年刊》28.2（2018年），第1099–1135页。doi：10.1214/17-AAP1326。[16] C. Cremer、X. Li 和 D. Duvenaud。“变分自编码器中的推断次优性”。载于：第*35*届国际机器学习会议论文集。第80卷。《机器学习研究论文集》

. PMLR, 2018, pp. 1078–1086 .

[17] M. Delattre、M. Lavielle 和 M.-A. Poursat。“混合效应模型中的BIC注记”。载于：《电子统计学杂志》8.1（2014），第456–475页。doi: 10.1214/14-EJS890.[18] B. Delyon、M. Lavielle 和 E. Moulines。“EM算法的随机逼近版本的收敛性”。载于：《统计学年鉴》27.1（1999），第94–128页。doi: 10.1214/aos/1018031103.[19] A. P. Dempster、N. M. Laird 和 D. B. Rubin。“通过EM算法从不完整数据中进行最大似然估计”。载于：《皇家统计学会会刊：*B*辑（方法论）》39.1（1977），第1–22页。doi: 10.1111/j.2517-6161.1977.tb01600.x.[20] J. Domke。“黑箱变分推断的可证明光滑性保证”。载于：第*37*届国际机器学习会议论文集。第119卷。《机器学习研究论文集》。PMLR，2020，第2587–2596页。[21] J. Domke、R. M. Gower 和 G. Garrigos。“黑箱变分推断的可证明收敛保证”。载于：《神经信息处理系统进展》。第36卷。2023。[22] 欧洲药品管理局。药品生命周期中人工智能（*AI*）应用反思文件。反思文件 EMA/CHMP/CVMP/83833/2023。阿姆斯特丹：欧洲药品管理局，2024。[23] J. Geweke。“使用蒙特卡罗积分对计量经济模型进行贝叶斯推断”。载于：《计量经济学》57.6（1989），第1317–1339页。doi: 10.2307/1913710.[24] S. Gruffaz、K. Kim、A. Durmus 和 J. Gardner。“用于期望最大化的有偏MCMC随机逼近”。载于：第*27*届国际人工智能与统计学会议论文集。第238卷。《机器学习研究论文集》。PMLR，2024，第2332–2340页。[25] B. Karimi、B. Miasojedow、E. Moulines 和 H.-T. Wai。“有偏随机逼近方案的非渐近分析”。载于：第三十二届学习理论会议论文集。第99卷。《机器学习研究论文集》。PMLR，2019，第1944–1974页。[26] H. Karimi、J. Nutini 和 M. Schmidt。“Polyak-Łojasiewicz条件下梯度方法和近端梯度方法的线性收敛”。载于：《机器学习与数据库中的知识发现》。第9851卷。《计算机科学讲义》。Springer，2016，第795–811页。doi: 10.1007/978-3-319-46128-1_50.[27] H. K. Khalil。《非线性系统》。第3版。新泽西州上萨德尔河：Prentice Hall，2002。[28] K. Kim、J. Oh、K. Wu、Y. Ma 和 J. R. Gardner。“论黑箱变分推断的收敛性”。载于：《神经信息处理系统进展》。第36卷。2023。[29] D. P. Kingma 和 M. Welling。“自动编码变分贝叶斯”。载于：第*2*届国际学习表征会议（*ICLR*）。2014。[30] A. J. Kleywegt、A. Shapiro 和 T. Homem-de-Mello。“随机离散优化的样本平均近似方法”。载于：《*SIAM*优化杂志》12.2（2002），第479–502页。doi: 10.1137/S1052623499363220.[31] G. Koch、M. Pfister、I. Daunhawer、M. Wilbaux、S. Wellmann 和 J. E. Vogt。“定量药理学与机器学习携手推进临床数据分析”。载于：《临床药理学与治疗学》107.4（2020），第926–933页。doi: 10.1002/cpt.1774.[32] E. Kuhn 和 M. Lavielle。“将EM的随机逼近版本与MCMC程序相结合”。载于：《*ESAIM*：概率与统计》8（2004），第115–131页。doi: 10.1051/ps:2004007.[33] E. Kuhn 和 M. Lavielle。“非线性混合效应模型中的最大似然估计”。载于：《计算统计学与数据分析》49.4（2005），第1020–1038页。doi: 10.1016/j.csda.2004.07.002。

[34] H. J. Kushner 和 G. G. Yin。《随机逼近与递归算法及其应用》。第2版。第35卷。随机建模与应用概率。纽约：Springer，2003。doi: 10.1007/b97441。[35] Z. Li、M. Prague、R. Thiébaut 和 Q. Clairon。《基于常微分方程推断非线性混合效应模型的变分自编码器》。2026年。[36] M. J. Lindstrom 和 D. M. Bates。“用于重复测量数据的非线性混合效应模型”。载于：《生物统计学》46.3（1990），第673–687页。doi: 10.2307/2532087。[37] J. Lu、K. Deng、X. Zhang、G. Liu 和 Y. Guan。“用于药代动力学建模的神经常微分方程及其在预测新给药方案方面相较其他机器学习模型的优势”。载于：iScience 24.7，102804（2021）。doi: 10.1016/j.isci.2021.102804。[38] R. Madabushi、P. Seo、L. Zhao、M. Tegenge 和 H. Zhu。“综述：模型指导的药物开发方法在药物开发生命周期和监管决策中的作用”。载于：Pharmaceutical Research 39.8（2022），第1669–1680页。doi: 10.1007/s11095- 022-03288-w。[39] C. C. Margossian 和 D. M. Blei。“摊销变分推断：何时使用以及为何使用？”载于：《人工智能中的不确定性》第四十届会议论文集。第244卷。《机器学习研究论文集》。PMLR，2024，第2434–2449页。[40] D. R. Mould 和 R. N. Upton。“群体建模、模拟和基于模型的药物开发基本概念”。载于：CPT：群体药代动力学与系统药理学 1.9（2012），第1–14页。doi: 10.1038/psp.2012.4。[41] A. B. Owen。《蒙特卡罗理论、方法与示例》。https://artowen.su.domains/ mc/。斯坦福大学，2013。[42] J. C. Pinheiro 和 D. M. Bates。《S 和 S-PLUS 中的混合效应模型》。统计学与计算。纽约：Springer，2000。doi: 10.1007/b98882。[43] D. J. Rezende、S. Mohamed 和 D. Wierstra。“深度生成模型中的随机反向传播与近似推断”。载于：第31届国际机器学习会议论文集。第32卷。《机器学习研究论文集》。PMLR，2014，第1278–1286页。[44] S. Richardson、I. Irurzun Arana、A. Nowojewski、D. Zhou、J. Leander、W. Tang、R. Dearden 和 M. Gibbs。“一种用于自动化群体药代动力学建模的机器学习方法”。载于：Communications Medicine 5（2025）。doi: 10.1038/s43856-025-01054-8。[45] H. Robbins 和 S. Monro。“一种随机逼近方法”。载于：《数学统计年刊》22.3（1951），第400–407页。doi: 10.1214/aoms/1177729586。[46] J. Rohleff、F. Bachmann、M. Kasper、B. Steffens、G. Koch 和 J. Schropp。“利用神经常微分方程变分自编码器，在非线性混合效应模型中自动建模时变协变量”。载于：CPT：群体药代动力学与系统药理学（2026）。已投稿。[47] J. Rohleff、F. Bachmann、U. Nahum、D. Bräm、B. Steffens、M. Pfister、G. Koch 和 J. Schropp。“借助变分自编码器重新定义参数估计与协变量选择：一次运行即可”。载于：CPT：群体药代动力学与系统药理学 14.12（2025），第2232–2243页。doi: 10.1002/psp4.70129。[48] J. Rohleff、G. Koch、A. Krause 和 J. Schropp。“非线性混合效应建模中变分自编码器、SAEM 和 FOCEI 在准确性与速度方面的基准测试”。载于：CPT：群体药代动力学与系统药理学（2026）。已投稿。[49] J. Rohleff、G. Koch 和 J. Schropp。“定量药理学中非线性混合效应建模的演进：迈向基于人工智能的变分自编码器”。载于：《药代动力学与药效学杂志》53.5，37（2026）。doi: 10.1007/s10928-026-10044-9。[50] L. B. Sheiner、B. Rosenberg 和 V. V. Marathe。“利用常规临床数据估计药代动力学参数的群体特征”。载于：《药代动力学与生物药剂学杂志》5.5（1977），第445–479页。doi: 10.1007/BF01061728。

[51] C. K. Sønderby、T. Raiko、L. Maaløe、S. K. Sønderby 和 O. Winther。“阶梯式变分自编码器”。载于：《神经信息处理系统进展》第29卷。2016年，第3738–3746页。[52] M. Tarek 和 P. Afonso。《使用变分期望最大化拟合大型非线性混合效应模型》。2026年。[53] 美国食品药品监督管理局。《利用人工智能支持药品和生物制品监管决策的考量》。指导草案。案卷编号：FDA-2024-D-4689。马里兰州银泉市：美国卫生与公众服务部，2025年。[54] Y. Wang。“各种 NONMEM 估计方法的推导”。载于：《药代动力学与药效学杂志》34.5（2007），第575–593页。doi: 10.1007/s10928-007-9060- 6。

