分类号：TP277　　单位代码：10335
密　级：无　　学　　号：12032042

浙江大学

博士学位论文

中文论文题目：基于泛函优化的概率隐变量建模理论及软测量应用

英文论文题目：Functional-Optimized Latent Variable Models for Soft Sensing

申请人姓名：陈智超
指导教师：宋执环 教授
学科 (专业)：控制科学与工程
研究方向：数据驱动工业过程建模
所在学院：控制科学与工程学院

论文递交日期　二〇二五年六月

致谢

“岁月不居，时节如流”；六年前，我背着行囊初次踏入玉泉校区，在工控新楼报到参加夏令营的场景仍历历在目；五年前，当我拖着行李箱正式入学时，那份初来乍到的忐忑与期待仿佛仍在心头萦绕。回望这五年的研究生求学路，若没有家人、老师、同窗以及朋友们，我想我绝无可能独自走完这段充满挑战与收获的人生旅程。

首先，我要衷心感谢我的导师宋执环教授。宋老师治学严谨、学识渊博，每次交流都让我受益匪浅。他鼓励我独立思考，既给予探索的自由，又引导我凝练核心问题，常说：“要让人家一提起这个工作，就知道是你做的。”在论文写作中，他逐字逐句审阅，提出细致修改意见，极大提升了论文质量。此外，宋老师始终关心我的职业发展，支持我前往企业实习的选择。他的言传身教，让我终身受益。同时，我也要衷心感谢课题组的葛志强教授。在保研阶段，正是葛老师在夏令营时给予我宝贵的机会，让我有幸加入课题组。葛老师在科研上精益求精，对我的学术成长给予了很多指导，使我受益匪浅。

在本文完成的过程中，首先要向楼下教研室的王浩同学致以最诚挚的谢意。正是在他的悉心指导下，我完成了人生第一个神经网络的实现，这段经历不仅让我掌握了关键技术，更点燃了我对科研的热情。从细节打磨到思路的激烈碰撞——康拓诺维奇对偶、近端算子，每一次交流都让我在机器学习领域获得突破性成长；从投稿的准备到 rebuttal 的应对，每一个环节都凝聚着他的智慧与付出。特别感谢微软科学研究中心的刘畅研究员，在实习期间，他系统地指导我构建科研体系——从组织研究框架到代码撰写技巧，从生成模型的理论基础到连续性方程的数学推导，再到变分蒙特卡洛和采样算法求解 Schrödinger 方程的实践，每一个环节都让我受益匪浅。衷心感谢蚂蚁集团的丁雷雷工程师和黄健敏工程师, 在实习期间他们毫无保留地传授工程经验——从 Scala 链路的底层实现到高并发代码的优化，这些工程实践极大地提升了我的技术视野。特别感谢好未来集团的刘天乔工程师。从各向同性表示的理论探讨，到 VAE 的本质，再到 LLM 的基本原理，每一次与他在 KFC/线上会议的对话都让我获益匪浅。在求学道路上，药学院张昊天同学的独到见解让我深受启发，从“一滴墨水扩散”的桥过程比喻，到力学与对称性导论的深刻阐释，他为我打开了跨学科研究的新视角。计算机学院王方懿康同学对梯度流系统的精辟分析，让我对连续性方程、扩散模型的理解达到了新的高度，为我的研

究开辟了创新思路。感谢楼下教研室的戴清阳同学。在贝叶斯推断的理论框架构建、摊销变分推断、平均变分，戴清阳同学都给予了我积极的讨论。感谢楼下教研室的张建峰同学，在我学术生涯的起步阶段，正是他不厌其烦地为我拨开迷雾——在百忙之中抽出时间，耐心地为我解析期刊论文中的模型推导过程。衷心感谢中山大学何畅副教授和挪威科技大学于浩水助理教授对我首篇英文论文的悉心指导，让我获益匪浅。

感谢浙科钱金传老师，北航江肖禹副研究员，新国立孔祥印博士，石东邵伟明副教授，杭师姚乐副教授、沈冰冰副教授、朱哲人助理研究员、刘熠副教授、魏驰航副教授和曾九孙教授，湖师杨泽宇副教授，中南袁小锋教授，桂理朱其萍老师，港城马嘉泽助理教授，中大万一千教授、杨祖金教授、周贤太教授、芮泽宝教授、薛灿副教授，本校许超教授、刘兴高教授、宋春跃教授、谢磊教授、郜传厚教授、梁军教授、吴争光教授，湾大孙庆强助理教授，省职机陈家枢老师，浙工刘毅教授、冯宇教授等对本文的帮助。

衷心感谢我的舍友王泽林、张宇隆、邱弈臻，五年间从骨折到断片时的照顾，你们的帮助让我深深感念这份珍贵的情谊。感谢张新民研究员、周乐教授、范赛特、王静波、卓越、陈光捷、张鑫宇、郑东磊、余家鑫、廖思奋、张宏毅、李德阳、曾令权、何柏村、庄新镇、周睿、崔琳琳、刘颖、彭程、曹文彬、张小波、陈博戬等师兄师姐，同级的刘昭然、张驰野、翟瑞锟、欧阳静、何懿萌、乔佳宇、张敏、沈紫嫣、李乐清、陈雨薇、李丹宁，以及潘黎铖、仇诚、马奕然、陈国斐、朱辉翔、宋泽宇、顾敖广、叶宣宏、罗月阳、盛伟伟、张丙鑫、袁金松、张雨桐、程雨斌、郑晨、张雪睿、孙姝、魏诗怡、程雨斌、陈季宇、陈佳威、皮松岩、丁青、王永靖、叶晴艺等师弟师妹，感谢你们这些年来的一起奋斗。感谢“先进金学”群聊的——王磊、陈洁薇、谢锦豪、胡杨、刘俊杰、吕帅等同学，你们的陪伴与鼓励让这段学术之旅充满温暖与动力。特别感谢杜汶键、赵文深，不厌其烦地为我解答数学问题；盖文峰和邱子琪的持续鼓励、罗健彬博士的耐心倾听与专业指导，以及陆至彬、陈运全在科研上的交流，都让我受益匪浅。此外，特别感谢黄耀鹏师兄、李兴涛博士、游日敏博士、陈楚楚博士、刘静远同学、钟逸凡同学、张贺博士、李佳晖博士、方钲清博士、张栋豪博士等人的鼓励和支持。

最后，感谢我的父亲和母亲。五年博士路，只归家一次，每每念及，愧疚难当。电话那头总说”放心去闯”——你们永远是我最坚强的后盾，最温暖的港湾。

陈智超

2025 年 6 月于求是园

摘要

数据驱动的软测量技术是现代工业过程建模的关键工具，能够精准估计难以直接测量的关键变量，在降低测量成本和优化产品质量等方面发挥了重要作用。概率隐变量模型凭借其在建模测量噪声、过程动态特性和过程非线性特性中的卓越能力，已成为数据驱动软测量领域的核心方法。然而，伴随着深度学习技术的迅速发展以及工业大数据环境的日益复杂，概率隐变量模型的建模能力面临新的挑战与更高要求：在数据层面，工业过程逐渐向高精度与智能化方向发展，复杂工业流程中常需在极端操作条件下进行卡边操作，这使得工业传感器易发生故障，从而导致过程数据存在缺失和不完整性，概率隐变量模型需要具备高效的缺失数据补全和恢复能力以保证建模的可靠性。在模型层面，深度学习的引入显著增强了概率隐变量模型对系统非线性和时变特性的建模能力；然而，深度学习模型本身的不可逆性和无约束性也引入了隐变量推断的难题，并进一步提高了对训练方法灵活性的要求。

综上所述，如何在大数据和深度学习的背景下设计和开发高效的概率隐变量建模方法，提升其在工业应用场景中的性能和适应性，已经成为软测量领域亟需解决的关键问题。泛函优化作为一种兼具理论深度与应用广度的数学工具，凭借其强大的包容性、丰富的优化手段以及完善的收敛性理论，为解决上述隐变量建模中的核心问题提供了重要的理论支持和实践保障。在此基础上，本文以泛函优化框架作为研究方法，围绕概率隐变量建模理论与软测量应用，进行了下列方面的研究：

（1） 针对软测量数据预处理阶段中常见的缺失数据问题，本文从泛函优化的角度出发，分析了概率隐变量模型在直接补全缺失数据时可能引发的补全结果发散性和训练推理目标不一致性问题。设计了一种抑制发散的代价泛函，并将沃瑟斯坦空间引入为概率密度函数的优化空间。通过推导并求解该代价泛函，提出了一种全新的缺失数据补全流程，该流程能够在优化过程中有效抑制发散现象，并保证补全结果的稳定性和一致性。在此基础上，进一步证明了条件概率分布补全与联合概率分布补全在泛函优化框架下的等价性。实验结果表明，该方法在工业数据集上的补全精度显著优于相关方法。

（2） 隐变量模型的参数训练过程中，通常隐式地将隐变量分布限制在特定的归一化概率密度函数族（如高斯分布族）内，导致模型的表达能力和适应性受限。针对上述隐式约束问题，本文提出了一种基于“量子化”变分分布的隐变量推断方法，将隐变量分布表示为一组有限的“粒子”，并通过引入随时间变化的微扰过程，逐步改变这些粒子在空间中的位置，从而实现对任意实数集支撑分布的变分推断。在此基础上，结合最优控制理论，将隐变量分布推断问题转化为无穷时域最优控制问题，并通过最优控制的求解和实现设计了一种高效的隐变量分布推断算法和对应的参数学习算法。此外，还从理论上论证了所提出的参数学习算法的收敛性，并在实验层面验证了该方法在软测量建模中的适用性和优越性。

（3） 在隐变量模型的实际应用中，部分工业场景中隐变量的优化需要满足约束域条件（如图神经网络中的归一化邻接矩阵推断问题），前一小点所提出的推断方法在此类场景下往往难以有效适用。为系统解决这一问题，本文提出了一种基于镜像下降方法的隐变量推断框架。具体而言，深入分析了前一章所提出的方法失效的根本原因，即优化过程隐含依赖于欧几里得空间的假设。针对这一局限性，引入了布雷格曼散度重新定义隐变量推断过程，并基于镜像下降方法设计了一种适用于约束域场景的高效隐变量推断算法。在实际应用中，以图神经网络的归一化邻接矩阵推断任务为例，具体化提出的算法并严格证明了其收敛性。实验结果表明该方法在建模精度和优化效率上均显著优于传统方法，验证了其在软测量建模任务中的适用性和潜在价值。

（4） 针对动态工业过程中的隐变量建模需求，本文基于最优控制理论提出了一种全新的非线性动态隐变量推断与模型优化方法。通过随机微分方程和交替方向乘子法的视角，将非线性动态隐变量模型的损失函数重新表述为有限时域最优控制问题并系统剖析了最优控制结构中隐含的推断网络设计原理。在此框架下，进一步明确了推断网络设计与最优控制问题求解之间的内在联系，从理论上为非线性动态隐变量模型的开发提供了新的指导。在此基础上，引入了矩展开策略重新设计了适配深度学习后端的模型实现方法并证明了模型在训练过程中的收敛性。实验结果表明，所提方法在软测量建模任务中表现出优越的性能。

关键词：数据驱动建模 泛函优化 变分推断 概率隐变量模型 软测量建模

Abstract

Data-driven soft sensor technology enables the accurate estimation of key variables that cannot be measured directly. It plays a vital role in reducing measurement costs and optimizing product quality. Probabilistic latent variable models (PLVMs) have emerged as a core methodology in the soft sensing field, because of their capabilities in modeling measurement noise, dynamics, and nonlinearities. However, the rapid advancement of deep learning technologies and the increasing complexity of industrial big data environments present new challenges and higher demands for the modeling capabilities of PLVMs. At the data level, industrial processes are evolving towards higher precision and intelligence. Complex industrial operations often require pushing processes towards their operational limits under extreme conditions, increasing the susceptibility of sensor failure. This frequently results in missing and incomplete data, necessitating that PLVMs possess efficient capabilities for data imputation to ensure modeling reliability. At the model level, the incorporation of deep learning significantly enhances the capability of PLVMs to capture the nonlinear and time-varying characteristics of systems. However, the irreversibility and unconstraint of deep learning models also introduce new challenges for latent variable inference and demand greater flexibility in training methods.

Consequently, designing and developing efficient PLVMs within the context of big data and deep learning-methods that enhance performance and adaptability in industrial application scenarios has become a critical issue demanding resolution in the soft sensing domain. Functional optimization, a mathematical tool characterized by both theoretical depth and broad applicability, offers significant theoretical support and a practical foundation for addressing the aforementioned core problems in latent variable modeling. Its strengths lie in its strong inclusiveness, diverse optimization techniques, and well-established convergence theories. Based on this, this thesis adopts the functional optimization framework as its research methodology to investigate probabilistic latent variable modeling theory and its soft sensor applications, focusing on the following aspects:

(1) To tackle the prevalent issue of missing data during the pre-processing stage of soft sens-

ing, this thesis analyzes, from a functional optimization perspective, the potential for divergent imputation results and inconsistencies between training and inference objectives when directly applying PLVMs for data imputation. To mitigate these problems, a divergence-suppressing cost functional is designed, and the Wasserstein space is introduced as the optimization domain for probability density functions. Through the derivation and solution of this functional, a novel data imputation procedure is proposed, which suppresses divergence during optimization and ensures the stability and consistency of the imputed results. Furthermore, the equivalence between imputation based on conditional probability distributions and imputation based on joint probability distributions is established within the functional optimization framework. Experimental results across multiple industrial datasets demonstrate that this method achieves significantly superior imputation accuracy compared to prevalent approaches.

(2) During the parameter training of PLVMs, the latent variable distribution is often implicitly restricted to specific normalized probability density function families (e.g., Gaussian), thereby limiting the model's expressive power and adaptability. To address this problem, this thesis introduces a latent variable inference method based on a "quantized" variational distribution. This method represents the latent distribution as a finite set of "particles" and employs a time-varying perturbation to progressively adjust the spatial positions of these particles, enabling variational inference for distributions supported on arbitrary subsets of the real numbers. Building upon this, leveraging optimal control theory, the latent variable inference problem is reformulated as an infinite-horizon optimal control problem. By solving this control problem, an efficient latent variable inference algorithm and the corresponding parameter learning algorithm are designed. The convergence of the proposed parameter learning algorithm is theoretically proven, and experimental results validate the method's applicability and superiority in complex industrial scenarios.

(3) In the practical application of PLVMs, certain industrial scenarios necessitate that latent variable optimization adheres to specific domain constraints (e.g., inferring normalized adjacency matrices in graph neural networks). The inference method proposed in the pre-

vious section often struggles in such settings. To systematically address this challenge, this thesis proposes a latent variable inference framework based on the mirror descent method. Specifically, it first analyzes the root cause of the previous method's failure in constrained domains, identifying its implicit reliance on Euclidean geometry during the optimization process. To this end, Bregman divergence is introduced to redefine the latent variable inference process, leading to the design of an efficient inference algorithm suitable for constrained domains based on mirror descent. As a practical application example, the proposed algorithm is concretely applied to the task of inferring normalized adjacency matrices for graph neural networks, and its convergence is rigorously proven. Experimental results demonstrate that this method significantly outperforms traditional approaches in both modeling accuracy and optimization efficiency, verifying its broad applicability and potential value in complex industrial settings.