A 附录

A.1 化简为充分统计量

形式 (3.4) 在递推 (3.2) 过程中保持不变。假设在第 $k-1$ 次迭代时，辅助函数具有以下形式

$$Q^{(k-1)}(\theta) \;=\; \big\langle s^{(k-1)}, \zeta(\theta) \big\rangle - \kappa(\theta) + b^{(k-1)},$$

其中 $s^{(k-1)} \in \mathbb{R}^{n_S}$ 以及标量 $b^{(k-1)} \in \mathbb{R}$ 均独立于 θ。如果将 $Q^{(0)}$ 初始化为这种形式，则对于 $k=1$ 该结论成立。将分解式 (2.10) 代入 (3.2) 可得

$$\begin{aligned} Q^{(k)}(\theta) &= (1-\gamma_k)\, Q^{(k-1)}(\theta) + \gamma_k\, \widehat{\mathcal{L}}_{\psi^{(k)}}\big(\theta; \varepsilon^{(k)}\big) \\ &= (1-\gamma_k)\Big[\big\langle s^{(k-1)}, \zeta(\theta)\big\rangle - \kappa(\theta) + b^{(k-1)}\Big] \\ &\quad + \gamma_k\Big[\big\langle S(x, \psi^{(k)}; \varepsilon^{(k)}), \zeta(\theta)\big\rangle - \kappa(\theta) + \mathcal{H}(\psi^{(k)}; \varepsilon^{(k)}) + C\Big] \\ &= \big\langle s^{(k)}, \zeta(\theta)\big\rangle - \kappa(\theta) + b^{(k)}, \end{aligned}$$

即形式 (3.4) 得以保持，其中

$$\begin{aligned} s^{(k)} &= s^{(k-1)} + \gamma_k\Big(S\big(x, \psi^{(k)}; \varepsilon^{(k)}\big) - s^{(k-1)}\Big), \\ b^{(k)} &= b^{(k-1)} + \gamma_k\Big(\mathcal{H}\big(\psi^{(k)}; \varepsilon^{(k)}\big) + C - b^{(k-1)}\Big). \end{aligned}$$

A.2 预备知识：统计量空间与 M 步算子

引理1的证明（统计量空间）

证明。 根据假设 1(i),

$$\big|r_{ij}(z)\big| \;\le\; \frac{|x_{ij}| + \sup|f|}{g_{\min}} \;<\; \infty \qquad \text{for all } z \in \mathbb{R}^{n_z},$$

因此， S 的第三个分块关于 ε一致有界，而前两个分块在紧集 Q 上关于 q连续，且与 ε无关。因此

$$C_S \;:=\; \sup_{\mathsf{q}\in Q,\ \varepsilon} \big\|S(x, \mathsf{q}; \varepsilon)\big\| \;<\; \infty, \tag{A.1}$$

而 S（有界集的闭凸包）是紧的且为凸集。
在单次评估 $S_{1\text{时}},\ _i = \mu_i$和 $S_2 = \sum_i(\mu_i\mu_i^\top + \Sigma_i)$，因此

$$\Sigma^{\mathsf{tot}}\big(S(x, \mathsf{q}; \varepsilon)\big) \;=\; \sum_{i=1}^{N} \Sigma_i \;\succeq\; N\tau_{\min} I.$$

$v \mapsto vv^\top$ 关于 Loewner序是凸的，因此 Σ^{tot} 关于该序是凹的，并且对于 S值域中各点的任意凸组合 $s = \sum_m \alpha_m S^{(m)}$，

$$\Sigma^{\mathsf{tot}}(s) \;\succeq\; \sum_m \alpha_m\, \Sigma^{\mathsf{tot}}\big(S^{(m)}\big) \;\succeq\; N\tau_{\min} I,$$

由连续性可延拓至闭包。同样，根据假设 1(iv)， $s \mapsto s_3$ 是线性的，并且 $S_3 \ge s_{\min}$ 逐点成立，因此 $s_3 \ge s_{\min}$ 在 S 上成立。最后， $\gamma_k \in [0,1]$ 将 (3.5) 化为凸组合

$$s^{(k)} \;=\; (1-\gamma_k)\, s^{(k-1)} \;+\; \gamma_k\, S\big(x, \mathsf{q}^{(k)}; \varepsilon^{(k)}\big),$$

因此， $s^{(k)} \in \mathsf{S}$ 可由 $s^{(0)} \in \mathsf{S}$ 通过归纳法得到。 □

引理2（M步算子）的证明

证明。 (i) 。 $H_s = -\nabla^2_\theta J(\theta_s,\ s) \succ 0$ 根据假设 2。代理函数 J 关于 θ 在 Θ 上是光滑的，且 M 连续，因此 $s \mapsto H_s$ 在紧集 S 上连续，$s \mapsto \|H_s\|$ 也连续，并且根据 $H_s \succ 0$，$s \mapsto \|H_s^{-1}\|$也连续。因此

$$\sup_{s\in\mathcal{S}} \left\|H_s\right\| \ < \ \infty, \qquad \sup_{s\in\mathcal{S}} \left\|H_s^{-1}\right\| \ < \ \infty.$$

(ii) 。 $\Theta_{\mathcal{S}} = \mathrm{M(S)}$ 是紧集的连续像，因此是紧集，并且包含于 Θ 中。连续的正值函数 $\theta \mapsto \lambda_{\min}(\Omega)$、 $\|\Omega\|^{-1}$、 σ^2、 σ^{-2} 因此在其上取得正最小值，所以对于某个 $c > 0$

$$c\, I \ \preceq \ \Omega \ \preceq \ c^{-1} I, \qquad c \ \le \ \sigma^2 \ \le \ c^{-1}, \qquad \theta = (\hat{z}_{\mathsf{pop}}, \Omega, \sigma^2) \in \Theta_{\mathcal{S}}.$$

由于 C^1 位于紧凸集 S 上，映射 M 在此处是利普希茨的。

(iii) 。由于 $\nabla_\theta J(\theta、\ s) \ = D\zeta(\theta)^\top s - \nabla_\theta \kappa(\theta)$ 关于 s 是仿射的，因此有 $\nabla_s \nabla_\theta J \ = \ D\zeta(\theta_s)^\top$，对一阶条件 $\nabla_\theta J(\mathrm{M}(s),\ s) \ = 0$ 关于 s 求导得到

$$\nabla^2_\theta J(\theta_s, s)\, \nabla_s \mathcal{M}(s) + D\zeta(\theta_s)^\top \ = \ -H_s\, \nabla_s \mathcal{M}(s) + D\zeta(\theta_s)^\top \ = \ 0,$$

即， $\nabla_s \mathrm{M}\,(s) \ = H_s^{-1} D\zeta(\theta_s)^\top$。 □

A.3 步骤 1：变分 E 步

引理 3 的证明（$\mathcal{U}(\theta)$上的光滑性和最大化元）

证明。 在 Q 上， L_i 的各项满足 $l_{\min} \le (L_i)_{aa}$ 和 $|(L_i)_{ab}| \le l_{\max}$，而根据假设 1(ii)， f 和g 的三阶及以下导数均有界；这也使得可以在期望号下对 (5.2) 求两次导数。因此， $\nabla^2_{\mathsf{q}}\mathrm{L}$ 在紧集 $\Theta_{\mathcal{S}} \times \mathrm{Q}$ 上连续，并且

$$\mathcal{K}_{\max} \ := \ \sup_{\Theta_{\mathcal{S}} \times Q} \left\|\nabla^2_{\mathsf{q}} \mathcal{L}\right\| \ < \ \infty.$$

因此， $\nabla_{\mathsf{q}}\mathrm{L}(\theta, \cdot)$ 在 Q 上是 $\mathrm{K}_{\max}$-利普希茨函数，所以 $\mathrm{L}(\theta, \cdot)$ 是 $\mathrm{K}_{\max}$-光滑的。集合 $\mathrm{U}\,(\theta)$ 是一个球，因而是凸集，并且 $r < l_{\min}$ 将其包含于 $\mathbb{R}^{N n_z} \times \mathrm{T}_+^N$ 内。根据假设3，

$$\nabla^2_{\mathsf{q}} \mathcal{L}(\theta, \mathsf{q}) \ \preceq \ -\mathcal{K}_{\min} I \quad \text{on } \mathcal{U}(\theta), \qquad \nabla_{\mathsf{q}} \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big) \ = \ 0,$$

所以 $\mathsf{q}^\star(\theta)$ 是严格凹函数 $\mathrm{L}(\theta, \cdot)$ 在 $\mathrm{U}\,(\theta)$ 上的唯一最大化元。 □