(4) Addressing the requirements for PLVMs modeling in dynamic industrial processes, this thesis proposes a novel method for nonlinear dynamic latent variable inference and model optimization grounded in optimal control theory. Adopting the perspective of stochastic differential equations and alternating direction method of multipliers, the loss function of the nonlinear dynamic latent variable model (NDPLVM) is reformulated as a finite-horizon optimal control problem. The underlying design principles for inference networks embedded within the optimal control structure are systematically analyzed. Within this framework, the connection between inference network design and the solution of the optimal control problem is elucidated, offering new theoretical guidance for the development of NDPLVMs. Subsequently, to facilitate implementation with deep learning backends, the moment expansion strategy is introduced to redesign the model realization, and the convergence of the model during training is proven. Experimental results indicate that the proposed method exhibits superior performance in soft sensing tasks.

Key Words: Data-Driven Modeling, Functional Optimization, Variational Inference, Probabilistic Latent Variable Models, Soft Sensor Modeling

主要符号对照表

缩写/符号	英文全称	中文全称
$\langle\cdot,\cdot\rangle$	inner product	内积
$\|\|\|_2$	L^2-norm/Euclidean norm	L^2-范数/欧几里得范数
$\|\|\|_{\mathscr{H}_{\mathrm{D}}}$	norm of the reproducing kernel Hilbert space $\mathscr{H}_{\mathrm{D}}$ defined on $\mathbb{R}^{\mathrm{D}}$	定义在 $\mathbb{R}^{\mathrm{D}}$ 空间上的再生核希尔伯特空间 $\mathscr{H}_{\mathrm{D}}$ 的范数
$:=$	definition symbol	定义符号
$\top$	transpose of a matrix	矩阵转置
∇	gradient in Euclidean space	欧氏空间中的梯度
$\nabla\cdot$	divergence in Euclidean space	欧氏空间中的散度
∂	symbol of partial differential	偏微分符号
$\Delta^{\mathrm{D}-1}$	D-dimensional simplex space	D-维单纯形空间
AdaGCL	adaptive graph contrastive learning	自适应图对比学习
ADMM	alternating direction method of multipliers	交替方向乘子法
$\mathcal{A}$	normalized adjacency matrix	归一化邻接矩阵
$\widehat{\mathcal{A}}$	adjacency matrix	邻接矩阵
$\mathscr{A}$	measurable set	可测集
$\arg\max/\arg\min$	the maximum/minimum value argument	目标函数的最大/最小值变量
AR-TCN	auto-regressive temporal convolution network	自回归-时间卷积网络
$\mathcal{B}$	batch size	批次大小
$\mathbb{B}[z\|\|z_\tau]$	Bregman divergence of point z wrt. z_τ	点 z 关于点 z_τ 的布雷格曼散度
$\mathcal{C}$	constant	公式中出现的常数

缩写/符号	英文全称	中文全称
$\mathscr{C}([0,\mathrm{T}],\mathbb{R}^{\mathrm{D}})$	continuous path space over $[0,\mathrm{T}]$ in $\mathbb{R}^{\mathrm{D}}$	从时间区间 $[0,\mathrm{T}]$ 到 $\mathbb{R}^{\mathrm{D}}$ 的连续路径空间
$\mathscr{C}([0,\infty),\mathbb{R}^{\mathrm{D_{LV}}})$	continuous path space over $[0,\infty)$ in $\mathbb{R}^{\mathrm{D_{LV}}}$	从时间区间 $[0,\infty)$ 到 $\mathbb{R}^{\mathrm{D_{LV}}}$ 的连续路径空间
CNN	convolution neural network	卷积神经网络
CSDI_T	conditional score-based diffusion models_tabular	基于条件得分的扩散模型-表格版
$\mathcal{D}$	dataset	数据集
$\mathcal{D}_{\text{train}}$	training dataset	训练集
$\mathcal{D}_{\text{test}}$	testing dataset	测试集
$\mathcal{D}_{\text{valid}}$	validation dataset	验证集
D	degree matrix	度矩阵
d	symbol of total differential	全微分符号
DA-LSTM	dual-attention long short term memory	双重注意力长短期记忆网络
DBPSFA	deep Bayesian probabilistic slow feature analysis	深度贝叶斯概率慢特征分析
DGDL	dynamic graph-based deep learning	动态图深度学习模型
DiGa	digamma function	双伽马函数
Dir	Dirichlet distribution	迪利克雷分布
DMVAER	dynamical mixture variational autoencoder regression	动态混合变分自编码器回归模型
DPMM	Dirichlet process mixture model	迪利克雷过程混合模型
DSM	denoise score matching	去噪得分匹配
$\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z)\|\|\mathcal{P}(z)]$	Kullback-Leibler divergence of distribution $\mathcal{Q}(z)$ wrt. $\mathcal{P}(z)$	分布 $\mathcal{Q}(z)$ 关于分布 $\mathcal{P}(z)$ 的库尔贝克-莱布勒散度
$\mathcal{E}$	epoch	迭代轮次

缩写/符号	英文全称	中文全称
$\mathbb{E}_{\mathcal{Q}(z)}[f(z)]$	expectation of function $f(z)$ wrt. distribution $\mathcal{Q}(z)$	关于分布 $\mathcal{Q}(z)$ 函数 $f(z)$ 的期望
E^2AG	entropy-regularized ensemble adaptive graph	熵正则集成自适应图
ER	entropy regularized	熵正则
$\boldsymbol{F}$	objective function	目标函数
$\mathbb{F}$	normalized function family	归一化函数族
$\mathscr{F}$	event set	事件集合
$\mathcal{F}[\mathcal{Q}(z)]$	funtional wrt. probability density function $\mathcal{Q}(z)$	概率密度函数 $\mathcal{Q}(z)$ 的泛函
$f_\theta(z)$	prior transition function	隐空间的先验转移函数
$g_\theta(z)$	emission function	隐空间的发射函数
GCN-RW	graph convolution network with random weight	随机权重的图卷积网络
GMVAE	Gaussian mixture variational autoencoder	高斯混合变分自编码器
$\mathscr{G}$	graph neural network operator	图神经网络算子
GNN	graph neural network	图神经网络
GPU	graph processing unit	图像处理单元
GRU	gated recurrent unit	门控循环单元
H_0	null hypothesis	零假设
H_1	alternative hypothesis	备择假设
$\mathcal{H}$	forecasting horizon	预测序列长度
$\mathscr{H}$	reproducing kernel Hilbert space	再生核希尔伯特空间
$\boldsymbol{H}$	Hamiltonian	哈密顿量
$\mathbb{H}[\mathcal{Q}(z)]$	entropy functional wrt. probability density function $\mathcal{Q}(z)$	概率密度函数 $\mathcal{Q}(z)$ 的熵泛函

缩写/符号	英文全称	中文全称
$\mathrm{HU}_{\mathrm{score}}$	hidden unit number of score function	得分函数的隐层单元数
$\mathbf{H.O.T.}$	higher order term	高阶项
$\mathcal{I}$	identity matrix	单位矩阵
$\mathbb{I}$	indicator function	指示函数
i.i.d	independent identically distribution	独立同分布
imp	imputation	补全
$\inf$	infimum	下确界
Informer	Informer	信息变压器网络
InfO	infinite-horizonal optimal control-based variational inference	基于无穷时域最优控制的变分推断
iTransformer	inverted-Transformer	转置变压器网络
$K(\cdot,\cdot)$	kernel function	核函数
KL divergence	Kullback-Leiber divergence	库尔贝克-莱布勒散度
KEMS	kernelized mirror descent-based iteration over simplex	基于核镜像下降的单纯形迭代
KnewImp	kernelized negative entropy-regularized Wasserstein distance-based imputation	基于核负熵正则化沃瑟斯坦距离的填补
k-NN	k-nearest neighbor	k-近邻
KSD	kernelized Stein discrepancy	核斯坦因差异度度量
L	volatility term of stochastic differential equation	随机微分方程的波动项
$\mathcal{L}$	loss function	神经网络的损失函数
LogTrans	LogTrans	局部变压器网络
LSPT-D	local similarity preserved transport for direct imputation	局部相似度保持补全
MakeDiag$(\cdot)$	make diagonal matrix function	构造对角矩阵的函数
MAE	mean absolute error	平均绝对值误差

缩写/符号	英文全称	中文全称
MAPE	mean absolute percentage error	平均绝对百分比误差
max/min	the maximum/minimum value	目标函数的最大/最小值
$\boldsymbol{M}$	mask matrix	遮罩矩阵
$\mathcal{M}$	convex set	凸集
$\mathscr{M}$	message passing function	信息传递函数
M	particle number	粒子数
MIRACLE	missing data imputation refinement and causal learning	缺失数据的输入、细化和因果学习模型
MIWAE	missing data importance-weighted autoencoder	缺失数据重要性加权自编码器
MLL	moment of log-likelihood	似然函数的矩
MPNN	message passing neural network	信息传递神经网络
MSE	mean squared error	均方误差
MUDVAE-SDVAE	modified unsupervised deep variational autoencoder-supervised deep variational autoencoder	改进的无监督变分自编码器-监督深度变分自编码器
$\mathcal{N}$	normal distribution	正态分布
$\mathscr{N}$	set of neighbor node	相邻节点集合
NaN	not a number	无效数字
NDPLVM	nonlinear dynamic probabilistic latent variable model	概率隐变量模型
NER	negative entropy regularization	负熵正则
NPLVR	nonlinear probabilistic latent variable regression	非线性概率隐变量回归
NDPLVM	nonlinear dynamical probabilistic latent variable model	非线性动态概率隐变量模型
ODE	ordinary differential equation	常微分方程

缩写/符号	英文全称	中文全称
p_{miss}	missing rate	缺失率
$p_\theta(\cdot)$	generative network with parameter set θ	以 θ 为参数的生成网络
$\mathcal{P}(z)$	prior distribution	先验分布
$\mathbb{P}(z)$	prior measure	先验测度
$\mathscr{P}_2(\mathrm{D})$	D-dimensional Wasserstein space	D-维沃瑟斯坦空间
PDE	partial differential equation	偏微分方程
PDTM	probabilistic discriminative time-series model	概率性判别时间序列模型
PLVM	probabilistic latent variable model	概率隐变量模型
$\mathcal{Q}(z)$	variational distribution	变分分布
$\mathbb{Q}(z)$	posterior measure	后验测度
$\mathscr{Q}$	spectral density matrix	谱密度矩阵
$q_\varphi(\cdot)$	inference network with parameter φ	以 φ 为参数的推断网络
$\mathbb{R}$	real number domain	实数域
R^2	coefficient of determination	决定系数
RBF	radial basis function	径向基函数
RKHS	reproducing kernel Hilbert space	再生核希尔伯特空间
RMSE	rooted mean squared error	均方根误差
RowSum$(\cdot)$	the row sum function	行和函数
SDGNN	symmetry-preserving dual stream graph neural networks	保持对称性的双流图神经网络
S-GCN	stacked graph convolution network	堆叠图神经网络
$\mathbb{S}(\mathcal{Q}(z),\mathcal{P}(z))$	kernelized Stein discrepancy of $\mathcal{Q}(z)$ wrt. $\mathcal{P}(z)$	分布 $\mathcal{Q}(z)$ 关于分布 $\mathcal{P}(z)$ 的核斯坦因差异度度量
Sink	Sinkhorn imputation	辛克宏补全
sup	supremum	上确界
St	Student's-t distribution	学生-t 分布

缩写/符号	英文全称	中文全称
s.t.	subject to	约束条件
T	number of iterations	迭代的总次数
$\mathscr{T}$	number of alternating iterations of the KnewImp algorithm	KnewImp 算法的交替迭代次数
$\mathcal{T}$	historical sequence length	历史序列长度
$\boldsymbol{T}(\cdot)$	transportation map	传输映射
tanh	hyperbolic tangent function	双曲正切函数
TDM	transformed distribution matching	转换分布对齐
TGLFA	two-stream graph convolutional network-incorporated latent feature analysis	融入图卷积网络的双流潜在特征分析
Transformer	Transformer	变压器网络
Trace	trace	迹
u	process variables	过程变量
$u(z)$	deterministic control policy	确定性控制策略
VAE	variational autoencoder	变分自编码器
(VB)EM	(variational Bayesian) expectation maximization	（变分贝叶斯）期望最大化
$\mathcal{V}$	von Mises statistic	冯·米塞斯统计量
$\mathcal{W}_p$	p-Wasserstein distance	p-沃瑟斯坦距离
x	data that contains process variables and quality variables	包含过程变量与质量变量的数据
$\boldsymbol{X}$	the dataset containing process variables and quality variables	包含过程变量与质量变量的数据集
y	label	质量变量
α	significance level	显著性水平
$\Gamma(\cdot)$	gamma function	伽马函数

缩写/符号	英文全称	中文全称
γ	positive constant	正值常数
δ	first variation	一阶变分
δ	Dirac measure/Dirac delta function	狄拉克测度/狄拉克德尔塔函数
ε	infinitesimal/discretization step	无穷小量/离散步长
η	learning rate	学习率
θ	parameter of generative network	生成网络参数
Θ	the union of the parameter of generative network and inference network	生成网络参数和推断网络参数的并集
Λ	eigen value of kernel function	核函数的特征值
λ	Lagrangian multiplier/adjoint state	拉格朗日乘子/伴随态
μ	mean of the normal distribution	正态分布的均值
$\nu(z)$	stochastic control policy	随机性控制策略
ν	degree of freedom	自由度
ρ	quadratic penalty coefficient	二次惩罚系数
Ξ	orthonormal bias	核函数的归一化正交基
Σ	covariance of the normal distribution	正态分布的协方差
τ	iteration index	优化轮次指标
ϕ	perturbation direction	微扰方向
φ	parameter of inference network	推断网络参数
$\Psi(\cdot)$	Bregman potential	布雷格曼势能
$\widehat{\psi}$	feature importance weight	特征重要性权重
$\psi(z)$	ansatz	拟设
Ω	support of probability distribution/sample space	概率分布的支撑/样本空间

目录

图目录

表目录

1 绪论

摘要： 概述了本文的研究背景与研究意义，系统梳理了工业软测量建模的研究内容，详细论述了隐变量模型的定义及其发展历程。通过回顾国内外基于隐变量模型的软测量建模研究现状和泛函优化技术的发展历史，从泛函优化的视角深入分析了当前软测量建模面临的问题与技术难点。在此基础上，阐述了本文的主要研究内容与创新点。

关键词： 工业过程建模 软测量 隐变量模型 泛函优化

1.1 研究背景及意义

在全球经济格局深刻变革与技术范式加速迭代的双重驱动下，工业智能化转型已成为国家竞争力的核心议题。作为国民经济发展的战略基石，工业体系的数字化与绿色化协同演进，不仅关乎产业升级的内生动力，更直接决定了国家在全球价值链中的位势。当前，主要工业化国家纷纷以差异化路径布局新一轮工业革命：美国通过《制造业战略计划 2024》[1]（Manufacturing USA 2024）强化人工智能与先进制造的融合，旨在重构全球产业链主导权；欧盟则以“工业 5.0”[2]为框架，强调人机协作与生产过程的可持续性；我国通过《工业领域碳达峰实施方案》等政策[3]，以人工技术创新为突破口，聚焦钢铁、水泥等高耗能行业的低碳工艺革新，从而利用自动化和智能化理论与技术，实现“碳达峰”和“碳中和”的“双碳”战略目标[4]。

然而，无论是推动美国的人工智能与先进制造的融合、欧盟强调的可持续生产，还是我国面向“双碳”目标的工艺优化，其核心都高度依赖对复杂工业装置内部状态和关键运行参数的实时、精准感知[5]。但是在实际生产环境中，由于多变量强耦合、非线性动态变化以及极端工况约束，许多决定操作参数的关键工艺变量无法通过直接传感器获取或者测量成本极其昂贵——如炼铁过程中的高炉炉温[6-7]、固定床反应器的产物气浓度[8]、精馏塔馏出物的杂质含量[9]以及聚烯烃的熔融指数[10-11]等。

为破解这一难题，“软测量”（soft sensor）技术应运而生。软测量是一种以数据驱动和模型分析为基础，通过可测变量间的关联信息，计算并预测不可直接测量的关键工艺质量指标或系统状态，从而实现间接测量的重要方法。作为工业智能化生产的核心支撑

技术之一，软测量不仅能够提升复杂工业过程的实时监测能力，还在优化工艺参数、节能降耗以及提升产品质量等方面展现了巨大潜力。因此，软测量已成为工业数字化转型及绿色制造领域的重要技术工具。

近年来，随着工业互联网（industrial internet）在工业过程中的广泛应用，工业过程积累的数据量达到了前所未有的高度。同时，先进的机器学习算法和基于图像处理单元（graph processing unit，GPU）的高性能计算设备的快速发展，使得基于数据驱动特别是以概率隐变量模型为代表的新型软测量建模方法展现出显著优势[12-14]。相比传统经验建模，这类方法具有易于维护、易于迭代、成本低廉等特点，正在逐渐成为主流研究方向[15]。这类数据驱动的软测量方法不仅解决了复杂非线性工业系统中的建模挑战，还为复杂系统的全生命周期管理提供了坚实的技术支撑[16]。通过精确提取高维数据中的隐含信息，这些方法能够在动态复杂的生产环境中高度适应，显著增强了工业过程的实时感知能力与预测能力。基于这种能力，软测量技术不仅实现了关键工艺变量的精准预测与反馈控制[17]，还有效促成了生产过程的节能降耗与质量提升，为智能化与绿色化相结合的新型工业生产方式提供了重要支撑。因此，推动软测量建模研究，不仅是对当前工业智能化需求的深度响应，也是实现“双碳”战略目标的技术必然路径。

1.2 软测量建模研究现状

1.2.1 软测量建模的基本原理

软测量的概念最早由 Brosilow 等人[18-20]提出，该方法首先选择一组易于测量且与关键变量/系统状态（本文称之为质量变量）高度相关的辅助变量（本文称之为过程变量），通过建立关键变量/系统状态与辅助变量之间的数学关系，实现对关键变量/系统状态的在线实时估计。具体而言，软测量建模技术通过测量易于获取的过程变量（通常为传感器可直接采集的变量），并构建过程变量量与质量变量之间的映射关系，从而实现对质量变量的实时预测。

如图1.1所示，软测量建模通常包括以下四个主要阶段[21]：

（1） **需求分析：**根据工业过程的实际需求以及基本机理分析，初步确定需要测量的辅助变量。这一阶段结合领域知识和工程经验，确保选取的辅助变量能够有效反映工业过程的动态行为。

（2） **数据采集：**完成需求分析后，部署传感器进行数据采集，并将采集到的数据存储到数据库中。针对采集的数据，采用统计分析和预处理方法剔除噪声与离群点、补全缺失值点，确保数据质量。这一阶段以高质量数据的获取为目标，支撑后续软测量建模工作。

（3） **软测量模型构建：**在完成数据准备后，通过统计分析、机理建模或数据驱动方法，构建辅助变量与目标变量之间的映射关系，并基于此模型对目标变量进行实时估计。这一阶段是软测量建模的核心研究内容，涉及特征选择、模型训练、验证与优化等关键步骤。

（4） **反馈决策：**将软测量模型的输出结果应用到工业生产过程中，用于实时监控工艺状态，并结合优化算法对生产过程进行动态调整和配置优化，提升工业系统的整体性能和效率。

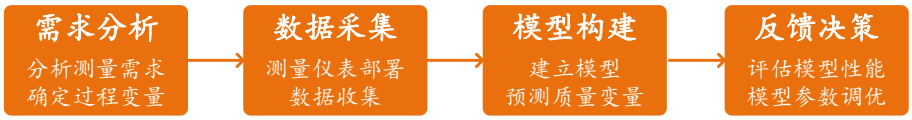


图 1.1 软测量建模的流程示意图

1.2.2 软测量模型的分类

根据软测量建模的实现原理，工业过程的软测量模型大致可以分为三类：基于机理建模的软测量模型、数据驱动的软测量模型，以及结合两者优势的混合软测量模型。这三类模型在数据依赖程度和机理需求上存在显著差异，并适用于不同的工业场景：

（1） **基于机理的软测量模型：**基于机理建模的软测量模型是最早应用于工业过程中关键变量估计与预测的一类方法，其核心思想是通过深入分析工业过程的运行机理，将过程中的物理和化学本质关系以数学方程的形式精准描述出来。在建模过程中，这类方法通常建立在完善的机理知识体系之上，结合物理学或化学定律推导出辅助变量与目标变量之间的明确关系，并通过 Aspen Plus[22]、gPROMS[23]、Simulink[24]、GAMS[25]等数值计算软件后端解算这些方程组从而实现对目标变量的实时预测。与其他建模方法相比，基于机理的软测量模型在工业应用中有着下列鲜明的优势：

- 机理模型能够直接反映工业过程的物理或化学特性，模型本身具备明确的物理意义，因此具有更强的可解释性。这种透明性使得模型能够在设计、优化和故障诊断等场景中发挥重要作用。
- 机理模型往往具有较高的普适性，只需对系统参数进行调整，就能够适用于相似的工业设备或工艺条件，实现模型的跨场景应用。

然而，基于机理建模的实现也存在诸多挑战：

- 此类方法对工业过程的机理知识需求极高，要求建模者对目标过程的内在化学反应、传递现象或力学行为有深入了解，且知识体系的构建往往需要大量的实验研究和领域积累。
- 对于规模庞大、结构复杂的现代工业过程，仅用物理和化学定律描述过程行为显得力不从心，建模和求解的难度显著上升。尤其是在多维非线性耦合场景下，机理模型的推导和参数优化极为复杂，可能导致计算负担明显增加。
- 大型工业系统由于物料特性、操作策略、外部扰动等多因素的影响，过程机理往往无法被全面掌握，这进一步限制了基于机理建模的适用范围。

尽管如此，在复杂性适中的工业场景中，基于机理建模的软测量方法依然展现出卓越的成效与价值。例如，Li 等人[26]基于具有数值求解能力的开源流程模拟平台，构建了化工储能过程中的模型预测控制算法，并在目标函数中显式引入了难以直接测量的电池退化因素，从而实现了针对储能电池全生命周期的优化控制方案，显著提高了储能系统的预测精度和运行可靠性。这些案例表明，在充分了解工业过程机理的前提下，基于机理建模的方法不仅能够提供高精度的预测结果，还能够通过对过程变量变化规律的深入解析，为生产过程的优化设计提供可靠的理论支撑。同时，也凸显了开放性建模工具和仿真平台在开发高精度软测量模型中的重要作用。

然而，随着工业系统逐步向规模化、大型化和智能化方向发展，工业过程日益呈现出复杂性、多样性和高动态特性，这使得基于机理的建模方法面临诸多挑战。在许多场景中，过程知识需要重新积累，例如电解槽工艺中膜的逐渐变薄这一动态过程[27]，可能引发复杂的非线性变化。此外，对于非线性强耦合的工业系

统，仅依赖单一的机理模型难以全面表征过程行为，导致建模精度显著下降，甚至可能无法有效开展建模工作。

（2） **基于数据的软测量模型：**数据驱动的软测量模型因不依赖于工业过程的机理知识和系统内部机制，通常被称为“黑箱模型”。这类模型的核心思想是通过建立端到端的数学映射关系，根据输入变量直接推测输出结果，而无需明确了解输入与输出之间的物理或化学机理。相比于基于机理建模的方法，数据驱动模型的开发能够显著缩短研究人员在机理知识学习上的时间成本，并且随着传统机器学习与深度学习技术的快速发展，数据驱动的软测量模型在精度、适应性和复杂场景下的表现上不断优化，能够满足现代工业中多样化的应用需求。具体而言，数据驱动的软测量模型可以进一步分为两大类[15]：基于统计的方法和基于深度学习的方法。

- 基于统计的方法主要依赖于传统的数学统计理论，利用历史数据挖掘输入变量与输出变量之间的相关性和映射关系。常见的统计学习方法包括偏最小二乘回归[28-29]（partial least square regression）、支持向量机[30]（support vector machine）等。这些方法在处理小规模数据集、线性或弱非线性问题时具有良好的性能。例如，支持向量机在解决小样本、高维数据的回归问题中表现出色，而随机森林[31-32]通过集成学习的方式能够有效提高模型的泛化能力。基于统计方法的优点在于其计算效率高、模型结构相对简单，并且在一定程度上具有可解释性。然而，对于强非线性、动态耦合的复杂工业过程，其建模能力可能受到限制。

- 随着计算能力的提升和工业大数据的广泛应用，深度学习方法在数据驱动建模中的应用愈发广泛。这些方法通过构建多层神经网络，能够从海量数据中自动提取复杂的非线性特征，并建立精确的输入输出映射关系。典型的深度学习模型包括堆叠自编码器[9,33-34]（stacked autoencoder）、卷积神经网络[35]（convolution neural network，CNN）、长短期记忆网络[36]（long short term memory，LSTM）以及变压器[37]（Transformer）网络等。例如，LSTM 在处理时间序列数据时表现优异[38]，能够捕捉工业过程中的动态行为；CNN 则擅长从空间数据中提取特征[39]，适用于捕捉局部时间窗口的特征并且加速工业时间序列数据的运算。深度学习模型的优势在于其强大的非线性拟合能力和

对复杂场景的适应性，但与此同时，其“黑箱”特性导致模型的可解释性较弱，且对数据量和计算资源的需求较高。

（3） **基于混合预测的软测量模型：**基于机理建模的方法以工业过程的物理、化学特性为基础，具有良好的可解释性和普适性，但对过程知识的掌握、经验积累和数学建模能力有较高要求。而数据驱动的建模方法则以工业数据为核心，通过端到端的映射关系进行变量预测，其建模效率高、迭代灵活，但依赖于大量高质量数据，并常受到泛化能力和物理约束的限制。为综合两者的优势，同时克服其各自的不足，研究人员提出了混合软测量模型方法，将数据驱动建模与机理建模相结合，通过引入过程机理知识对数据驱动方法进行辅助增强，从而既提升模型预测的精度，也增强了模型的可解释性。

混合软测量模型的核心思想通常是以数据驱动建模方法为主体框架，将机理知识以特定方式嵌入模型构建或训练过程。例如，过程机理可以通过物理约束或约束函数的形式加入模型的目标函数，也可以作为先验知识指导数据驱动模型的结构设计、参数优化或数据预处理。另一方面，机理建模的不足（如无法全面表征的复杂动态特性）也可以通过数据驱动模型进行补偿，提升整体建模性能。

在具体应用中，混合软测量模型已被广泛用于处理工业化流程中具有高非线性、强耦合特性的问题。例如，Puli 等人[40]提出了一种物理约束的概率慢特征分析方法，将物理约束（如线性等式和不等式）融入到动态特征提取中，显著提升了模型的解释性和预测精度，同时在工业案例中实现了硫含量预测误差的显著降低。Lou 等人[41]则提出了一种数据/机理混合驱动的建模与全局序列优化框架，通过结合灰狼优化算法，有效解决了高炉冶炼系统的动态非线性建模和多目标优化问题。这些研究表明，混合软测量模型不仅能够提升模型精度，还能保留机理建模的物理意义，从而为工业过程设计优化提供额外的信息支撑。

尽管混合软测量模型表现出许多优势，但其在实际应用中仍面临着一些挑战。首先，不同工业场景中过程机理的抽象层次和复杂性差异显著，这要求研究者对工业机理及其与数据驱动模型的融合方式有深入理解。其次，如何设计有效的融合框架将机理知识与数据驱动算法紧密结合仍是一个关键难题。例如，不同类型的工业过程可能需要选用特定的数据驱动方法（如神经网络、支持向量机等）或

不同的模型融合策略（如加权融合、模型嵌入或协同优化等）。此外，混合软测量模型的实现通常需要高性能计算资源，尤其是当机理模型和数据驱动模型都涉及复杂非线性运算时，计算效率也成为一个需要重点解决的瓶颈问题。

1.3 概率隐变量模型的定义、特性和发展历程

本文主要研究工作是如何利用泛函优化手段构建概率隐变量建模理论并探究其在软测量建模中的应用，因此本节将对概率隐变量模型的定义、特性和发展历程进行介绍。

1.3.1 概率隐变量模型的定义和特性

隐变量（latent variable），又称潜变量，指无法被直接观测或测量的变量。相对地，可以直接观测或测量的变量称为可观测变量（observable variable）或显变量。尽管隐变量本身不可直接观测，但它们通过影响显变量的表现间接参与了数据的形成过程。因此，隐变量的存在可以通过显变量的统计特性加以推断和量化[42]。换言之，隐变量可以被理解为驱动可观测变量变化的关键因素，它们蕴含了数据生成机制的重要信息。

概率隐变量模型[42-43]（probabilistic latent variable model）是一种强大的数据建模工具，其目标是将可观测数据与隐变量联系起来，并通过建模数据生成过程来揭示潜在结构。概率隐变量模型提出的核心思想是：假设真实数据的生成过程是由少量隐变量所主导的，而这些隐变量则通过某种方式影响了观测变量。从这种假设出发，概率隐变量模型通过构建显变量和隐变量之间的概率关系来描述数据的整体分布。通过这种方式，即便面对高维复杂数据，概率隐变量模型仍能够捕捉底层的生成机制，并以结构化的方式表达系统的复杂特性。相比直接对观测数据建模，概率隐变量模型采用了一种扩展思路，即通过引入隐变量将复杂观测数据分布利用易于处理的隐变量分布进行边缘化[44]（marginalize out），从而在处理复杂分布时显得更加灵活。

在工业过程建模领域，概率隐变量模型因其独特的数学特性，在处理典型工业数据时展现出显著优势。具体而言，面对多源传感器数据、设备日志和操作记录构成的高维非线性系统，这类模型通过以下机制实现有效建模：