引理 4 的证明（变分包络恒等式）

证明。 根据假设3，映射 $\mathsf{q}^\star$ 在包含紧集 $\Theta_{\mathcal{S}}$ 的某个开集上 C^1 ，因此其导数在那里有界，且$\mathsf{q}^\star$是利普希茨的；由于 S 在 Q 上是利普希茨的， $\bar{S} = S \circ \mathsf{q}^\star$也是利普希茨的。由于$\mathsf{q}^\star(\theta) \ \in \mathrm{Q}$,

$$\bar{S}(\theta) \ = \ \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,I)} \left[S\big(x, \mathsf{q}^\star(\theta); \varepsilon\big) \right] \ \in \ \mathcal{S}$$

作为 S 在 Q 上的值域中的点的平均值。链式法则和一阶条件$\nabla_{\mathsf{q}}\mathrm{L}(\theta, \mathsf{q}^\star(\theta)) \ = 0$ 给出

$$\begin{aligned}
\nabla_\theta \mathcal{L}^\star(\theta) =& \nabla_\theta \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big) + \nabla_\theta \mathsf{q}^\star(\theta)^\top \underbrace{\nabla_{\mathsf{q}} \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big)}_{=0} \\
=& \nabla_\theta \mathcal{L}\big(\theta, \mathsf{q}^\star(\theta)\big) \\
=& \nabla_\theta J\big(\theta, \bar{S}(\theta)\big) \\
=& D\zeta(\theta)^\top \bar{S}(\theta) - \nabla_\theta \kappa(\theta),
\end{aligned}$$

其中第二行使用了 $\nabla_\theta \mathrm{L}(\theta、\ \mathsf{q}) \ = \nabla_\theta J(\theta、\ S(\mathsf{q}))$，因为 θ仅通过 J进入证据下界。 ¯右端在 $\Theta_{\mathcal{S}}$上是利普希茨的，因为 J 是光滑的，而 S 在紧集上是利普希茨的。□

A.4 步骤 2：Robbins-Monro 形式与内循环误差

引理 5 的证明（噪声条件）
证明。令 $\mathcal{G}_k^{(1)} \supseteq \sigma(s^{(0)}, \psi^{(0)}, \{\varepsilon^{(l)}\}_{l\le k-1})$ 和 $\widetilde{\mathcal{F}}_k = \sigma(\mathsf{G}_k,\ \varepsilon^{(k)})$。迭代量 $\mathsf{q}^{(k)} = \mathsf{q}(\psi_L)$ 是 $\psi^{(k-1)}$ 以及第 k 次迭代的内循环噪声的可测函数，因此是 G_k- 可测的，而 $\varepsilon^{(k)}$ 是重新独立抽取的。因此

$$\mathbb{E}\big[S(x,\mathsf{q}^{(k)};\varepsilon^{(k)})\,\big|\,\mathcal{G}_k\big] \;=\; \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\big[S(x,\mathsf{q}^{(k)};\varepsilon)\big] \;=\; S(\mathsf{q}^{(k)}), \qquad \mathbb{E}\big[\Delta_k^{\mathrm{mc}}\,\big|\,\mathcal{G}_k\big] \;=\; 0.$$

由于根据（A.1），$\|S(x,\ \mathsf{q};\ \varepsilon)\| \le C_S$，因此 $\|\Delta_k^{\mathrm{mc}}\| \le 2C_S$，所以部分和 $\mathrm{N}_n = \sum_{k\le n}\gamma_k\Delta_k^{\mathrm{mc}}$ 关于 (F_n) 构成鞅，并且

$$\sum_{k\ge1}\mathbb{E}\Big[\big\|\gamma_k\Delta_k^{\mathrm{mc}}\big\|^2\Big] \;\le\; 4C_S^2\sum_{k\ge1}\gamma_k^2 \;<\; \infty,$$

因此 N_n 是 L^2-有界的，并且几乎必然收敛。最后，$\mathsf{q}^{(k)}$ 和 $\mathsf{q}^\star(\theta_{(k-1)})$ 都属于 Q，在 Q 上 S 满足利普希茨条件，利普希茨常数为 $L_S<\infty$，因此

$$\big\|\Delta_k^{\mathrm{in}}\big\| \;=\; \big\|S(\mathsf{q}^{(k)}) - S(\mathsf{q}^\star(\theta^{(k-1)}))\big\| \;\le\; L_S\,\delta_k.$$

□

命题 1 的证明（消失的内循环误差）
证明。 根据 S 中两点的更新 $s^{(k-1)} = s^{(k-2)} + \gamma_{k-1}\big(S(x,\ \mathsf{q}^{(k-1)};\ \varepsilon^{(k-1)}) - s^{(k-2)}\big)$，

$$\big\|s^{(k-1)} - s^{(k-2)}\big\| \;\le\; \gamma_{k-1}\,\mathrm{diam}(\mathcal{S}), \qquad \mathrm{diam}(\mathcal{S}) = \sup_{s,s'\in\mathcal{S}}\|s-s'\| < \infty.$$

由于 $\mathsf{q}^\star$ 和 M 均为利普希茨函数（引理 2 和引理 4），结合 $c_1 = L_{\mathsf{q}^\star}L_{\mathcal{M}}\,\mathrm{diam}(\mathsf{S})$ 可得

$$\begin{aligned}\big\|\mathsf{q}^{(k-1)} - \mathsf{q}^\star(\theta^{(k-1)})\big\| &\le \delta_{k-1} + \big\|\mathsf{q}^\star(\theta^{(k-2)}) - \mathsf{q}^\star(\theta^{(k-1)})\big\| \\ &\le \delta_{k-1} + L_{\mathsf{q}^\star}L_{\mathcal{M}}\big\|s^{(k-1)} - s^{(k-2)}\big\| \\ &\le \delta_{k-1} + c_1\,\gamma_{k-1}.\end{aligned}$$

将其代入压缩 $\delta_k \le \rho\,\|\mathsf{q}^{(k-1)} - \mathsf{q}^\star(\theta_{(k-1)})\| + c_0\alpha_k$，并利用 $\alpha_k \le c_\alpha\gamma_k$（假设5），

$$\delta_k \;\le\; \rho\,\delta_{k-1} + \rho\,c_1\gamma_{k-1} + c_0c_\alpha\gamma_k,$$

并且有 $u_k = \delta_k/\gamma_k$，

$$u_k \;\le\; \rho\,\frac{\gamma_{k-1}}{\gamma_k}\,u_{k-1} + \rho\,c_1\frac{\gamma_{k-1}}{\gamma_k} + c_0c_\alpha.$$

令 $\epsilon>0$，其中 $\rho(1+\epsilon)<1$。由于 $\gamma_{k-1}/\gamma_k\to1$，存在 k_0，使得 $\gamma_{k-1}/\gamma_k \le 1+\epsilon$ 对于 $k\ge k_0$, so

$$u_k \;\le\; \rho(1+\epsilon)\,u_{k-1} + \big(\rho\,c_1(1+\epsilon) + c_0c_\alpha\big) \qquad (k\ge k_0),$$

并迭代这一仿射压缩，得到

$$\limsup_{k\to\infty}\;\frac{\delta_k}{\gamma_k} \;\le\; \frac{\rho\,c_1(1+\epsilon) + c_0c_\alpha}{1-\rho(1+\epsilon)}.$$

左侧不依赖于 ϵ，因此令 $\epsilon\downarrow0$，可得

$$\limsup_{k\to\infty}\;\frac{\delta_k}{\gamma_k} \;\le\; \frac{\rho\,c_1 + c_0c_\alpha}{1-\rho} \;<\; \infty.$$

因此，根据假设 4，$\delta_k = O(\gamma_k)\to0$和 $\sum_k\gamma_k\delta_k = O\big(\sum_k\gamma_k^2\big)<\infty$。

□

命题 2（渐近伪轨道）的证明

证明。 将更新式 (5.4) 写为 $s^{(k)} = s^{(k-1)} + \gamma_k g_k$，其中 $g_k := \mathrm{F}(s^{(k-1)}) + \Delta_k^{\mathrm{mc}} + \Delta_k^{\mathrm{in}}$。根据引理 5 和 $\mathsf{q}^{(k)}$，$\mathsf{q}^\star(\theta_{(k-1)}) \in \mathsf{Q}$，

$$\|g_k\| \;\le\; c_g \;:=\; \sup_{\mathcal{S}} \|\mathcal{F}\| + 2C_S + L_S \operatorname{diam}(\mathcal{Q}). \tag{A.2}$$

固定 $T > 0$ 和 $n \in \mathbb{N}$，令 $m := \max\{k : t_k \le t_n + T\}$，并对 $n \le k \le m$，设

$$E_k := \big\| s^{(k)} - \Phi^{\mathcal{F}}_{t_k - t_n}\big(s^{(n)}\big) \big\|, \qquad W_{n,k} := \sum_{j=n+1}^{k} \gamma_j \big(\Delta_j^{\mathrm{mc}} + \Delta_j^{\mathrm{in}}\big).$$