- **降维与信息提取：**工业过程中的数据通常由成百上千个传感器采集，形成了高维、

复杂的数据空间。然而，这些高维数据往往包含大量冗余信息或噪声，直接对其建模会带来维度灾难和计算复杂性。概率隐变量模型的降维能力可以有效缓解这一问题。通过将高维观测数据映射到低维隐空间，概率隐变量模型能够提取数据的核心特征、去除冗余信息并同时保留对系统行为至关重要的关键变量。

- **生成性假设与工业数据特性契合：**概率隐变量模型的生成性假设认为观测数据是由少量隐变量通过特定的生成机制生成的。这种假设与工业过程数据的特性高度一致。在工业场景中，观测变量往往是冗余的，且由少数核心变量驱动。例如，精馏塔的运行状态可能由回流比、进料量、冷凝器温度和再沸器温度决定，而多个传感器的测量结果可能仅是这些核心变量的不同投影或冗余表示。概率隐变量模型的生成性假设能够自然地建模这种数据生成机制，并通过推断隐变量揭示工业系统中关键变量的作用机制。这种建模方式不仅能够提高模型的解释性[45-46]，还能为工业过程的优化和控制提供科学依据。

- **处理不确定性与噪声的能力：**工业数据常常受到测量噪声、数据缺失以及操作不确定性的影响[47]。这些不确定性会显著降低传统数据驱动建模方法的性能。而概率隐变量模型基于概率框架，能够自然地处理数据中的不确定性和噪声。通过将复杂的原始数据分布分解为隐变量空间上的一系列条件分布，概率隐变量模型可以对隐变量和观测数据的不确定性进行建模，从而在噪声环境下保持较高的鲁棒性。此外，概率隐变量模型的概率图还使其能够对缺失数据进行有效处理，从而提高模型的可靠性。

- **高效的建模与训练：**传统隐变量模型（如概率主成分分析[48]、因子分析[42,49]等）通常具有较为简单的训练过程，能够在较短的时间内完成模型构建。这种高效性使得隐变量模型非常适合作为工业数据建模的基模型，尤其是在需要快速响应和实时决策的场景中。尽管现代概率隐变量模型[50]（如变分自编码器[51]、生成对抗网络[52-53]等）引入了更复杂的非线性结构，但随着 GPU 计算能力的提升和以摊销变分推断[54-55]为代表的优化算法发展，其训练效率也得到了显著提升。这使得概率隐变量模型在不牺牲建模能力的情况下，满足工业过程对实时性和高效性要求。

1.3.2 概率隐变量模型的发展历程

1.3.2.1 传统概率隐变量模型

传统概率隐变量模型是概率隐变量建模的早期形式，其理论基础与多元统计模型高度重合。这类模型通过将隐变量表示为随机变量，并使用概率分布描述隐变量与观测变量之间的关系，从而为数据建模提供了一种灵活且具有解释性的框架。从隐变量的类型来看，根据隐变量支撑空间的区别，传统概率隐变量模型可以分为连续隐变量模型和离散隐变量模型两大类模型：

- **连续隐变量模型：**在连续隐变量模型中，隐变量分布的支撑（support）被假设为实数集，换言之隐变量被表示为连续数值，其目标是通过隐变量的引入对高维数据进行降维或特征提取。典型的连续隐变量模型包括因子分析[56]（factor analysis）、概率主成分分析[57]（probabilistic principal component analysis）、线性动态系统[58]（linear dynamical system）等。这些模型通过假设观测数据是由隐变量的线性组合加上噪声生成的，能够将高维观测数据映射到低维隐空间，从而提取数据的核心特征。例如，在工业数据分析中，概率主成分分析和因子分析由于对数据中的噪声和不确定性的处理能力，而被广泛地与即时学习[59-60]和流式变分[49]等自适应软测量技术相结合应用于关键质量指标的预报上。

- **离散隐变量模型：**在离散隐变量模型中，隐变量分布的支撑通常被假设为单纯形上的离散点集合。隐变量仅取离散值，常用于表示观测数据的类别结构或聚类关系。典型的离散隐变量模型包括高斯混合模型[43]（Gaussian mixture model）和隐马尔可夫模型[43]（hidden Markov model）。这些模型通过将复杂的原始数据分布分解为多个简单分布的组合，能够高效且直观地对处于多工况的工业操作场景进行建模。例如，高斯混合模型可以将工业生产过程中的多工况收集得到数据分解为多个具有不同特性的子过程[61-63]，从而为质量指标预测提供更为准确的预测结果。

1.3.2.2 深度概率隐变量模型

尽管传统概率隐变量模型因其理论成熟、结构简洁而在工业数据建模中得到了广泛应用，但随着工业过程复杂性的不断提升，这类模型逐渐暴露出诸多局限性。例如，现

代工业系统的数据往往具有高维非线性特性和动态特性，传统模型的线性假设（如主成分分析、因子分析）难以充分表达变量之间的复杂依赖关系。此外，传统模型在面对大规模、多模态数据（如时间序列、图像数据等）时表现力受到显著限制，进而难以满足对工业系统建模精度和鲁棒性的更高要求。因此，为应对复杂工业过程建模的挑战，研究者们开始尝试将深度学习框架与概率隐变量模型相结合，催生了以变分自编码器[64]（variational autoencoder）、动态变分自编码器[65]（dynamical variational autoencoder）、生成对抗网络[66]为代表的深度概率隐变量模型（deep probabilistic latent variable model）。

尽管深度网络结构的引入显著提升了模型的表达能力（expressiveness）和容量（capacity），但也伴随着训练过程中的新挑战：传统概率隐变量模型因其简洁的生成结构，可以采用以矩阵求逆为代表的数学运算对隐变量分布进行直接推断从而构建模型的训练流程。该过程通常依赖“变分分布”对隐变量分布进行近似，因此被称为变分推断[67-68]。然而，在深度学习框架下，由深度神经网络构成的生成函数通常不具备可逆性，这使得隐变量分布的推断和训练流程变得更加复杂，模型需要引入额外的网络并修改对应的变分推断算法以实现对隐变量分布的推断。例如，在变分自编码器类模型中，为解决这一问题，模型需要通过“摊销变分推断”[54,69]的策略，引入额外的推断网络以并将隐变量的分布限制在某个归一化概率密度函数族中，以构建模型的训练。而在生成对抗网络中，需要采用对抗训练（adversarial training）策略以估计隐变量的“密度比”[70]（density ratio）进而实现对隐变量分布的推断进而实现对模型的训练。总而言之，深度网络非可逆性的特性导致隐变量推断和训练流程的设计面临更高的要求。针对这一问题，不断有新的方法提出，这些努力不仅推动了深度概率隐变量模型的发展，也为复杂工业过程建模提供了更强大的工具和理论支持。

1.4 基于概率隐变量模型的软测量建模研究现状

本节对基于隐变量模型的软测量建模研究现状进行梳理，主要包含三部分：基于传统概率隐变量模型的软测量建模现状、基于深度概率隐变量模型的软测量建模现状以及概率隐变量模型在软测量建模的应用中有待解决的问题。

1.4.1 基于传统概率隐变量模型的软测量建模

以概率主成分分析为代表的概率隐变量模型最初主要应用于无监督学习任务。然而，Yu 等人[71]的研究首次将其引入到监督学习场景中，提出了监督概率主成分分析模型和半监督概率主成分分析模型，并在多标签和多类别分类任务中验证了其有效性。在此基础上，Ge 等人[72-74]结合 Ghahramani 和 Beal[75]提出的混合因子分析模型，将概率主成分分析扩展为混合概率主成分回归，并进一步提出了其监督版本和半监督版本。该方法成功应用于脱丁烷精馏塔过程的软测量建模中，显著提升了建模精度。针对工业过程测量数据中可能存在的离群值问题，Zhu 等人[76]将隐变量的先验分布从高斯分布族扩展至学生-*t* 分布族，提出了半监督鲁棒主成分回归模型。这一改进有效解决了过程测量数据中的离群点问题，并成功应用于田纳西-伊斯曼过程的建模。需要指出的是，上述模型通常基于线性假设来构造数据生成过程，并假定系统处于稳定的操作条件。然而，这种设定难以充分捕捉工业过程中的非线性特性和动态变化。为此，Ge 和 Chen[58]在 Ghahramani 和 Roweis[77]提出的动态线性系统参数估计框架的基础上，提出了监督动态线性系统回归模型，并在含有缺失数据场景下的脱丁烷精馏塔过程验证了其有效性。与此同时，Ma 和 Huang[78]基于切换动力系统理论[79]将切换线性动态系统模型应用于油砂过程的软测量建模，进一步提高了模型对动态工业过程的适应能力。

此外，通过引入自适应机制构建局部隐变量模型是克服传统概率隐变量模型局限性的一种有效途径。例如，Yuan 等人[59-60]将即时学习引入概率主成分回归框架和线性系统框架，并在脱丁烷精馏塔过程中验证了其方法的有效性。Yang 和 Ge[49]则将流式变分贝叶斯框架[80]应用于概率因子回归模型中，以实现对甲烷化炉出口气体浓度的软测量建模。值得注意的是，在基于隐变量的建模框架下通过引入随机非线性映射可以在数据生成结构中构造非线性特征从而提升模型处理复杂非线性过程的能力。例如，Yang 和 Ge[81]在非线性概率因子分析模型中引入随机权重和激活函数，构建了非线性因子回归的软测量模型。Zheng 等人[82]则将极限学习机引入因子分析框架，设计了对应的软测量模型和算法。

尽管传统概率隐变量模型在软测量建模领域已经取得了广泛的成功应用，但其仍然存在显著的局限性。具体而言，这类模型通常基于对数据生成过程的线性假设，而在应对实际工业过程中的复杂非线性行为时，这种假设往往难以充分刻画过程特征[14]。虽

然通过引入自适应机制、混合模型或随机特征提取方法等策略可以在一定程度上缓解这一问题，但这些改进方式通常伴随着显著增加的设计复杂度和计算成本。例如，在即时学习框架中，为了提升样本检索效率，往往需要引入 MapReduce 等技术作为后端支持[83-84]。此外，过于复杂的自适应机制如果设计不当，可能会对模型性能产生负面影响，从而进一步限制了传统概率隐变量模型在更复杂实际场景中的应用。

1.4.2 基于深度概率隐变量模型的软测量建模

正如1.4.1节所述，传统概率隐变量模型由于其线性假设，通常难以有效建模工业过程中复杂的非线性特性。为此，研究者们提出利用神经网络结构对隐变量的数据生成过程进行参数化[85]，从而构建深度概率隐变量模型。需要注意的是，诸如多层感知机、卷积神经网络、循环神经网络以及自注意力机制等常用的神经网络结构通常是不可逆的[86]，这导致了深度概率隐变量模型在训练过程中对隐变量分布的推断面临较大挑战。

为了解决这一问题，Kingma 和 Welling 在变分自编码器[87]中首次提出引入另一组神经网络，用于模拟最优的隐变量分布，并根据这一分布设计推断网络的函数结构。通过重参数化技巧，该方法实现了生成网络和推断网络的联合训练，从而完成深度概率隐变量模型的优化。这种变分推断技术也被称为摊销变分推断[54-55]。基于变分自编码器的思想，Shen 等人[13]引入堆叠自编码器中的逐层预训练与微调策略，利用多层感知机进行堆叠，提出了非线性概率隐变量回归模型及其半监督版本，并在一段炉实验中验证了所提出方法的有效性。Xie 等人[88]进一步扩展了这一框架，分别在有标签数据集和无标签数据集上训练了两个变分自编码器，并通过隐空间的相似性构建了软测量建模方法。此外，他们还针对缺失数据场景设计了数据补全流程，并在聚合物软测量建模中验证了其有效性。Chai 等人[66]采用了类似的思想，但通过对抗训练的方式对模型进行优化。在此基础上，他们进一步推导出了适用于缺失数据场景的损失函数，并在多相流数据建模中取得了良好的效果。需要指出的是，这一类变分自编码器模型的隐空间通常假设为高斯分布，这在应对多工况操作条件下的工业过程建模时存在一定的局限性。为了解决这一问题，崔琳琳等人[89]在变分自编码器框架中引入高斯混合模型，提出了相应的软测量建模方法。Guo 等人[90]进一步将高斯混合模型作为隐空间分布，并结合即时学习机制对变分自编码器进行改进，提出了新的软测量模型。然而，这类模型通常假设工业过程处于稳态操作条件，因此在动态特性显著的工业场景中表现有限。针对这一问题，Shen

和 Ge[91]将变分自编码器的摊销推断方法引入线性动态系统框架，并在后续研究中引入质量变量相关的过程变量加权机制[92]、金字塔结构[93]和周期-趋势分解策略[94]，以提升模型的建模效果。此外，通过在模型结构中引入循环神经网络中的记忆机制[36,95]可以有效突破线性动态系统中的马尔可夫假设，从而让软测量模型能够更好地描述过程动态特性。在这一思想的指导下，Lu 等人[96]引入门控循环单元（gated recurrent unit，GRU）到动态系统框架中，并在加氢裂化过程中验证了所提出方法的有效性。在此基础上，Jiang 等人[97]考虑到工业过程的平稳变换特性，将门控单元引入慢特征分析，提出了能够处理缺失数据的软测量建模方法，并在后续工作中引入了切换动力系统[98]以适应工业过程的多工况特性。Yao 等人[99]将门控循环单元引入有限混合模型，以应对工业过程中的多工况与动态性问题，并在甲烷化炉实验中证明了其方法的优越性。

尽管研究者通过在模型结构设计层面引入多种神经网络结构，开发了不同的深度概率隐变量模型，并在软测量建模任务中取得了显著进展，但仍然有以下几个方面的问题亟待解决：首先，虽然深度概率隐变量模型的框架延续了传统概率隐变量模型在缺失数据学习方面的优势，但当前的缺失数据补全流程通常是基于联合分布的极大似然设计的，而非直接面向条件概率分布的极大似然优化。这种方式在理论上和实践上均对缺失数据建模能力形成了一定局限，尤其在处理复杂非线性工业系统时，其效果可能难以达到预期。其次，根据“没有免费的午餐”定理[100]，即使复杂的深度模型结构显著提升了数据表示的灵活性，其模型训练仍然面临巨大挑战。这主要归因于深度神经网络不可逆性带来的隐变量分布的推断难度：虽然摊销变分推断通过神经网络参数化隐变量分布，为深度概率隐变量模型的训练提供了一条可行的解决路径，但这一方法出于梯度计算和损失函数优化的需求，通常需要将隐变量分布限制在易于归一化的概率密度函数族中。这样的假设虽然简化了优化和计算流程，但在很大程度上限制了模型对下游建模任务的适应性和性能提升。最后，在推断网络的设计上，摊销变分推断通常依赖于分析最优隐变量分布的函数结构，深度概率隐变量模型最初主要应用于无监督学习任务，在这一任务场景下，推断网络通常通过直接“反转”生成网络结构来构建[101]。然而，针对动态场景的软测量建模任务，推断网络的输入往往仅限于过程变量。这种设计可能与摊销变分推断的设计理念存在一定的差异，在一定程度上可能影响隐变量推断的准确性。因此，如何设计最优的推断网络结构以更好适配软测量建模任务，仍然是一个值得进一步进行研究的问题。

1.5 泛函优化的研究现状

1.5.1 概念溯源与核心应用

与传统的优化问题旨在寻找最优的数值或参数向量不同，泛函优化的目标是确定一个函数，满足一定的约束条件的前提下，使得某个依赖于该函数的整体性度量，即“泛函”（functional）达到最大值或最小值。这种在无限维函数空间中[102]的思想最早可追溯至 18 世纪的变分法[103]（calculus of variations）研究中。以欧拉（Euler）和拉格朗日（Lagrange）为代表的数学家，通过研究两点之间最短路径和最速降线（brachistochrone）等经典问题，建立了通过泛函极值确定最优函数的基本方法。这一思想在 20 世纪得到了显著拓展，其中两个关键理论进展对概率建模与动态系统优化产生了深远影响：

1. **概率分布的几何度量与守恒：**在经典力学中，刘维尔定理[104]（Liouville’s theorem）揭示了在哈密顿系统（Hamiltonian system）下，相空间（phase space）中概率密度随时间演化的守恒特性。这一定理为统计力学奠定了基础[105]，并确立了概率密度函数在动态系统中的归一化约束。与此同时，康托诺维奇（Kantorovich）在 20 世纪 40 年代开创的最优传输理论[106]（optimal transportation theory），通过最小化将一个概率分布“传输”到另一个分布所需的总成本，定义了沃瑟斯坦距离（Wasserstein distance）。这一度量能够有效捕捉概率分布的几何特征，提供了一种类似于欧几里得范数（Euclidean norm）的“距离”概念来量化不同概率分布之间的差异。

2. **动态系统的最优控制：**极大值原理[107]（maximum value principle）与动态规划[108]（dynamic programming）将早期变分法拓展至带约束的控制系统优化领域。这些理论为设计动态系统的最优控制策略提供了强大的数学框架，广泛应用于工程、经济等多个领域[109]。

综上所述，泛函优化提供了一套处理以函数为优化对象的强大分析框架，这一数学工具，特别是概率分布的几何度量（沃瑟斯坦距离）、概率演化与守恒定律（刘维尔定理）以及最优控制理论为理解和优化概率密度函数提供了关键工具，对于本文后续重构概率隐变量建模理论及其在软测量建模中的应用具有重要的指导意义。

1.5.2 有待通过泛函优化解决的问题

综合1.4.1和1.4.2节所述内容，尽管传统概率隐变量模型和深度概率隐变量模型在软测量建模领域已经取得了显著进展，但仍然存在不少亟待解决的挑战与问题。从预处理到模型构建的整个过程，概率隐变量模型在精确性与灵活性需求方面仍未完全满足工业软测量建模的实际应用需求。在此背景下，泛函优化理论提供了一个极具潜力的统一视角，有望为概率隐变量模型性能的提升构建更坚实的理论基础。需要注意的是，将泛函优化应用于隐变量驱动的软测量建模研究尚在起步阶段，许多关键问题尚未得到充分研究。具体而言，以下四个方面值得引入泛函优化技术进行进一步探索与分析：

（1） **数据预处理：**正如1.4.1和1.4.2节所述，现有概率隐变量模型在学习过程中将缺失数据视为隐变量，并通过最大化缺失数据和观测数据的联合似然进行推断。这种方法在实现缺失数据补全的同时，亦完成了对模型的参数优化。然而需要指出的是，缺失数据补全本质上是一个条件推断问题，因此所需极大化的似然项应为缺失数据相对于观测数据的条件分布而非联合分布。此外，当前概率隐变量模型驱动方法在进行缺失数据补全时关注的核心问题是缺失数据的“补全分布”是否对齐，而忽略了缺失数据“补全值”是否对齐。换言之，现有方法未能考虑补全值的精确性。因此，引入概率密度函数优化这一泛函优化技术重新审视概率隐变量模型在缺失数据补全过程中所采用的训练与补全策略并建立新的推断算法以促进补全流程对补全值的精确推断是一个具备重要研究意义的问题。

（2） **稳态概率隐变量模型：**正如1.4.2节所述，虽然引入深度神经网络结构显著增强了概率隐变量模型对数据生成过程的表达能力，但由于深度神经网络的不可逆性，额外引入推断网络对隐空间分布进行参数化成为必然选择。然而，为了便于计算模型的损失函数，推断网络的分布通常被限制在归一化概率密度函数族内（如高斯分布族），这种假设在降低计算复杂度的同时，显著限制了模型的灵活性，进而削弱了模型的表达能力。因此，如何应用以概率密度函数优化和最优控制理论为代表的泛函优化技术，从模型损失函数的构造出发，设计更具灵活性的隐变量推断策略以及模型训练策略并证明改进策略的收敛性是一个亟待研究的方向。

（3） **限制域的概率隐变量模型：**在稳态模型的研究基础之上，还需进一步扩展概率隐变量模型在复杂深度学习结构的适应性。具体而言，现有的大多数概率隐变量模型假设隐变量分布支撑于全实数空间（例如高斯分布的支撑为全实数空间），但这一假设在实际工业应用中并不总是成立，尤其是当隐变量受限于特定的几何或约束条件时。例如，在图结构数据、变压器网络或高斯混合模型中，隐变量的可能取值通常需要受到单纯形域的约束。如果直接引入基于推断网络参数化的限制域支撑分布，会带来复杂的正则项计算问题。因此，如何在保证隐变量推断灵活性的前提下，引入约束优化策略对隐变量推断过程进行重构，进而建立相应的模型训练流程，并证明改进后流程在理论上的收敛性，是一个亟待研究的重要方向。

（4） **动态概率隐变量模型：**在稳态模型和限制域模型的研究基础上，还需要进一步扩展概率隐变量模型在动态过程数据建模中的适应性。具体而言，动态数据建模面临两个关键挑战：一是不同时间点的观测数据之间的耦合关系，二是不同时刻的隐变量之间的动态相关性。因此，隐变量推断网络的设计不仅需要充分考虑不同时刻的观测数据，还需建模隐变量在时间维度上的依赖性与耦合特性。在此背景下，如何借助概率密度函数优化、最优控制理论等泛函优化技术，为动态概率隐变量模型的结构设计确立理论准则，并进一步推导适用于动态过程建模的新型概率隐变量模型及其实现方式，是一个极具研究价值的重要方向。

因此，本文将在接下来的章节中从泛函优化的角度针对这些问题进行系统性的探讨，并提出相关理论分析和对应的解决方案。

1.6 本文的研究内容与创新点

1.6.1 各章节研究内容

针对1.5.2节提出的问题，本文将深入研究如何利用泛函优化方法对概率隐变量模型在软测量建模中的数据预处理和模型构建环节进行系统性重构以提升其性能。所提出的理论框架具有广泛的一般性和统一性，不仅能够将分析结果与研究结论推广至更通用的概率隐变量建模方法，还为新的隐变量模型构建方法及其软测量模型的开发提供了重要启发。图1.2概述了本文的整体研究框架与技术路线。其中，橙色框标识了本文旨在解决

的关键科学问题，即1.5.2节提出的概率隐变量模型驱动软测量建模中面临的若干挑战；黑色框（对应第3章至第6章）列出了针对各项挑战所提出的具体解决方法或模型；浅蓝色框则展示了支撑这些方法和模型的核心理论基础。基于该框架，下文将对各章节的研究内容进行详细阐释。

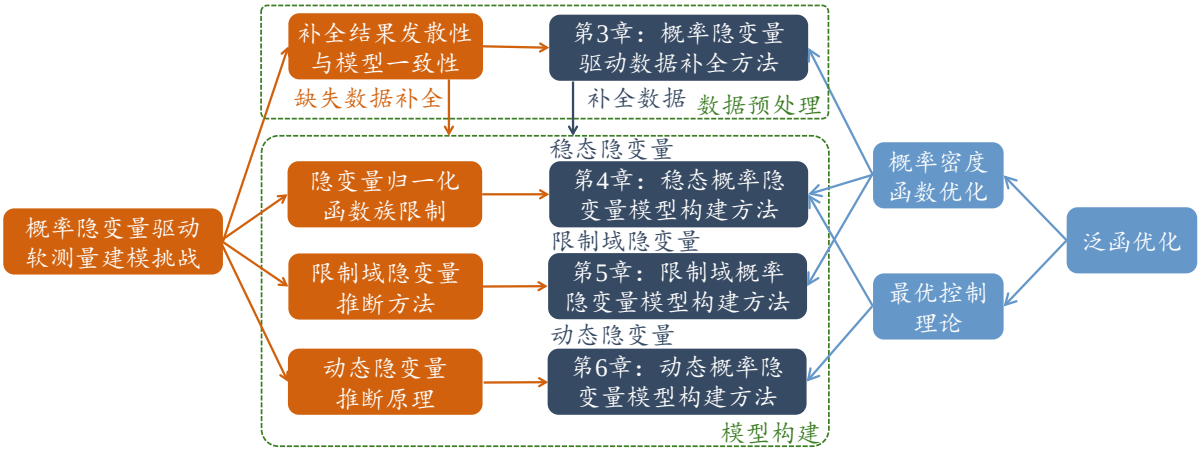


图 1.2 本文各章研究内容和利用泛函优化的方式及其所解决的软测量建模挑战示意图

本文研究内容主要围绕概率隐变量模型在工业软测量建模中的两个关键阶段展开：数据预处理（聚焦于缺失数据问题）和模型构建（应对隐变量归一化函数族限制、限制域支撑的推断问题和动态隐变量推断问题）。数据预处理为隐变量模型的有效训练奠定了基础，而模型构建则是实现精确软测量建模的核心环节。该框架亦揭示了各章节研究所依托的核心理论基础——概率密度函数优化与最优控制理论，两者共同构成了泛函优化方法论在此背景下的具体应用。具体而言，第3章针对工业过程数据中的缺失值补全问题，提出了一种结合在沃瑟斯坦空间中进行概率密度函数优化的新算法，旨在解决现有基于概率隐变量模型的数据补全方法所导致的补全结果发散及训练-推理目标不一致性问题。在得到补全数据的前提下，第4章聚焦于建模过程中隐变量分布的隐式约束挑战，通过将隐变量分布“量子化”为可优化的有限粒子集并结合空间泛函优化与无穷时域最优控制理论，构建了一种新的隐变量推断方案。针对隐变量需满足特定约束域支撑的推断场景，第5章深入分析了第4章方法框架在此情况下的局限性，并在此基础上引入布雷格曼散度，以图神经网络的归一化邻接矩阵推断为研究案例，设计了一种基于镜像下降的隐变量推断方法，以处理支撑为限制域的隐变量推断问题。在稳态概率隐变量模型和限制域概率隐变量模型的基础上，第6章面向动态隐变量模型构建问题，将概率隐变量模型的推断问题重新表述为有限时域最优控制问题，并运用及交替方向乘子法

（alternating direction multiplier method，ADMM）和随机微分方程理论为动态非线性概率隐变量模型的推断网络设计与深度学习后端实现提供了理论指导。

1.6.2 各章节创新点介绍

本文共包含七章，其中第1章为绪论部分，介绍了软测量建模的问题和需求，梳理了概率隐变量模型的发展历程，以及现有基于概率隐变量模型的软测量建模研究的现状及存在的问题，同时概括了本文的主要研究内容和贡献。第2章为预备知识，提供了理解后续章节推导所需的关键理论，包括概率隐变量模型的相关算法、泛函优化理论以及本文验证算法有效性和优越性所采用的数据集。第7章则对全文的研究工作和贡献进行了总结，并探讨了未来可能的研究方向。

在这些基础性章节的铺垫之上，第3章至第6章构成了论文的主要研究内容。以下将结合第3章至第6章的研究内容，详细阐述本文的创新点：

- **第一部分（第3章）：**从泛函优化的角度，剖析了利用概率隐变量模型直接补全缺失数据可能引发的结果发散与训练-推理目标不一致问题。在此基础上，通过设计一种抑制发散的代价泛函并引入沃瑟斯坦空间作为概率密度函数的优化空间推导出了一种全新的缺失数据补全流程。此外，在泛函优化框架下进一步证明了条件概率分布补全与联合概率分布补全诱导的优化流程的等价性。最后，综合上述理论洞见，提出了一种基于概率隐变量模型的全新数据补全算法并在工业数据集上对所提出算法的有效性进行了实验验证。

- **第二部分（第4章）：**聚焦于隐变量模型参数训练过程中存在的隐变量隐式约束问题，即为了提升训练过程在计算机实现时的便利性，通常将隐变量限制于特定的概率密度函数族（如高斯分布族）。为解决此问题，本文提出将隐变量分布进行“量子化”处理，将隐变量分布表示为一组有限的“粒子”，并通过引入时间变化的微扰过程，逐步优化这些粒子在空间中所处的位置，从而实现对任意实数集支撑分布的有效逼近。在此基础上，本文进一步结合概率隐变量模型的目标泛函并引入最优控制方法，将隐变量分布的推断问题转化为无穷时域最优控制策略的求解问题。通过求解无穷时域最优控制问题，设计了一种新的隐变量分布推断方法以及相应的概率隐变量模型参数学习算法，并从理论上证明了所提出概率隐变量

模型参数学习算法的收敛性。在最后部分对所提出的隐变量推断算法和概率隐变量模型参数学习算法的有效性和实用性进行了实验验证。

- **第三部分（第5章）：**在第4章的基础上，进一步深入研究了支撑域为约束域情形下的稳态概率隐变量模型（如图神经网络和变压器网络）训练问题，并提出了一种基于镜像下降方法的全新概率隐变量模型训练框架。具体而言，针对第4章框架在隐变量处于约束域场景下失效的根本原因（即迭代过程对欧几里得空间的隐式依赖性）进行了详细分析。为克服该局限性，引入布雷格曼散度对迭代过程进行修改，并基于镜像下降方法设计了一种适用于约束域场景的隐变量推断算法。随后，以图神经网络的图结构推断任务为应用案例，将该方法具体化并从理论上证明了算法的收敛性。此外，从多个层面开展了实验以验证所提方法的有效性和优越性。

- **第四部分（第6章）：**探索了运用最优控制作为泛函优化手段对非线性动态概率隐变量模型的隐空间推断和模型在深度学习后端的实现问题展开了系统性研究。研究工作从随机微分方程的视角出发，将隐变量模型的损失函数重新表述为有限时域最优控制问题，并利用最优控制理论和 ADMM 导出了动态概率隐变量模型中的推断网络设计原理。这一理论框架明确了推断网络设计原则与最优控制问题解之间的内在联系，为推断网络的开发提供了新的理论指导。为满足软测量建模对模型精度的实际需求，研究中引入了矩展开策略，并据此重新设计了适配深度学习后端的实现方法。此外，还对所提出方法的收敛性进行了证明；并从建模精度、消融分析、敏感性分析和收敛性分析这四个层面的实验对所提方法的有效性与优越性进行了验证。