将更新从 $j = n+1$、……、k 求和，得到 $s^{(k)} = s^{(n)} + \sum_{j=n+1}^{k} \gamma_j \mathrm{F}(s^{(j-1)}) + W_{n,\,k}$。减去积分形式 $\Phi^{\mathcal{F}}_{t_k - t_n}(s^{(n)}) = s^{(n)} + \int_{t_n}^{t_k} \mathrm{F}(\Phi^{\mathcal{F}}_{u-t_n}(s^{(n)}))\, du$，并利用 $\|\Phi^{\mathcal{F}}_{t_{j-1}-t_n}(s^{(n)}) - \Phi^{\mathcal{F}}_{u-t_n}(s^{(n)})\| \le c_{\mathcal{F}} \gamma_j$，其中 $u \in [t_{j-1}, t_j]$ 满足 $c_{\mathcal{F}} = \sup_{\mathcal{S}} \|\mathrm{F}\|$，

$$\begin{aligned} E_k &\le \sum_{j=n+1}^{k} \int_{t_{j-1}}^{t_j} \Big\| \mathcal{F}\big(s^{(j-1)}\big) - \mathcal{F}\big(\Phi^{\mathcal{F}}_{u-t_n}(s^{(n)})\big) \Big\|\, du \;+\; \big\| W_{n,k} \big\| \\ &\le L_{\mathcal{F}} \sum_{j=n+1}^{k} \gamma_j E_{j-1} \;+\; L_{\mathcal{F}} c_{\mathcal{F}} \sum_{j=n+1}^{k} \gamma_j^2 \;+\; \sup_{n \le l \le k} \big\| W_{n,l} \big\|. \end{aligned}$$

由 $E_n = 0$ 和 $\sum_{j=n+1}^{m} \gamma_j \le T + \gamma_{n+1}$，离散 Gronwall 不等式给出

$$\max_{n \le k \le m} E_k \;\le\; e^{L_{\mathcal{F}}(T + \gamma_{n+1})} \Big(L_{\mathcal{F}} c_{\mathcal{F}} \sum_{j>n} \gamma_j^2 + \sup_{n \le l \le m} \big\| W_{n,l} \big\| \Big), \tag{A.3}$$

其中

$$\sup_{n \le l \le m} \big\| W_{n,l} \big\| \;\le\; \underbrace{\sup_{n \le l \le m} \Big\| \sum_{j=n+1}^{l} \gamma_j \Delta_j^{\mathrm{mc}} \Big\|}_{\to 0 \text{ a.s. (Lemma 5)}} \;+\; \underbrace{L_S \sum_{j>n} \gamma_j \delta_j}_{\to 0 \text{ (Proposition 1)}} .$$

此外，由于 $\sum_j \gamma_j^2 < \infty$，(A.3) 右端几乎必然趋于 0，这就是网格点处的 (5.7)。对于 $t \in [t_n, t_{n+1}]$，插值与 $s^{(n)}$ 的差至多为 $\gamma_{n+1}\|g_{n+1}\| \le c_g \gamma_{n+1} \to 0$，且 $\Phi^{\mathcal{F}}$ 关于时间是利普希茨的，因此结论一致成立。有界利普希茨半流的渐近伪轨道的极限集非空、紧、连通且不变 [7, 34],；由于 S 是紧集，有界性也成立。□

A.5 第 3 步：李雅普诺夫稳定性

引理 6（李雅普诺夫衰减）的证明

证明。 这里结合了三个恒等式。首先，$\mathrm{M}(s)$ 是 $J(\cdot, s)$ 的驻点，因此 (4.2) 可写为

$$D\zeta(\theta_s)^{\top} s \;=\; \nabla_\theta \kappa(\theta_s). \tag{A.4}$$

其次，在 θ_s 处计算包络恒等式 (5.3)，并代入 (A.4)，即可通过平均场表示 $\mathrm{L}^\star$ 的梯度，

$$\begin{aligned} \nabla_\theta \mathcal{L}^\star(\theta_s) &= D\zeta(\theta_s)^{\top} \bar{S}(\theta_s) - \nabla_\theta \kappa(\theta_s) \\ &= D\zeta(\theta_s)^{\top} \big(\bar{S}(\theta_s) - s \big) \\ &= D\zeta(\theta_s)^{\top} \mathcal{F}(s). \end{aligned} \tag{A.5}$$

第三，敏感性 (5.1) 和链式法则给出

$$\nabla_s \mathcal{V}(s) \;=\; -\nabla_s \mathcal{M}(s)^\top \nabla_\theta \mathcal{L}^\star(\theta_s) \;=\; -D\zeta(\theta_s) H_s^{-1} \nabla_\theta \mathcal{L}^\star(\theta_s). \tag{A.6}$$

由于 J 在（θ, s）上光滑，M 在 S 上是利普希茨的，且上确界$_s \|H_s^{-1}\| < \infty$，因此 (A.6) 中的每个因子在紧集 S 上都是利普希茨的，∇_sV 也如此。结合 (A.5) 和 (A.6),

$$\begin{aligned}\big\langle \nabla_s \mathcal{V}(s),\, \mathcal{F}(s) \big\rangle &= -\nabla_\theta \mathcal{L}^\star(\theta_s)^\top H_s^{-1}\, D\zeta(\theta_s)^\top \mathcal{F}(s) \\ &= -\nabla_\theta \mathcal{L}^\star(\theta_s)^\top H_s^{-1}\, \nabla_\theta \mathcal{L}^\star(\theta_s) \\ &= -\big\|\nabla_\theta \mathcal{L}^\star(\theta_s)\big\|^2_{H_s^{-1}},\end{aligned}$$

由此得到 (5.8)，因为 $H_s^{-1} \succ 0$。(5.9) 中的界取 $c_{\mathsf{dec}} = \inf_{s\in\mathcal{S}} \|H_s\|^{-1} > 0$ 和$c_{\mathsf{grad}} = \sup_{s\in\mathcal{S}} \|D\zeta(\theta_s) H_s^{-1}\| < \infty$时成立；根据引理2和 S 的紧性，这两个量均有限且为正。□

命题 3 的证明（离散李雅普诺夫衰减）

证明。 如命题 2 的证明所示， $s^{(k)} = s^{(k-1)} + \gamma_k g_k$取 $g_k = \mathrm{F}\,(s^{(k-1)}) + \Delta_k^{\mathrm{mc}} + \Delta_k^{\mathrm{in}}$和$\|g_k\| \le c_g$。令 $c_{\mathcal{V}} = \sup_{\mathcal{S}} \|\nabla \mathrm{V}\| < \infty$。由于 ∇V 在凸集 S 上是利普希茨的，根据下降引理、引理 6 中的衰减 ∇V (s)，F (s) $\le -c_{\mathsf{dec}} \|\nabla_\theta \mathrm{L}^\star(\theta_s)\|^2$ 以及 $\|\Delta_k^{\mathrm{in}}\| \le L_S \delta_k$，取$c_2 = L_{\nabla\mathcal{V}} c_g^2/2$，可得

$$\begin{aligned}\mathcal{V}\big(s^{(k)}\big) &\le \mathcal{V}\big(s^{(k-1)}\big) + \gamma_k \big\langle \nabla\mathcal{V}(s^{(k-1)}), g_k \big\rangle + \tfrac{L_{\nabla\mathcal{V}}}{2}\gamma_k^2 \|g_k\|^2 \\ &\le \mathcal{V}\big(s^{(k-1)}\big) - \gamma_k\, c_{\mathsf{dec}} \big\|\nabla_\theta \mathcal{L}^\star(\theta^{(k-1)})\big\|^2 + \gamma_k \big\langle \nabla\mathcal{V}(s^{(k-1)}), \Delta_k^{\mathrm{mc}} \big\rangle \\ &\quad + \gamma_k \big\|\nabla\mathcal{V}(s^{(k-1)})\big\|\, L_S \delta_k + c_2 \gamma_k^2.\end{aligned} \tag{A.7}$$

该逐步不等式仅使用假设 1-3。这里， $\|\nabla \mathrm{V}\,(s^{(k-1)})\|$由 $c_{\mathcal{V}}$界定。从 1 求和至 n,

$$\begin{aligned}\mathcal{V}\big(s^{(n)}\big) + c_{\mathsf{dec}} \sum_{k=1}^{n} \gamma_k \big\|\nabla_\theta \mathcal{L}^\star(\theta^{(k-1)})\big\|^2 \;\le\; \mathcal{V}\big(s^{(0)}\big) + \underbrace{\sum_{k=1}^{n} \gamma_k \big\langle \nabla\mathcal{V}(s^{(k-1)}), \Delta_k^{\mathrm{mc}} \big\rangle}_{=:\,\mathcal{Z}_n} \\ + c_{\mathcal{V}} L_S \underbrace{\sum_{k=1}^{n} \gamma_k \delta_k}_{<\infty} + c_2 \underbrace{\sum_{k=1}^{n} \gamma_k^2}_{<\infty}.\end{aligned} \tag{A.8}$$