1.7 本章小结

本章首先介绍了研究背景和研究意义，系统梳理了软测量的基本原理和模型分类。随后，本章介绍了概率隐变量模型的定义、特性和发展历程。在此基础上，本章阐述了概率隐变量模型在软测量领域的研究现状和泛函优化方法的溯源与核心应用，分析了当前研究中需要引入泛函优化技术进行解决的问题，并引出了本文的主要研究内容。在上述论述和分析的基础上，本章概述了各章节的具体研究内容与主要创新点。

2 预备知识

摘要： 系统梳理了支撑全文研究的核心理论基础、评价指标与数据集，围绕概率隐变量模型、泛函优化方法和工业数据集特性三个层面展开。在概率隐变量模型的部分，本章阐明了概率隐变量模型的基本框架与训练思想；在泛函优化的介绍中，涵盖了概率密度函数优化的模型假设空间及最优控制问题和极大值原理的关键内容；针对实验所用工业数据集，介绍了数据集采集的工业背景和变量构成。通过对概率建模、泛函优化策略及数据集背景的介绍，为后续章节的创新方法设计与验证奠定了理论基础和实践基石。

关键词： 预备知识 概率隐变量模型 泛函优化 软测量数据集

2.1 概率隐变量模型

2.1.1 概率隐变量模型的构建框架及其学习算法

工业过程常见的概率隐变量模型结构可以抽象为图2.1所示的结构。将工业过程观测数据的集合记为 $\mathcal{D}=\{(u_i,y_i)|u_i\in\mathbb{R}^{\mathrm{D_{PV}}\times 1},y_i\in\mathbb{R}^{\mathrm{D_{QV}}\times 1},i=1,2,\ldots,\mathrm{N}\}$，$u_i$ 为工业过程中易于测量的过程变量（process variable），其中 PV 是“process variable”的缩写，y_i 为工业过程中难以测量的质量变量（quality variable），其中 QV 是“quality variable”的缩写。定义观测数据 $x_i:=[u_i,y_i]\in\mathbb{R}^{\mathrm{D}\times 1}$，其中 $\mathrm{D}=\mathrm{D_{PV}}+\mathrm{D_{QV}}$。

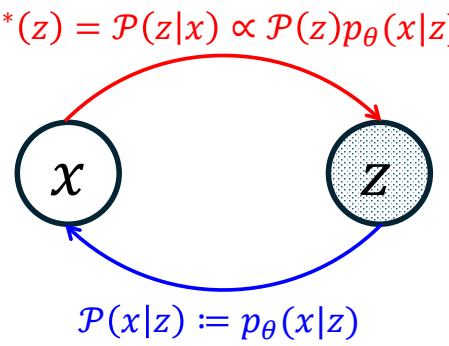


图 2.1 概率隐变量模型结构和其学习算法示意图

需要指出的是由于工业过程的强非线性性和动态特性，x 的分布 $\mathcal{P}(x)$ 可能呈现出多峰（multi-modal）和动态（dynamic）特性，因此隐变量模型的建模核心是引入一组服

从简单分布 $\mathcal{P}(z):\Omega\to\mathbb{R}^+\cup\{0\}$ 的变量 $z\in\mathbb{R}^{\mathrm{D_{LV}}}$（如图2.1蓝色阴影部分所示），其中 Ω 为分布 $\mathcal{P}(z)$ 的支撑（support），LV 是“latent variable”的缩写，将复杂的分布 $\mathcal{P}(x)$ 通过下式进行边缘化：

$$\mathcal{P}(x)=\int_{\Omega}\mathcal{P}(x|z)\mathcal{P}(z)\,\mathrm{d}z, \tag{2-1}$$

进而对简单的条件分布 $\mathcal{P}(x|z)$ 进行学习，从而简化计算流程。因此，考虑引入以 θ 为参数的函数（如神经网络等）$p_\theta(x|z)$。根据极大似然原理，可以引入如下的优化命题进行模型的训练：

$$\arg\max_\theta\ \log\int_{\Omega}p_\theta(x|z)\mathcal{P}(z)\,\mathrm{d}z, \tag{2-2}$$

由于上式存在高维概率积分难以进行计算，因此考虑引入变分分布 $\mathcal{Q}(z):\Omega\to\mathbb{R}^+\cup\{0\}$ 对该目标函数进行重构：

$$\begin{aligned}
&\arg\max_\theta\ \log\int_{\Omega}p_\theta(x|z)\mathcal{P}(z)\,\mathrm{d}z\\
\Rightarrow&\arg\max_\theta\ \log\int_{\Omega}p_\theta(x|z)\frac{\mathcal{P}(z)}{\mathcal{Q}(z)}\mathcal{Q}(z)\,\mathrm{d}z\\
\overset{\text{(i)}}{\Rightarrow}&\arg\max_{\theta,\mathcal{Q}(z)}\ \int_{\Omega}\mathcal{Q}(z)[\log p_\theta(x|z)\frac{\mathcal{P}(z)}{\mathcal{Q}(z)}]\,\mathrm{d}z\\
\Rightarrow&\arg\max_{\theta,\mathcal{Q}(z)}\ \mathbb{E}_{\mathcal{Q}(z)}[\log\frac{p_\theta(x|z)\mathcal{P}(z)}{\mathcal{Q}(z)}],
\end{aligned} \tag{2-3}$$

其中步骤“(i)”是基于如下所示的琴生不等式[110]（Jensen’s inequality）：

$$\log\int_{\Omega}\mathcal{Q}(z)f(z)\mathrm{d}z\geq\int_{\Omega}\mathcal{Q}(z)\log f(z)\,\mathrm{d}z, \tag{2-4}$$

而 $\mathbb{E}_{\mathcal{Q}(z)}[\log\frac{p_\theta(x|z)\mathcal{P}(z)}{\mathcal{Q}(z)}]$ 通常被称为证据下界（evidence lower bound，ELBO）。通过不断地提升 ELBO，最终可以使得模型得到优化。因此，为了对 ELBO 进行有效的“提升”从而实现对隐变量模型的参数 θ 的训练，通常需要先进行变分分布 $\mathcal{Q}(z)$ 的优化。为了达到这个目标，ELBO 可以重构为如下的形式：

$$\begin{aligned}
&\arg\max_{\mathcal{Q}(z)}\ \mathbb{E}_{\mathcal{Q}(z)}[\log\frac{p_\theta(x|z)\mathcal{P}(z)}{\mathcal{Q}(z)}]\\
\Rightarrow&\arg\max_{\mathcal{Q}(z)}\ \mathbb{E}_{\mathcal{Q}(z)}[\log\frac{\mathcal{P}(z|x)}{\mathcal{Q}(z)}]\\
\Rightarrow&\arg\min_{\mathcal{Q}(z)}\ \underbrace{\mathbb{E}_{\mathcal{Q}(z)}[\log\frac{\mathcal{Q}(z)}{\mathcal{P}(z|x)}]}_{:=\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z)\|\mathcal{P}(z|x)]}\\
\Rightarrow&\arg\min_{\mathcal{Q}(z)}\ \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z)\|\mathcal{P}(z|x)],
\end{aligned} \tag{2-5}$$

其中 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(z)\|\mathcal{P}(z|x)\right]$ 为分布 $\mathcal{Q}(z)$ 和分布 $\mathcal{P}(z|x)$ 之间的库尔贝克-莱布勒散度（Kullback-Leibler divergence，KL 散度）。由于 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(z)\|\mathcal{P}(z|x)\right] \geq 0$ 恒成立[111]，并且当且仅当在 $\mathcal{Q}(z) \propto \mathcal{P}(z|x)$ 等号成立，其中 $\mathcal{P}(z|x)$ 又满足关系 $\propto p_\theta(x|z)\mathcal{P}(z)$。因此，如图2.1红色箭头部分所示，一个最优的变分分布 $\mathcal{Q}^*(z)$ 应当满足下式：

$$\mathcal{Q}^*(z) \propto \mathcal{P}(z|x), \tag{2-6}$$

或者下式：

$$\mathcal{Q}^*(z) \propto p_\theta(x|z)\mathcal{P}(z). \tag{2-7}$$

可以看出，一个最优的变分分布 $\mathcal{Q}^*(z)$ 是函数 $p_\theta(x|z)$ 的“反函数”（reverse function），对 $\mathcal{Q}(z)$ 的优化可以视作将函数 $p_\theta(x|z)$ 进行反函数求解的过程。由于这个过程是对函数进行求解，属于变分问题[103]，因此 $\mathcal{Q}(z)$ 被称为变分分布（variational distribution）而对 $\mathcal{Q}^*(z)$ 的优化过程被称为变分推断[67-68]（variational inference）而根据 $\mathcal{Q}^*(z)$，即可利用式(2-3)对 θ 进行优化。在这个基础上，通过交替地优化 $\mathcal{Q}(z)$ 和 θ 即可对 ELBO 进行提升，从而达到最大似然的效果。根据上述思想，可以总结如算法2.1所示的概率隐变量模型学习变分贝叶斯期望-最大化算法（variational Bayesian expectation-maximization，VBEM），这一过程如图2.1中的蓝色箭头和红色箭头所示。

算法 2.1 概率隐变量学习的 VBEM 算法伪代码

输入： 数据：$\{x_i|i=1,...,\mathrm{N}, x_i \in \mathbb{R}^{\mathrm{D_{LV}}}\}$，条件概率模型 $p_\theta(x|z)$，模型参数 θ，以及先验分布 $\mathcal{P}(z)$，优化迭代次数迭代次数 T，收敛阈值 ϵ。

输出： 优化后的模型参数 θ^*。

$\mathcal{L}_0 \leftarrow +\infty$
for $\tau \leftarrow 0$ to $\mathrm{T}-1$ **do**
  $\mathcal{Q}_{\tau+1}(z) \leftarrow \dfrac{p_\theta(x|z)\mathcal{P}(z)}{\mathcal{P}(x)}|_{\theta=\theta_\tau}$
  $\theta_{\tau+1} \leftarrow \arg\max_{\theta,\mathcal{Q}(z)} \mathbb{E}_{\mathcal{Q}(z)}\left[\log \frac{p_\theta(x|z)\mathcal{P}(z)}{\mathcal{Q}(z)}\right]|_{\mathcal{Q}(z)=\mathcal{Q}_{\tau+1}(z)}$
  $\mathcal{L}_{\tau+1} \leftarrow \mathbb{E}_{\mathcal{Q}(z)}\left[\log \frac{p_\theta(x|z)\mathcal{P}(z)}{\mathcal{Q}(z)}\right]|_{\theta=\theta_{\tau+1},\mathcal{Q}(z)=\mathcal{Q}_{\tau+1}(z)}$
  $\theta^* \leftarrow \theta_{\tau+1}$
  if $\mathcal{L}_{\tau+1}-\mathcal{L}_\tau \leq \epsilon$ **then**
    打断循环
  end if
end for
return θ^*

2.1.2 深度学习时代的概率隐变量模型

如图2.2上半部所示，在深度学习时代，为了增强数据的表示能力研究者通常会引入神经网络来对条件分布 $p_\theta(x|z)$ 进行建模。然而，这种建模策略引发了一个严峻的问题——深度神经网络的可逆性通常难以保证，这导致了对函数 $\mathcal{Q}(z)$ 的优化变得困难。尽管如此，根据式(2-6)的分析，可以发现最优的 $\mathcal{Q}^*(z)$ 的函数形式实际上是某个输入为 x 的函数。根据这一思想，Kingma 和 Welling[51]引入了一个参数为 φ 的神经网络 $q_\varphi(z|x)$，用于近似最优分布 $\mathcal{Q}^*(z)$，如图2.2下半部所示。在假设 $q_\varphi(z|x)$ 能模拟 $\mathcal{Q}^*(z)$ 的前提下，即 $\mathcal{P}(z|x)$ 可以用 $q_\varphi(z|x)$ 进行参数化的前提下，式(2-3)所定义的 ELBO 可以重新表述为以下优化问题：

$$\begin{aligned} &\arg\max_{\theta,\varphi}\ \mathbb{E}_{q_\varphi(z|x)}[\log\frac{p_\theta(x|z)\mathcal{P}(z)}{q_\varphi(z|x)}] \\ \Rightarrow &\arg\max_{\theta,\varphi}\ \mathbb{E}_{q_\varphi(z|x)}\left[\log p_\theta(x|z)\right]-\mathbb{D}_{\mathrm{KL}}\left[q_\varphi(z|x)\|\mathcal{P}(z)\right]. \end{aligned} \tag{2-8}$$

通过同时优化参数 θ（即优化 $p_\theta(z|x)$）和 φ（即优化 $\mathcal{Q}(z)$），可以有效提升 ELBO。这样一类模型被称为变分自编码器（variational autoencoder，VAE），而这种变分推断技术被称为“摊销变分推断”[54-55]（amortized variational inference）。

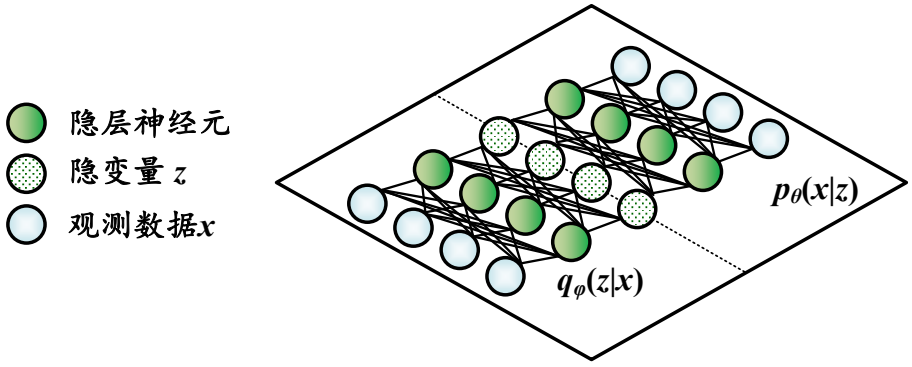


图 2.2 VAE 的模型结构示意图

在这个基础上，算法2.1可以重构为算法2.2所示的流程对 θ 和 φ 进行同步更新，从而实现对 VAE 模型的训练。利用摊销变分推断的思想，通过研究 $\mathcal{Q}^*(z)$ 的函数结构，引入神经网络 q_φ 对变分分布 $\mathcal{Q}^*(z)$ 进行参数化，在假设 q_φ 能够完美近似 $\mathcal{Q}^*(z)$ 的前提下，即可以构建起带有深度学习结构的概率隐变量模型的训练流程。随着这一研究思路的发展，后续研究者逐渐开发出了变分循环神经网络[112]（variational recurrent neural

network）、动态变分自编码器[65]（dynamical variational autoencoder）等模型，从而扩展了 VAE 在动态隐变量推断场景下的应用能力。

算法 2.2 基于摊销变分推断的 EM 算法伪代码

输入： 数据：$\{x_i|i=1,...,\mathrm{N}, x_i \in \mathbb{R}^{\mathrm{D_{LV}}}\}$、条件概率模型 $p_\theta(x|z)$，模型参数 θ、以及先验分布 $\mathcal{P}(z)$、优化迭代次数迭代次数 T，收敛阈值 ϵ。