The 过程 (Z$_n$) 是一个鞅，因为 ∇V ($s^{(k-1)}$) 是 G$_k$-可测的，且 $\mathbb{E}[\Delta_k^{\mathrm{mc}} | \mathsf{G}_k] = 0$，并且

$$\sum_{k\ge 1} \mathbb{E}\Big[\big|\mathcal{Z}_k - \mathcal{Z}_{k-1}\big|^2\Big] \;\le\; 4c_{\mathcal{V}}^2 C_S^2 \sum_{k\ge 1} \gamma_k^2 \;<\; \infty,$$

因此 Z$_n$ 几乎必然收敛。由于 V 在 S 上有界，(A.8) 左侧有上界，从而得到 (5.10)。此外，V ($s^{(n)}$) 是一个收敛鞅、一个收敛余项和一个有下界的不增项之和，因此也收敛。□

命题 4 的证明（消失的平稳性残差）

证明。 记 $R_k := \|\nabla_\theta \mathrm{L}^\star(\theta_{(k)})\|^2$。映射 $s \mapsto \|\nabla_\theta \mathrm{L}^\star(\mathrm{M}\,(s))\|^2$ 在紧集 S 上满足利普希茨条件，其常数为 L_R，并且根据 (A.2)， $\|s^{(k)} - s^{(k-1)}\| \le c_g \gamma_k$，因此

$$\big|R_k - R_{k-1}\big| \;\le\; L_R c_g\, \gamma_k. \tag{A.9}$$

写作 $c := L_R c_g$，可假设其为正数，因为否则 R 为常数，并且根据 (5.10) 以及 $\sum_k \gamma_k = \infty$，恒等于零。

令 $\epsilon > 0$。根据命题 3，级数 $\sum_k \gamma_k R_{k-1}$ 收敛，因此存在 n_1，使得

$$\sum_{k>n} \gamma_k R_{k-1} \;<\; \frac{\epsilon^2}{c} \qquad \text{for every } n \ge n_1. \tag{A.10}$$

可以证明，对于每个 $n \ge n_1$，都有 $R_n < 2\epsilon$。反设对某个 $n_0 \ge n_1$有 $R_{n_0} \ge 2\epsilon$。如果对所有 $k \ge n_0$ 都有 $R_k \ge \epsilon$，则根据假设 4，有 $\sum_{k>n_0} \gamma_k R_{k-1} \ge \epsilon \sum_{k>n_0} \gamma_k = \infty$，这与（A.10）矛盾。令 $m > n_0$ 为满足 $R_m < \epsilon$的最小索引，则对于 $n_0 \le k < m$有 $R_k \ge \epsilon$。根据（A.9），

$$\epsilon \le R_{n_0} - R_m \le \sum_{k=n_0+1}^{m} \left| R_k - R_{k-1} \right| \le c \sum_{k=n_0+1}^{m} \gamma_k,$$

因此 $\sum_{k=n_0+1}^{m} \gamma_k \ge \epsilon / c$。由于 $R_{k-1} \ge \epsilon$对于 $n_0 < k \le m$,

$$\sum_{k>n_0} \gamma_k R_{k-1} \ge \sum_{k=n_0+1}^{m} \gamma_k R_{k-1} \ge \epsilon \sum_{k=n_0+1}^{m} \gamma_k \ge \frac{\epsilon^2}{c},$$

再次与 (A.10) 矛盾。因此，对所有 $n \ge n_1$， $0 \le R_n < 2\epsilon$均成立；又由于 $\epsilon > 0$是任意的，$R_k \to 0$几乎必然成立。□

A.6 步骤 4：平稳最优性条件

引理7的证明（极限集刻画）

证明。 第二个恒等式是假设3的一阶条件。根据包络恒等式 (5.3),

$$0 = \nabla_\theta \mathcal{L}^\star(\theta^\star) = \nabla_\theta \mathcal{L}(\theta^\star, \mathsf{q}^\star) = \nabla_\theta J\big(\theta^\star, \bar{S}(\theta^\star)\big),$$

这给出了第一个恒等式；结合 $s^\star = \bar{S}(\theta^\star)$, 则得到第三个恒等式。逆向阅读同一推导链即可得出逆命题。□

A.7 定理1-3的证明

主要结果还用到另外两个引理。第一个引理考虑一个以速率 γ_k 向某一水平压缩的序列，并证明即使该压缩受到扰动，其上极限仍位于该水平之下。

引理 9（扰动收缩）。 设 $a > 0$、 $\gamma_k \ge 0$ 满足 $\sum_k \gamma_k = \infty$，且 $\beta_k \in \mathbb{R}$ 满足上极限$_k \beta_k \le \beta$。若 $u_k \in \mathbb{R}$ 对所有充分大的 k 满足 $a\gamma_k \le 1$，且

$$u_k \le (1 - a\gamma_k)\, u_{k-1} + \gamma_k \beta_k, \tag{A.11}$$

则 上极限$_k u_k \le \beta / a$。

证明。 设 $\epsilon > 0$、 $c = (\beta + \epsilon)/a$，并取 k_0，使得对所有 $k > k_0$， $a\gamma_k \le 1$、 $\beta_k \le ac$ 与 (A.11) 成立。则

$$u_k - c \le (1 - a\gamma_k)\, u_{k-1} + a\gamma_k\, c - c = (1 - a\gamma_k)\,(u_{k-1} - c),$$

并且结合 $1 - x \le e^{-x}$ 与 $\sum_k \gamma_k = \infty$, 超出量 $w_k = \max\{u_k - c, 0\}$ 满足

$$w_k \le (1 - a\gamma_k)\, w_{k-1} \le e^{-a\gamma_k}\, w_{k-1} \le w_{k_0} \exp\Big(- a \sum_{j=k_0+1}^{k} \gamma_j \Big) \xrightarrow[k\to\infty]{} 0.$$

因此，对于每个 $\epsilon > 0$, 上极限$_k u_k \le (\beta + \epsilon)/a$。 □

第二引理是定理 2 和定理 3 中局部结论的核心。它针对一般目标函数 f 进行表述；在定理 2 中，f 为 $\mathrm{L}^\star$，在定理 3 中则为 LL。沿着迭代序列，令

$$e_k = f(\theta^\star) - f\big(\theta^{(k)}\big), \qquad G_k = \big\| \nabla_\theta f\big(\theta^{(k)}\big) \big\|$$

与最大值的差距和梯度范数。

引理 10（局部捕获）。设 f 在包含 $\Theta' \subset \Theta$ 的开集上连续可微，并且在点 $\theta^\star \in \Theta'$ 附近二阶连续可微，其中 $\nabla_\theta f\,(\theta^\star) = 0$ 且 $m = \lambda_{\min}\!\big(-\nabla^2_\theta f\,(\theta^\star)\big) > 0$。设 $\theta^{(k)} \in \Theta'$ 满足 $\|\theta^{(k)} - \theta^{(k-1)}\| \to 0$ 下降不等式

$$e_k \;\le\; e_{k-1} - \gamma_k\, c_{\mathsf{dec}}\, G_{k-1}^2 + \gamma_k\, c_{\mathsf{grad}}\, G_{k-1}\, b_k + \gamma_k \xi_k + c_2 \gamma_k^2, \tag{A.12}$$

其中 γ_k 满足假设 *4*，$c_{\mathsf{dec}} > 0$ 和 c_{grad}是常数，偏差满足 $b_k \ge 0$且上极限$_k\, b_k \le b$，噪声 ξ_k 满足 $\sum_k \gamma_k \xi_k$ 收敛。于是存在 $\theta^\star$ 的一个邻域$\mathcal{U}$，使得如果 $C_b := 2c_{\mathsf{grad}} b/(c_{\mathsf{dec}} m)$ 足够小，且对于无穷多个 k，$\theta^{(k)} \in \mathcal{U}$，

$$\limsup_{k\to\infty} \big\|\theta^{(k)} - \theta^\star\big\| \;\le\; C_b.$$

证明。令 $\Delta_k = \|\theta^{(k)} - \theta^\star\|$。

步骤 *1*，f 在 $\theta^\star$ 附近。由海森矩阵的连续性，存在 $r_\star > 0$，使得在以 $\theta^\star$为中心、半径为 $r_\star$ 的闭球上，$\nabla^2_\theta f \preceq -\frac{m}{2} I$ 成立。令 $\mathrm{B} = \{\theta \in \Theta' :\, \|\theta - \theta^\star\| \le r_\star\}$，$\theta \in \mathrm{B}$ 且 $t = \|\theta - \theta^\star\|$。分别在 $\theta^\star$ 和 θ 处应用泰勒定理，得到

$$\begin{aligned} \frac{m}{4} t^2 \;&\le\; f(\theta^\star) - f(\theta) \;\le\; \big\langle \nabla_\theta f(\theta),\, \theta^\star - \theta \big\rangle - \frac{m}{4} t^2 \\ &\le\; \big\|\nabla_\theta f(\theta)\big\|\, t - \frac{m}{4} t^2 \;\le\; \frac{1}{m} \big\|\nabla_\theta f(\theta)\big\|^2, \end{aligned} \tag{A.13}$$

其中，最后一步对 t取最大值。外部界由二次增长和 Polyak–Łojasiewicz 不等式 [26]给出。

第 *2* 步，在 B 中进行压缩。令 $\theta^{(k-1)} \in \mathrm{B}$ 和 $a = c_{\mathsf{dec}} m/2$。由杨氏不等式以及 (A.13) 中的 $G_{k-1}^2 \ge m\, e_{k-1}$，可得

$$-c_{\mathsf{dec}} G_{k-1}^2 + c_{\mathsf{grad}} G_{k-1} b_k \;\le\; -\frac{c_{\mathsf{dec}}}{2} G_{k-1}^2 + \frac{(c_{\mathsf{grad}} b_k)^2}{2 c_{\mathsf{dec}}} \;\le\; -a\, e_{k-1} + \frac{(c_{\mathsf{grad}} b_k)^2}{2 c_{\mathsf{dec}}}.$$

余项 $\varrho_k = \sum_{j>k} (\gamma_j \xi_j + c_2 \gamma_j^2)$ 趋于零，且满足 $\gamma_k \xi_k + c_2 \gamma_k^2 = \varrho_{k-1} - \varrho_k$。将二者代入 (A.12)，可得对于 $u_k = e_k + \varrho_k$

$$u_k \;\le\; (1 - a\gamma_k)\, e_{k-1} + \gamma_k \frac{(c_{\mathsf{grad}} b_k)^2}{2 c_{\mathsf{dec}}} + \varrho_{k-1} \;=\; (1 - a\gamma_k)\, u_{k-1} + \gamma_k \beta_k, \tag{A.14}$$

其中 $\beta_k = \frac{(c_{\mathsf{grad}} b_k)^2}{2 c_{\mathsf{dec}}} + a\varrho_{k-1}$，并且当 $v_\infty = \frac{m}{4} C_b^2$ 时，

$$\limsup_{k\to\infty} \beta_k \;\le\; \frac{(c_{\mathsf{grad}} b)^2}{2 c_{\mathsf{dec}}} \;=\; \frac{c_{\mathsf{dec}} m}{2} \cdot \frac{m}{4} \Big(\frac{2 c_{\mathsf{grad}} b}{c_{\mathsf{dec}} m}\Big)^2 \;=\; a\, v_\infty.$$

第 *3* 步，迭代序列保持在$\mathcal{B}$ 中。令 $v = \frac{m}{4}\big(\frac{r_\star}{2}\big)^2$ 且

$$\mathcal{U} \;=\; \Big\{\theta \in \Theta' :\, \|\theta - \theta^\star\| < \frac{r_\star}{2} \;\text{ and }\; f(\theta^\star) - f(\theta) < \frac{v}{2}\Big\},$$

根据 f的连续性，$\theta^\star$ 的一个邻域。若 $C_b < r_\star/4$，即 $v_\infty < v/4$，则称常数 C_b 足够小。由于 γ_k、ϱ_k 以及各步长趋于零，且 $\limsup_k \beta_k < \frac{3}{4} a v$，因此存在 k_1，使得对所有 $k \ge k_1$，

$$\big\|\theta^{(k)} - \theta^{(k-1)}\big\| < \frac{r_\star}{2}, \qquad a\gamma_k \le 1, \qquad |\varrho_k| \le \frac{v}{8}, \qquad \beta_k \le \frac{3}{4}\, a\, v.$$

取 $k_2 \ge k_1$，其中 $\theta^{(k_2)} \in \mathrm{U}$。于是 $\Delta_{k_2} < r_\star/2$ 且 $u_{k_2} \le \frac{v}{2} + \frac{v}{8} \le \frac{3}{4} v$。如果对于某个 $k > k_2$，有 $\Delta_{k-1} < r_\star/2$ 且$u_{k-1} \le \frac{3}{4} v$，则由 (A.14) 和 (A.13) 可得

$$\begin{aligned} \Delta_k \;&\le\; \Delta_{k-1} + \big\|\theta^{(k)} - \theta^{(k-1)}\big\| \;<\; \frac{r_\star}{2} + \frac{r_\star}{2} \;=\; r_\star, \\ u_k \;&\le\; (1 - a\gamma_k)\, \frac{3}{4} v + \gamma_k \frac{3}{4} a v \;=\; \frac{3}{4} v, \\ \frac{m}{4} \Delta_k^2 \;&\le\; e_k \;=\; u_k - \varrho_k \;\le\; \frac{3}{4} v + \frac{v}{8} \;<\; v \;=\; \frac{m}{4}\Big(\frac{r_\star}{2}\Big)^2. \end{aligned}$$

根据归纳法，对于所有 $k > k_2$， $\Delta_k < r_\star/2$ 和 (A.14) 均成立。

步骤 *4*，取极限。 将引理9应用于 (A.14)、 $\varrho_k \to 0$ 和 (A.13)，可得

$$\frac{m}{4}\limsup_{k\to\infty}\Delta_k^2 \;\le\; \limsup_{k\to\infty} e_k \;=\; \limsup_{k\to\infty} u_k \;\le\; v_\infty \;=\; \frac{m}{4}C_b^2.$$

□

定理1的证明（ELBO VAE-NLME 算法的收敛性）

证明。 第一个极限由命题 4 得出。根据引理2(ii)，集合 $\Theta_{\mathcal{S}}$是紧集，且 $\theta_{(k)}$=M ($s^{(k)}$) 包含于其中，因此 {$\theta^{(k)}$} 的每个子序列都有一个收敛的子子序列 $\theta_{(k_j)}\to\hat{\theta}$，由 ∇_θL*的连续性，其极限满足

$$\left\|\nabla_\theta \mathcal{L}^\star(\hat{\theta})\right\| \;=\; \lim_{j\to\infty}\left\|\nabla_\theta \mathcal{L}^\star\left(\theta^{(k_j)}\right)\right\| \;=\; 0, \qquad \text{i.e.,} \qquad \hat{\theta}\in\Theta^\star.$$

因此， $\Theta^\star$ 非空，且由于它是紧集 $\Theta_{\mathcal{S}}$的闭子集，故它也是紧集；这是因为 ∇_θL* 连续（引理 4）。由此得到第二个极限，并且根据引理7，($\theta^\star$，q*($\theta^\star$)) 是证据下界的一个驻点。对于最后一个结论，

$$\left\|\theta^{(k)}-\theta^{(k-1)}\right\| \;\le\; L_{\mathcal{M}}\left\|s^{(k)}-s^{(k-1)}\right\| \;\le\; L_{\mathcal{M}}\, c_g\,\gamma_k \;\xrightarrow[k\to\infty]{}\; 0$$

根据引理2(ii)和(A.2)，因此{$\theta^{(k)}$}的极限点集合是连通的 [34] 且包含于 $\Theta^\star$中。由于 $\Theta^\star$ 是紧集，具有孤立元素，因而是有限集，所以它归结为单个点 $\theta^\star$。最后，根据命题 2， L{$s^{(k)}$}的极限点集合是紧的且是不变的，并且由M的连续性，M(L) ={$\theta^\star$}。在 L上，常微分方程 $\dot{s} = \bar{S}(\mathrm{M}(s)) - s$ 因此写成 $\dot{s} = s^\star - s$ ，其流为 $\Phi_t^{\mathcal{F}}(s) = s^\star + e^{-t}(s - s^\star)$。对于 $s\in L$和 $t\ge 0$， 不变性给出 $s_t\in L$， 其中 $\Phi_t^{\mathcal{F}}(s_t) = s$，并且

$$\left\|s - s^\star\right\| \;=\; \left\|\Phi_t^{\mathcal{F}}(s_t) - s^\star\right\| \;=\; e^{-t}\left\|s_t - s^\star\right\| \;\le\; e^{-t}\,\mathrm{diam}(\mathcal{S}) \;\xrightarrow[t\to\infty]{}\; 0.$$