输出： 优化后的模型参数 θ^*。

$\mathcal{L}_0 \leftarrow +\infty$
for $\tau \leftarrow 0$ to $\mathrm{T}-1$ **do**
  $\theta_{\tau+1}, \varphi_{\tau+1} \leftarrow \arg\max_{\theta,\varphi} \mathbb{E}_{q_\varphi(z|x)}\left[\log p_\theta(x|z)\right] - \mathbb{D}_{\mathrm{KL}}\left[q_\varphi(z|x)\|\mathcal{P}(z)\right]$
  $\mathcal{L}_{\tau+1} \leftarrow \mathbb{E}_{q_\varphi(z|x)}[\log \frac{p_\theta(x|z)\mathcal{P}(z)}{q_\varphi(z|x)}]|_{\theta=\theta_{\tau+1},\varphi=\varphi_{\tau+1}}$
  $\theta^* \leftarrow \theta_{\tau+1}$
  if $\mathcal{L}_{\tau+1} - \mathcal{L}_\tau \leq \epsilon$ **then**
    打断循环
  end if
end for
return θ^*

2.2 泛函优化理论

2.2.1 概率密度函数的优化

考虑如下优化问题：

$$\arg\min_{z} \quad \boldsymbol{F}(z) \tag{2-9}$$

其中 z 是待优化变量，则利用梯度下降方法，可以得到第 τ 次迭代和 $\tau+1$ 次迭代的优化结果：

$$z_{\tau+1} = z_\tau - \varepsilon \nabla_z \boldsymbol{F}(z)|_{z=z_\tau}, \tag{2-10}$$

其中 ε 是步长。值得注意的是，式(2-10) 可以重写为以下等价的优化问题形式：

$$z_{\tau+1} = \arg\min_{z} \quad \frac{1}{2\varepsilon}\|z - z_\tau\|_2^2 + \boldsymbol{F}(z). \tag{2-11}$$

观察上述表达式，可以发现优化关于变量 z 的函数 $\boldsymbol{F}(z)$ 可以视作在以 z_τ 为中心、半径为 ε 的搜索空间内，寻找函数 $\boldsymbol{F}(z)$ 的最小值。在此过程中，对变量 z 隐含了欧几里得空间（Euclidean space）的假设，由于每次迭代都确保 $z_{\tau+1} \in \mathbb{R}^{\mathrm{D_{LV}}}$，因此这一迭代过程被认为是良定义的（well-defined）。

为了能够让式(2-11)所定义的优化过程能够处理约束域上的变量，可以考虑对式(2-9)、式(2-10)和式(2-11)进行如下变形：

$$\begin{aligned}&\arg\min_{z\in\mathbb{R}^{\mathrm{D_{LV}}}}\ \boldsymbol{F}(z)\\ \Rightarrow& z_{\tau+1}=z_\tau-\varepsilon\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}\\ \overset{\text{(i)}}{\Rightarrow}& z_{\tau+1}=\arg\min_{z\in\mathbb{R}^{\mathrm{D_{LV}}}}\ \frac{1}{2\varepsilon}\|z-z_\tau\|_2^2+[\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}]^\top z,\end{aligned}\tag{2-12}$$

其中步骤“(i)”的右端的目标函数为 $[\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}]^\top z$ 而非式(2-11)中的 $\boldsymbol{F}(z)$。在此基础上，如果希望处理在某个限制域，如对处在某个凸集（convex set）$\mathcal{M}$ 上的变量 z 的优化，一个可行的方案是更换步骤“(i)”中的相似度度量指标（discrepancy metric），从而重构 z 的迭代流程，使得 $z\in\mathcal{M}$ 恒成立，从而保证迭代过程的良定性。

为了实现该目标，考虑引入一个连续可微且严格凸的标量函数 $\Psi(z):\mathcal{M}\to\mathbb{R}\cup\{\infty\}$，对式(2-12)的步骤“(i)”进行重构，得到如下结果：

$$z_{\tau+1}=\arg\min_{z\in\mathcal{M}}\ \frac{1}{\varepsilon}\{\Psi(z)-\Psi(z_\tau)-[\nabla_z\Psi(z)|_{z=z_\tau}]^\top[z-z_\tau]\}+[\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}]z.\tag{2-13}$$

在式(2-13)的基础上可以定义如下式所示的布雷格曼散度（Bregman divergence）：

$$\mathbb{B}[z\|z_\tau]\coloneqq\Psi(z)-\Psi(z_\tau)-[\nabla_z\Psi(z)|_{z=z_\tau}]^\top[z-z_\tau],\tag{2-14}$$

与之相对应，标量函数 $\Psi(z)$ 又被称为布雷格曼势（Bregman potential）。注意到，当 $\Psi(z)=\frac{1}{2}\|z\|_2^2$ 时，式(2-13)即为式(2-12)，根据这个观察，可以发现，通过对布雷格曼势 $\Psi(z)$ 的修改，即可得到对限制域 $\mathcal{M}$ 良定义的迭代流程。因此，在式(2-13)的基础上，可以得到如下形如式(2-12)的镜像下降（mirror descent）的迭代表达式：

$$z_{\tau+1}=\nabla_\zeta\Psi^*\left[\nabla_z\Psi(z)-\tau\nabla_z\boldsymbol{F}(z)\right]|_{z=z_\tau},\tag{2-15}$$

其中 $\Psi^*(\zeta)$ 是 $\Psi(z)$ 的芬克尔共轭（Fenchel conjugate），其基本定义式如下：

$$\Psi^*(\zeta)\coloneqq\sup_{z\in\mathcal{M}}\ [\zeta z-\Psi(z)],\tag{2-16}$$

而 $\nabla_z\Psi(z)$ 是一个双射函数（bijective function），将 z 从域 $\mathcal{M}$ 投影至 $\Psi^*(\zeta)$ 的定义域 $\mathrm{dom}(\Psi^*)$，在这个过程，从 $\mathrm{dom}(\Psi^*)$ 到 $\mathcal{M}$ 的反向投影可以重构为下式：

$$[\nabla_z\Psi]^{-1}=\nabla_\zeta\Psi^*.\tag{2-17}$$

以此为基础，式(2-15)形成的镜像下降表达式可以分解为如下步骤：

- z_τ 被 $\nabla_z\Psi(z)$ 从限制域 $\mathcal{M}$ 内投影至 ζ_τ；
- 投影后的 ζ_τ 进行梯度下降 $\zeta_{\tau+1}=\zeta_\tau-\eta\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}$；
- 更新后的 $\zeta_{\tau+1}$ 在 $\nabla_\zeta\Psi^*(\zeta)$ 的作用下被投影回限制域 $\mathcal{M}$，完成本次迭代。

尽管布雷格曼散度与镜像下降为特定空间内的变量优化提供了统一框架，但许多关键应用（如变分推断）的核心是处理概率密度函数本身。因此，在探讨如何优化这些函数之前，有必要先将注意力转向概率密度函数的性质及其变化方式，以为后续引入针对性的优化方法奠定理论基础。为此，根据式(2-5)，概率隐变量建模的核心在于如何在令某个泛函 $\mathcal{F}[\mathcal{Q}(z)]:\mathscr{P}_2(\mathrm{D_{LV}})\to\mathbb{R}^+\cup\{0\}$ 指标的度量最小的前提下求解密度函数 $\mathcal{Q}(z):\Omega\to\mathbb{R}^+\cup\{0\}$，在这里，泛函 $\mathcal{F}[\mathcal{Q}(z)]$ 的定义域 $\mathscr{P}_2(\mathrm{D})$ 被称为沃瑟斯坦空间[113-114]（Wasserstein space），其定义式如下：

$$\mathscr{P}_2(\mathrm{D_{LV}})\coloneqq\{\mathcal{Q}(z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^+\cup\{0\}|\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}(z)\mathrm{d}z=1,\mathbb{E}_{\mathcal{Q}(z)}\left[z^\top z\right]<\infty\}. \tag{2-18}$$

常见的分布如高斯分布（Gaussian distribution）、混合高斯分布（mixture of Gaussian distribution）、迪利克雷分布（Dirichlet distribution）等，都可以被纳入沃瑟斯坦空间这一集合中。而在通过改变分布 $\mathcal{Q}(z)$ 优化泛函 $\mathcal{F}[\mathcal{Q}(z)]$ 的过程中，概率密度函数 $\mathcal{Q}(z)$ 必须时满足如下归一化约束条件：

$$\int_\Omega\mathcal{Q}(z)\,\mathrm{d}z=1. \tag{2-19}$$

换言之，在优化概率密度函数的过程中，始终需要保证条件式(2-19)的严格成立，以确保优化过程的数学良定性（well-definedness）。为了保证这一良定性，接下来的内容将详细阐述概率密度函数在优化过程中隐含的约束机制。假设有一组粒子 $z\in\mathbb{R}^{\mathrm{D_{LV}}}$ 服从分布 $\mathcal{Q}(z)$，对 z 引入如下所示的微扰（perturbation）$\boldsymbol{T}(z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^{\mathrm{D_{LV}}}$：

$$\boldsymbol{T}(z)=z+\varepsilon\phi(z), \tag{2-20}$$

其中 $\phi(z)$ 是微扰方向，ε 是无穷小量，而映射 $\boldsymbol{T}(z)$ 又被称为传输映射（transportation map）。在此基础上，微扰后的粒子 $z_{\boldsymbol{T}}\in\mathbb{R}^{\mathrm{D_{LV}}}$ 的坐标可以通过下式给出：

$$z_{\boldsymbol{T}}=\boldsymbol{T}(z)=z+\varepsilon\phi(z), \tag{2-21}$$

根据式(2-19)的要求，可以将该形变过程在二维场景和均匀分布时的情景在图2.3给出，则 $z_{\boldsymbol{T}}$ 对应的概率密度函数 $\mathcal{Q}_{\boldsymbol{T}}(z)$ 可以在式(2-19)的基础上通过下式给出：

$$\mathcal{Q}_{\boldsymbol{T}}(z)\left|\det \nabla_z \boldsymbol{T}(z)\right| = \mathcal{Q}(z). \tag{2-22}$$

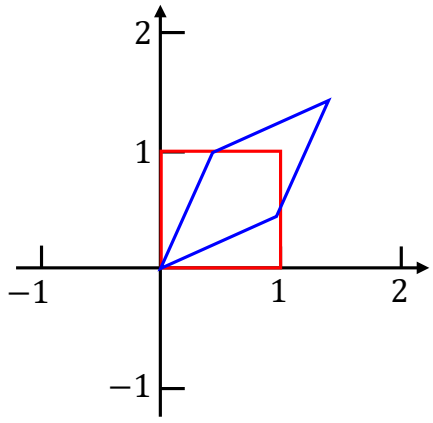


图 2.3 均匀分布的支撑变化示意图（红色为变形前，蓝色为变形后，变形前后 $\int_\Omega \mathcal{Q}(z)\mathrm{d}z = 1$）

进一步地，引入如下引理对概率密度函数 $\mathcal{Q}(z)$ 因微扰 $\phi(z)$ 所产生的演化进行描述：

引理 2.1： 设 $\phi(z) : \mathbb{R}^{\mathrm{D_{LV}}} \to \mathbb{R}^{\mathrm{D_{LV}}}$ 表示微扰方向，当微扰的大小 τ 趋近于 0 时候，概率密度函数在空间上不同时间切片 τ 的演化满足如下的偏微分方程（partial differential equation，PDE）：

$$\frac{\partial \mathcal{Q}_\tau(z)}{\partial \tau} = -\nabla_z \cdot \left[\mathcal{Q}_\tau(z)\phi(z)\right]. \tag{2-23}$$

证明： 在进行引理的证明前，需要说明的是，由于本文假设 $\mathcal{Q}(z) \in \mathscr{P}_2(\mathrm{D_{LV}})$，因此证明过程对函数进行一阶泰勒展开的操作是合理的；利用泰勒展开 $\dfrac{\partial \boldsymbol{T}(z)}{\partial z} = I + \tau\nabla_z\phi(z)$ 和 $\log\det(I+A) \approx \mathrm{Trace}(A)$（其中 Trace 为矩阵的迹），可以将式(2-23)重写为下列形式：

$$\log \mathcal{Q}(z) = \log \mathcal{Q}_{\boldsymbol{T}}(z) + \tau[\nabla_z \log \mathcal{Q}_{\boldsymbol{T}}(z)]^\top \phi(z) + \tau\nabla_z \cdot \phi(z), \tag{2-24}$$

因此，可以得到如下结果：

$$\frac{\mathcal{Q}_{\boldsymbol{T}}(z) - \mathcal{Q}(z)}{\tau} = \mathcal{Q}_{\boldsymbol{T}}(z)\frac{1 - \exp\left\{\tau[(\nabla_z \log \mathcal{Q}_{\boldsymbol{T}}(z))^\top \phi(z) + (\nabla_z \cdot \phi(z))]\right\}}{\tau}, \tag{2-25}$$

进一步利用泰勒展开 $\exp(z) \approx 1 + z$，式(2-25)右侧可化为：

$$\begin{aligned} &\mathcal{Q}_{\boldsymbol{T}}(z)\frac{1 - \exp\left\{\tau[(\nabla_z \log \mathcal{Q}_{\boldsymbol{T}}(z))^\top \phi(z) + \nabla_z \cdot \phi(z)]\right\}}{\tau} \\ =& -\left[\nabla_z \mathcal{Q}_{\boldsymbol{T}}(z)\right]^\top \phi(z) + \mathcal{Q}_{\boldsymbol{T}}(z)\nabla_z \cdot \phi(z). \end{aligned} \tag{2-26}$$

注意到，对式(2-25)的左侧取极限操作 $\lim_{\tau\to 0}$，可得下列方程：

$$\frac{\partial \mathcal{Q}_\tau(z)}{\partial \tau} = -\left[\nabla_z \mathcal{Q}_\tau(z)\right]^\top \phi(z) - \mathcal{Q}_\tau(z)\nabla_z \cdot \phi(z) = -\nabla_z \cdot \left[\mathcal{Q}_\tau(z)\phi(z)\right], \tag{2-27}$$

则式(2-23)得证。 **证毕**

前文已经定义了由概率分布构成的沃瑟斯坦空间，并讨论了分布在此空间中演化所遵循的 PDE。因此，自然需要一个差异度度量（dissimilarity）来量化这些分布之间的差异，这正是沃瑟斯坦距离[115]的作用。与 KL 散度等 f-散度[116]（f-divergence）不同，沃瑟斯坦距离提供了一种具有优良几何性质的度量工具。尤其是 2-沃瑟斯坦距离（$\mathcal{W}_2$ 距离），因其良好的数学特性和直观解释，在生成模型[117-118]和计算生物学[119-120]等领域得到了广泛应用。

据维拉尼的著作[121]，假设所讨论的分布支撑为 $\mathbb{R}^{\mathrm{D_{LV}}}$，$\boldsymbol{T}(z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^{\mathrm{D_{LV}}}$ 是一个将 $\mathcal{Q}(z)$ 映射到 $\mathcal{Q}_{\boldsymbol{T}}(z)$ 的传输映射。则 $\mathcal{W}_2$ 距离的平方可以定义为：

$$\mathcal{W}_2^2(\mathcal{Q}(z),\mathcal{Q}_{\boldsymbol{T}}(z)) = \inf_{\boldsymbol{T}:\boldsymbol{T}_\#\mathcal{Q}(z)=\mathcal{Q}_{\boldsymbol{T}}(z)} \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}(z)\|z-\boldsymbol{T}(z)\|_2^2 \mathrm{d}z. \tag{2-28}$$