因此 $L = \{s^\star\}$， 即 $s^{(k)}\to s^\star$。 □

定理 2 的证明（持续的内循环不准确性）

证明。 利用 $C = c_{\mathsf{grad}}L_S/c_{\mathsf{dec}}$ 可证明这些断言，其中 c_{dec}， c_{grad}是 (5.9) 中的常数，而 L_S 是 $\|\Delta_k^{\mathrm{in}}\| \le L_S\delta_k$ 中的常数。记 $G_k = \|\nabla_\theta \mathrm{L}^\star(\theta_{(k)})\|$ 和 $\xi_k = \nabla\mathrm{V}(s^{(k-1)})$， Δ_k^{mc}。将更紧的界 $\|\nabla\mathrm{V}(s^{(k-1)})\| \le c_{\mathsf{grad}}G_{k-1}$（见 (5.9)）代入逐步不等式 (A.7)，得到

$$\mathcal{V}\left(s^{(k)}\right) \;\le\; \mathcal{V}\left(s^{(k-1)}\right) - \gamma_k\, c_{\mathsf{dec}}\, G_{k-1}^2 + \gamma_k\, c_{\mathsf{grad}}\, G_{k-1}L_S\delta_k + \gamma_k\xi_k + c_2\gamma_k^2. \tag{A.15}$$

下限 (4.7)。 假设它以正概率在某个事件上不成立。那么存在 $\epsilon_1 > 0$、$\eta > c_{\mathsf{grad}}(L_S\delta+\epsilon_1)/c_{\mathsf{dec}}$ 和 k_0，满足 $G_k\ge\eta$，且对所有 $k\ge k_0$ 都有 $L_S\delta_k \le L_S\delta+\epsilon_1$。因此，对于 $k > k_0$，

$$\begin{aligned} -c_{\mathsf{dec}}\,G_{k-1}^2 + c_{\mathsf{grad}}\,G_{k-1}L_S\delta_k \;&\le\; -G_{k-1}\Big(c_{\mathsf{dec}}\,G_{k-1} - c_{\mathsf{grad}}\big(L_S\delta+\epsilon_1\big)\Big) \\ &\le\; -\eta\Big(c_{\mathsf{dec}}\,\eta - c_{\mathsf{grad}}\big(L_S\delta+\epsilon_1\big)\Big) \;=:\; -\epsilon_0 \;<\; 0, \end{aligned}$$

并对（A.15）求和，得到

$$\mathcal{V}\left(s^{(n)}\right) \;\le\; \mathcal{V}\left(s^{(k_0)}\right) - \epsilon_0\sum_{k=k_0+1}^{n}\gamma_k + \mathcal{Z}_n - \mathcal{Z}_{k_0} + c_2\sum_{k>k_0}\gamma_k^2 \;\xrightarrow[n\to\infty]{}\; -\infty,$$

由于鞅 $\mathrm{Z}_n = \sum_{k\le n}\gamma_k\xi_k$ 收敛， $\sum_k\gamma_k^2 < \infty$ 和 $\sum_k\gamma_k = \infty$。这与 V 在 S 上有界相矛盾。

邻域 (4.8)。 由于 V ($s^{(k)}$) = –L*($\theta_{(k)}$)，将 L*($\theta^\star$) 加到（A.15）上，得到当e_k =L*($\theta^\star$) – L*($\theta_{(k)}$) 时的（A.12），其中

$$b_k \;=\; L_S\delta_k, \qquad \limsup_{k\to\infty} b_k \;\le\; L_S\delta \;=\; b, \qquad C_b \;=\; \frac{2\,c_{\mathsf{grad}}\,L_S}{c_{\mathsf{dec}}\,m}\,\delta \;=\; \frac{2C}{m}\,\delta \;=\; C_\delta,$$

且 $\sum_k\gamma_k\xi_k$ =lim$_n$Z$_n$收敛。由于 $\|\theta_{(k)}-\theta_{(k-1)}\| \le L_{\mathcal{M}}c_g\gamma_k \to 0$与定理1的证明中一样，引理 10 适用于 f = L* 和 $\Theta' = \Theta_{\mathcal{S}}$。 □

对于定理3，先证明引理 8，并汇总重要性加权递归的性质。

引理 8（加权统计量的一致性）的证明

证明。将 $w_i(z) = p(x_i, z;\, \theta)/r_i(z)$ 记为未归一化权重，并将 u 记为 (3.8) 的一个分量，即 z_i、$z_i z_i^\top$ 或 $\sum_l r_{il}(z_i)^2$。由假设 1(i)，$p(x_i|z;\, \theta)$ 有界，因此 $p(x_i, z;\, \theta)$ 具有与先验相同的高斯尾部，而 t 密度 r_i 仅按多项式衰减。因此，w_i 和 $w_i u$ 在 $\mathbb{R}^{n_z}$ 上有界。此外，$|u(z)| \le c\,(1 + \|z\|^2)$，故 u 在 r_i 下具有有限的 $2p$ 阶矩，对每个 $p < \nu/4$ 均成立；特别地，对某个 $p > 1$ 也成立，因为 $\nu > 4$。分子和分母分别应用强大数定律，得到 [23, 定理 1]

$$\begin{aligned}\sum_{j=1}^{K} \tilde{w}_{ij}\, u(z_{ij}) \;&=\; \frac{\frac{1}{K}\sum_{j=1}^{K} w_i(z_{ij})\, u(z_{ij})}{\frac{1}{K}\sum_{j=1}^{K} w_i(z_{ij})} \;\xrightarrow[K\to\infty]{\text{a.s.}}\; \frac{\mathbb{E}_{z\sim r_i}\big[w_i(z)\, u(z)\big]}{\mathbb{E}_{z\sim r_i}\big[w_i(z)\big]} \\ &=\; \frac{\int u(z)\, p(x_i, z;\theta)\, dz}{\int p(x_i, z;\theta)\, dz} \;=\; \mathbb{E}_{z\sim p(\cdot|x_i;\theta)}\big[u(z)\big],\end{aligned}$$

$\widehat{\ \ }$ $\bar{\ \ }$并对 i 求和，几乎必然得到 $S_K \to S_p(\theta)$。对于前两个分块，界 (5.13) 可由 [1, 定理 2.3], 得出；该定理将自归一化重要性采样的偏差和均方误差界定为对于无界测试函数，与 $1/K$ 成比例的常数。此常数由 u 的 $2p$ 阶矩以及 $w_i u$ 和 w_i 在 r_i 下的矩给出，根据上述结果，这些矩均有限。对于有界的第三个分块，[1, 定理 2.1] 给出的常数与 $\rho(\theta,\mathsf{q})$ 成正比。由控制收敛定理可知，所有矩均关于 (θ, q) 连续，因此 c_{is} 在 $\Theta\times\mathsf{Q}$ 的紧子集上有界。□

引理 11(重要性加权递归)。 设定理 *3* 的假设成立。那么

(i) Δ_k^{mc} *is a martingale difference,* $\mathbb{E}[\Delta_k^{\text{mc}} \mid \mathcal{G}_k] = 0$*, with* $\mathbb{E}[\|\Delta_k^{\text{mc}}\|^2 \mid \mathcal{G}_k] \le \bar{c}_{\text{is}}/K$*, and the bias satisfies*

$$\big\|\Delta_k^{\text{is}}\big\| \;\le\; \frac{\bar{c}_{\text{is}}}{K}, \qquad \bar{c}_{\text{is}} \;=\; \sup\big\{c_{\text{is}}(\theta,\mathsf{q}) :\, \theta\in\Theta_{\mathcal{S}_p},\, \mathsf{q}\in\mathcal{Q}\big\} \;<\; \infty, \tag{A.16}$$

其中 c_{is} 来自引理 *8*,

(ii) $\mathcal{V}_p$ 在 $\mathcal{S}_p$ 上连续可微且具有 *Lipschitz* 梯度,

$$\big\langle \nabla_s \mathcal{V}_p(s),\, \mathcal{F}_p(s)\big\rangle \;=\; -\big\|\nabla_\theta \mathcal{LL}(\theta_s)\big\|^2_{H_s^{-1}} \;\le\; 0, \tag{A.17}$$

并且，若以 $\mathcal{LL}$ 替代$\mathcal{L}^\star$，则(5.9) 中的界仍成立，其中常数 $c_{\text{dec}}, c_{\text{grad}}$ 取遍$\mathcal{S}_p$。

证明。 (*i*) 。给定 G_k，迭代值 $\theta^{(k-1)} \in \Theta_{\mathcal{S}_p}$ 和编码器输出$\mathsf{q}^{(k)} \in \mathsf{Q}$ 均为固定值。条件方差有均方误差作为上界，因此在给定 G_k的条件下，两个界均为 (5.13)。上确界有限，因为 $\Theta_{\mathcal{S}_p} = \mathsf{M}(\mathsf{S}_p)$ 是紧集的连续像（假设 2 和 6），因而是紧集，且 Q 也是紧集。 (*ii*) 。此处的计算与引理 6 相同，只是以费舍尔恒等式 $\nabla_\theta \mathrm{LL}(\theta) = D\zeta(\theta)^\top \bar{S}_p(\theta) - \nabla_\theta \kappa(\theta)$ 替代包络恒等式。由 $\nabla_s \mathsf{M}(s) = H_s^{-1} D\zeta(\theta_s)^\top$以及链式法则可得，$\nabla_s \mathrm{V}_p(s) = -D\zeta(\theta_s) H_s^{-1} \nabla_\theta \mathrm{LL}(\theta_s)$，并且

$$\begin{aligned} D\zeta(\theta_s)^\top\big(\bar{S}_p(\theta_s) - s\big) \;&=\; \big[D\zeta(\theta_s)^\top \bar{S}_p(\theta_s) - \nabla_\theta\kappa(\theta_s)\big] - \big[D\zeta(\theta_s)^\top s - \nabla_\theta\kappa(\theta_s)\big] \\ &=\; \nabla_\theta \mathcal{LL}(\theta_s) - 0, \end{aligned}$$

其中，第二个括号根据 M步的一阶条件 (A.4) 消失。与 $-H_s^{-1}\nabla_\theta \mathrm{LL}(\theta_s)$ 取内积得到 (A.17)。常数由引理2(i)按照引理 6 得出。□

定理3的证明（VAE-NLME 算法的收敛性）

证明。这些断言通过 $C = c_{\text{grad}}\bar{c}_{\text{is}}/c_{\text{dec}}$ 得到证明，其中 c_{dec}， c_{grad}是引理 11(ii) 中的常数，而 $\bar{c}_{\text{is}}$是引理 11(i) 中的常数。记 $g_k = \mathrm{F}_P(s^{(k-1)}) + \Delta_k^{\text{mc}} + \Delta_k^{\text{is}}$，于是 $s^{(k)} = s^{(k-1)} + \gamma_k g_k$。与变分统计量不同，$g_k$ 并不有界，但根据引理 11(i)，其条件二阶矩是有界的，因为 F_P 在紧集 S_P上有界。根据假设 4，

$$\mathbb{E}\Big[\sum_{k\ge 1}\gamma_k^2\|g_k\|^2\Big] < \infty, \qquad \text{hence} \qquad \sum_{k\ge 1}\gamma_k^2\|g_k\|^2 < \infty \quad \text{and} \quad \gamma_k\|g_k\| \to 0 \tag{A.18}$$

几乎必然。特别地，步长 $\|\theta^{(k)} - \theta^{(k-1)}\| \le L_{\mathcal{M}}\gamma_k\|g_k\|$趋于零，其中 M 在 S_P 上的 Lipschitz 常数由 S_P 上的引理2(ii)给出。与命题 3 的证明一样，利用下降引理和引理 11 可得

$$\mathcal{V}_p\big(s^{(k)}\big) \le \mathcal{V}_p\big(s^{(k-1)}\big) - \gamma_k\, c_{\text{dec}}\, G_{k-1}^2 + \gamma_k\, c_{\text{grad}}\, G_{k-1}\big\|\Delta_k^{\text{is}}\big\| + \gamma_k\xi_k \tag{A.19}$$

其中 $G_k = \|\nabla_\theta \mathrm{LL}(\theta^{(k)})\|$，且 $\xi_k = \nabla\mathcal{V}_P(s^{(k-1)}), \Delta_k^{\text{mc}} + \frac{L_{\nabla\mathcal{V}_P}}{2}\gamma_k\|g_k\|^2$。级数 $\sum_k \gamma_k\xi_k$几乎必然收敛，其第一部分是具有平方可和增量的鞅，第二部分则由 (A.18) 得出。

断言 *(i)*。根据引理 11(i)，$\|\Delta_k^{\text{is}}\| \le \bar{c}_{\text{is}}/K$，因此 (A.19) 就是将 $L_S\delta_k$ 替换为 $\bar{c}_{\text{is}}/K$ 和 $c_2 = 0$ 后的 (A.15)。于是，由下界（4.7）的证明可得 $\liminf_k G_k \le c_{\text{grad}}\bar{c}_{\text{is}}/(c_{\text{dec}}K) = C/K$。

断言 *(ii)*。对于 $K_k \to \infty$ with $\sum_k \gamma_k/K_k < \infty$，偏差界给出

$$\big\|\Delta_k^{\text{is}}\big\| \le \frac{\bar{c}_{\text{is}}}{K_k} \xrightarrow[k\to\infty]{} 0, \qquad \sum_k \gamma_k\big\|\Delta_k^{\text{is}}\big\| \le \bar{c}_{\text{is}}\sum_k \frac{\gamma_k}{K_k} < \infty.$$

由于 G_k 在紧集 $\Theta_{\mathcal{S}_p}$上有界，对 (A.19) 求和可得 $\sum_k \gamma_k G_{k-1}^2 < \infty$，与命题 3 中一样。命题 4 的证明也可沿用，只需将 (A.9) 替换为

$$\big|R_m - R_n\big| \le L_R\big\|s^{(m)} - s^{(n)}\big\| \le L_R\Big(c\sum_{k=n+1}^{m}\gamma_k + \sup_{l>n}\Big\|\sum_{k=n+1}^{l}\gamma_k\Delta_k^{\text{mc}}\Big\|\Big)$$

其中 $c = \sup_{\mathcal{S}_p}\|\mathrm{F}_P\| + \bar{c}_{\text{is}}$，其最后一项在 $n \to \infty$时几乎必然趋于零，因为 $\sum_k \gamma_k\Delta_k^{\text{mc}}$收敛。因此 $G_k \to 0$。由于 ∇_θLL 连续且 $\Theta_{\mathcal{S}_p}$是紧集，$\theta^{(k)}$收敛到 $\Theta_{\mathcal{S}_p}$ 中 LL 的驻点集合；若该集合只有一个驻点，则收敛到 θ_{ML}。

断言 *(iii)*。将 $\mathrm{LL}(\theta_{\text{ML}})$ 加到 (A.19) 上，可得 (A.12)，其中 $e_k = \mathrm{LL}(\theta_{\text{ML}}) - \mathrm{LL}(\theta^{(k)})$，且$b_k = \|\Delta_k^{\text{is}}\|$、$c_2 = 0$ 以及上文中的噪声 ξ_k 均满足相应条件。根据 (A.18)，这些步长趋于零。通过在积分号下求导可知，LL 在 Θ 上二阶连续可微，因为 θ 阶导数 $p(x, z; \theta)$ 可由 z 的某个多项式与先验的高斯尾部之积控制，且该控制在 θ 的局部范围内一致成立。因此，引理 10 适用于 $f = \mathrm{LL}$ 和 $\Theta' = \Theta_{\mathcal{S}_p}$，偏差界给出

$$b = \frac{\bar{c}_{\text{is}}}{K}, \quad C_b = C_K \qquad \text{for fixed } K, \qquad\qquad b = 0, \quad C_b = 0 \qquad \text{for } K_k \to \infty.$$

□

A.8 残差评估

残差在每次迭代时计算。定理1的残差 $\|\nabla_\theta \mathrm{L}^\star\|$ 由包络恒等式（5.3）得到，其中 $\mathsf{q}^\star(\theta)$ 通过对 (μ_i, L_i) 进行自由优化获得；定理3的残差$\|\nabla_\theta$LL$\|$则由费舍尔恒等式（5.12）得到。二者都需要对个体参数取期望。对于高斯分布 $\mathrm{N}(\mu_i, L_iL_i^\top)$ 和函数 g，这些期望通过高斯-埃尔米特求积而非蒙特卡罗求和进行计算，

$$\text{Gauss-Hermite:} \quad \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\big[g(\mu_i + L_i\varepsilon)\big] \approx \sum_{j\in\{1,\dots,n\}^{n_z}} \omega_j\, g\big(\mu_i + L_i u_j\big),$$

$$\text{Monte Carlo:} \quad \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\big[g(\mu_i + L_i\varepsilon)\big] \approx \frac{1}{M}\sum_{m=1}^{M} g\big(\mu_i + L_i\varepsilon^{(m)}\big), \qquad \varepsilon^{(m)} \sim \mathcal{N}(0, I),$$

其中， $u_j = (u_{j_1}, \ldots, u_{j_{nz}})$ 和 $\omega_j = \omega_{j_1} \cdots \omega_{j_{nz}}$ 是由针对 $\mathrm{N}(0,1)$ 的 n 点高斯-埃尔米特求积公式构造的乘积求积公式的节点和权重。在后验分布下的期望使用相同的网格，但应用于扩展后的编码器分布 $\mathrm{N}(\mu_i, 1.6^2 L_i L_i^\top)$，并采用与（3.9）中相同的自归一化权重。

由于 z_i 仅为三维，网格包含 n^3 个点，即对于 ELBO 的 $n = 12$ 个节点为 1728 个点，而对于对数似然的 $n = 28$ 至 $n = 36$ 个节点则介于 21952 和 46656 个点之间。蒙特卡洛误差仅以 $M^{-1/2}$ [13, 30]的速率衰减，而每个维度使用 12 个节点时，与基于 2^{20} 个样本的蒙特卡洛平均结果在十位数上保持一致。因此，在这种情况下，高斯-埃尔米特求积以更低的成本实现了更高的精度。