其中，$\boldsymbol{T}_\#\mathcal{Q}(z)=\mathcal{Q}_{\boldsymbol{T}}(z)$ 表示 $\boldsymbol{T}(z)$ 将概率分布 $\mathcal{Q}(z)$ 映射到 $\mathcal{Q}_{\boldsymbol{T}}(z)$，也就是对所有的可测集 $\mathscr{A}$，满足 $\mathcal{Q}(\boldsymbol{T}^{-1}(\mathscr{A}))=\mathcal{Q}_{\boldsymbol{T}}(\mathscr{A})$，$\|z-\boldsymbol{T}(z)\|_2^2$ 表示 z 与其映射到的位置 $\boldsymbol{T}(z)$ 之间的欧氏距离的平方，而井号 # 表示推进测度（pushforward measure）。下确界（infimum）符号 inf 表示在所有满足映射关系 $\boldsymbol{T}_\#\mathcal{Q}(z)=\mathcal{Q}_{\boldsymbol{T}}(z)$ 的函数 $\boldsymbol{T}$ 中，寻找使得积分值最小的那个函数。将使得积分值最小的 $\boldsymbol{T}$ 记为 $\boldsymbol{T}^*$，即最优映射。进一步，可以引入一个与映射相关的微扰：$\phi(z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^{\mathrm{D_{LV}}}$，将映射表示为 $\boldsymbol{T}(z)\coloneqq z+\varepsilon\phi(z)$，其中 ε 为无穷小量，则式(2-28)所定义的 $\mathcal{W}_2$ 距离的平方可以重构为：

$$\mathcal{W}_2^2(\mathcal{Q}(z),\mathcal{Q}_{\boldsymbol{T}}(z)) = \inf_{\phi:(z+\varepsilon\phi)_\#\mathcal{Q}(z)=\mathcal{Q}_{\boldsymbol{T}}(z)} \varepsilon^2 \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}(z)\|\phi(z)\|_2^2 \mathrm{d}z. \tag{2-29}$$

2.2.2 最优控制问题与极大值原理

在前面的章节中，已介绍了空间尺度上的泛函优化，而本节将讨论时间尺度上的泛函优化，即最优控制[103,122]（optimal control）。最优控制是泛函优化在时间域中的重要应用形式，其目标在于通过设计有效的控制策略，确保动态系统在给定的时间区间内实现某种性能的最优化。与空间尺度上的泛函优化相比，最优控制特别强调动态系统随时间演化的过程，并致力于在时间范围内优化系统的行为。

最优控制问题的核心在于研究如何在满足动态约束条件和边界条件的前提下，选择合适的控制变量，以使目标泛函达到最优值。本小节将以庞特里亚金极大值原理[107]（Pontryagin's maximum value principle）为基础，介绍最优控制的求解方法。具体而言，考虑状态变量 $z \in \mathbb{R}^{\mathrm{D_{LV}}}$，其随时间的演化行为可以用如下形式的常微分方程（ordinary differential equation，ODE）进行描述：

$$\frac{\mathrm{d}z_t}{\mathrm{d}\tau} = f(z_t, t) + u(z_t) \tag{2-30}$$

其中，$f(z_t, t)$ 被称为系统的动态特性，它表征了在没有控制输入 $u(z_t)$ 的情况下，状态如何随着时间变化。通过适当选择控制输入 $u(z_t)$，可以有效地影响系统的动态行为，从而实现预定的控制目标。为了量化控制目标，定义形式如下的性能泛函 $\mathcal{F}$：

$$\arg\min_{u(z_t)} \ \mathcal{F} \coloneqq \int_{t_0}^{t_f} \boldsymbol{L}(z_t, u(z_t), t)\,\mathrm{d}\tau + \Phi(z_{t_f}), \tag{2-31}$$

其中，$\boldsymbol{L}(z_t, u(z_t), t)$ 是拉格朗日函数，描述在时间 t 时状态 z_t 和控制输入 $u(z_t)$ 的即时成本或效益，$\Phi(z_{t_f})$ 是终端成本函数，表示在终止时间 t_f 处的状态对目标的贡献。庞特里亚金极大值原理为求解最优控制问题提供了系统化的方法。其基本思想是通过构建哈密顿量 $\boldsymbol{H}(z_t, u(z_t), \lambda_t, t)$ 来转化为极值问题。哈密顿量定义为：

$$\boldsymbol{H}(z_t, u(z_t), \lambda_t, t) = \boldsymbol{L}(z_t, u(z_t), t) + \lambda_t^\top f(z_t, t), \tag{2-32}$$

其中，$\lambda_t \in \mathbb{R}^{\mathrm{D_{LV}}}$ 是伴随态（adjoint state），反映了状态变量对目标泛函的边际影响。

根据庞特里亚金极大值原理，解决最优控制问题需满足以下条件：

- **极大值条件：**哈密顿量 $\boldsymbol{H}(z_t, u(z_t), \lambda_t, t)$ 对控制变量 $u(z_t)$ 的偏导数应为零，即：

$$\frac{\partial \boldsymbol{H}(z_t, u(z_t), \lambda_t, t)}{\partial u(z_t)} = 0. \tag{2-33}$$

- **状态方程：**系统状态变量 z 满足动态方程，即：

$$\frac{\mathrm{d}z_t}{\mathrm{d}\tau} = f(z_t, t) + u(z_t) \tag{2-34}$$

- **伴随方程：**伴随态 λ_t 满足其动态方程，通常形式为：

$$\frac{\mathrm{d}\lambda_t}{\mathrm{d}\tau} = -\frac{\partial \boldsymbol{H}(z_t, u(z_t), \lambda_t, t)}{\partial z_t}. \tag{2-35}$$

- **边界条件：**在给定的时间区间内，需要满足初始和终止条件，通常形式为 $z_{t_0}=z_0$ 和 z_{t_f} 的相关条件。例如当 $t_f\to\infty$ 时，应该满足：

$$\lim_{t_f\to\infty}\lambda_{t_f}=0, \tag{2-36}$$

与之相对应地，终端条件 $\Phi(z_{t_f})$ 通常设置为 0。

通过以上步骤，可以得到一组关于 z_t 和 $u(z_t)$ 的系统方程。求解这组方程，便能够有效地计算最优控制策略 $u(z_t)$ 和最优状态轨迹 z_t。

2.3 软测量建模任务的相关评估指标与数据集介绍

2.3.1 软测量建模评估指标

软测量建模的核心任务在于如何构建以过程变量为输入质量变量为输出的回归模型。在此基础上，假设模型给出了预测值 $\hat{y}$，数据的真实值为 y，则模型的性能可通过以下指标进行量化评估：

- **均方根误差（Root Mean Square Error，RMSE）：**

$$\mathrm{RMSE}=\sqrt{\frac{1}{\mathrm{N_{test}}}\sum_{l=1}^{\mathrm{N_{test}}}(y_l-\hat{y}_l)^2}, \tag{2-37a}$$

其中 y、$\hat{y}$ 和 $\mathrm{N_{test}}$ 分别为质量变量的真实值、质量变量的预测值和测试集样本数。

- **决定系数（Coefficient of Determination，$\mathrm{R^2}$）：**

$$\mathrm{R^2}=1-\frac{\sum_{l=1}^{\mathrm{N_{test}}}(y_l-\hat{y}_l)^2}{\sum_{l=1}^{\mathrm{N_{test}}}(y_l-\bar{y})^2}, \tag{2-37b}$$

其中 $\bar{y}$ 为质量变量的均值。

- **平均绝对百分比误差（mean absolute percentage error，MAPE）：**

$$\mathrm{MAPE}=\frac{1}{\mathrm{N_{test}}}\sum_{l=1}^{\mathrm{N_{test}}}\left|\frac{(y_l-\hat{y}_l)}{y_l}\right|\times 100\%. \tag{2-37c}$$

- **平均绝对误差（mean absolute error，MAE）：**

$$\mathrm{MAE}=\frac{1}{\mathrm{N_{test}}}\sum_{l=1}^{\mathrm{N_{test}}}|y_l-\hat{y}_l|. \tag{2-37d}$$

对软测量建模任务而言，RMSE、MAPE 和 MAE 指标越小则模型的预测精度越大，而 R^2 越接近 1 则模型预测越准确。

2.3.2 软测量建模数据集

2.3.2.1 脱丁烷精馏塔数据集

图2.4为脱丁烷精馏塔（debutanizer column）的工艺流程示意图。该精馏塔隶属于意大利锡拉库萨（Syracuse）的某炼油厂，在脱硫及轻汽油分馏工艺中扮演着重要角色[123]。本文研究所采用的脱丁烷精馏塔数据集即源于该装置的实际运行数据。作为软测量建模领域广泛应用的基准数据集[15]，它常被用来评价不同模型的预测性能，相关细节可参见 Fortuna 等人的著作[124]。

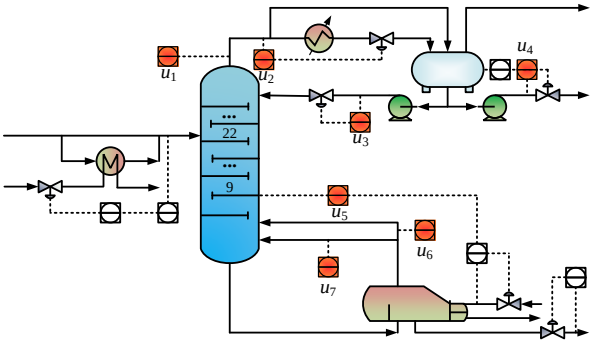


图 2.4 脱丁烷精馏塔的流程示意图

脱丁烷塔在炼油工艺中扮演着关键角色，它分离轻重组分以确保下游汽油产品的质量。为了实时监测和控制塔底丁烷浓度，本研究选择图2.4中红色区域标注的七个过程变量作为软测量模型的输入变量。这些变量的具体描述见表2.1。经过严格的数据采集和预处理，最终获得了包含 2394 组样本的脱丁烷塔数据集。

2.3.2.2 二氧化碳吸收塔数据集

在合成氨工业中，二氧化碳是主要的副产物之一。为了实现资源的有效利用和环境保护，通常需要在工艺流程中设置二氧化碳捕集与储存装置，并将捕集到的二氧化碳回

表 2.1 脱丁烷精馏塔的变量描述表

变量符号	变量描述
u_1	塔顶温度
u_2	塔顶压力
u_3	回流流率
u_4	塔顶馏出率
u_5	第 9 块塔板温度
u_6	塔釜温度 A
u_7	塔釜温度 B
y	塔釜丁烷含量

收用于尿素的生产。目前，热钾碱工艺是合成氨工业中广泛应用的二氧化碳捕集技术。该工艺利用热碱液与吸收气体在二氧化碳吸收塔内进行逆流接触，通过式(2-38)所示的化学反应高效地吸收二氧化碳。

$$\begin{cases} CO_2 + 2KOH = K_2CO_3 + H_2O \\ K_2CO_3 + H_2O + CO_2 = 2KHCO_3 \end{cases}. \tag{2-38}$$

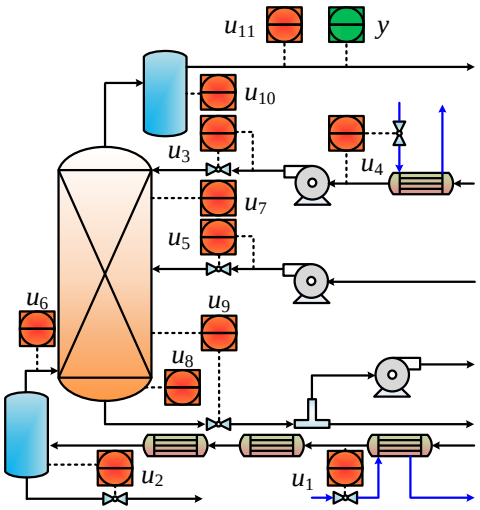


图 2.5 二氧化碳吸收塔的流程示意图

如图2.5所示，二氧化碳吸收塔作为该工艺的核心设备，其运行性能直接关系到整个系统的效率和稳定性。因此，建立能够实时预测吸收塔尾气二氧化碳出口含量的软测量模型对于维持和优化二氧化碳吸收过程至关重要。鉴于直接测量吸收塔尾气中的二氧化碳浓度面临一定的技术难题和较高的成本，本文以某合成氨工厂的二氧化碳吸收塔为研究对象，通过软测量建模方法进行分析。在建模过程中，结合工艺过程知识和详细的机理分析，系统地从流程图2.5中的红色关键测点中筛选出若干过程测量变量，作为软

测量模型的辅助变量。表2.2则对这些变量的具体信息进行了详细描述。经过严格的数据采集、清洗和预处理，最终获得了包含 6000 组样本的二氧化碳吸收塔数据集。

表 2.2 二氧化碳吸收塔的变量描述表

变量符号	变量描述
u_1	进口气体压力
u_2	缓冲罐液位
u_3	贫液入口温度
u_4	贫液体流量
u_5	半贫液体流量
u_6	进口气体温度
u_7	吸收塔压力
u_8	富液温度
u_9	吸收塔塔釜液位
u_{10}	气液分离器液位
u_{11}	出口气体压力
y	出口二氧化碳浓度

2.3.2.3 水煤气变换单元数据集

在合成氨工艺中，高纯度氢气和适宜的碳氢比是决定产品质量的关键指标。图2.6所示的水煤气变换单元（water gas shift unit）作为调控上述指标的核心单元操作，其运行性能直接影响整个工艺流程的效率与稳定性。具体而言，水煤气高低温变换过程通常采用两个绝热固定床列管式反应器串联配置。过程气体首先进入高温变换器，在适宜的反应速率下接近热力学平衡，随后经换热器冷却，并进入低温变换器以进一步提高反应转化率，实现热力学平衡的移动。通过精准控制式(2-39)所示的水煤气变换反应的反应程度，可获得满足下游生产工艺要求的变换气组成，从而为合成氨的顺利运行提供保障。

$$\mathrm{CO} + \mathrm{H_2O} \rightleftharpoons \mathrm{CO_2} + \mathrm{H_2}. \tag{2-39}$$

由于反应是放热反应，在高温条件下根据勒夏特列原理[125]（Le Chatelier’s principle）一氧化碳反应的转化率会下降，最终导致一氧化碳反应转化不彻底，与之相反的是，在低温条件下，根据阿伦尼乌斯（Arrhenius）经验公式，反应的速率会变慢，导致在一定的空速条件下一氧化碳转化率降低。因此，在实际操作过程中，往往采用不同的催化剂，分别在高温条件下和低温条件下从动力学和热力学的角度上达到最佳的转化率以产生最佳的一氧化碳转化率。上述工艺要求对反应器出口气中的一氧化碳浓度进行实时的质量预报以达到最优的监测性能。但是在实际生产过程中，由于气相色谱的化验时间延迟

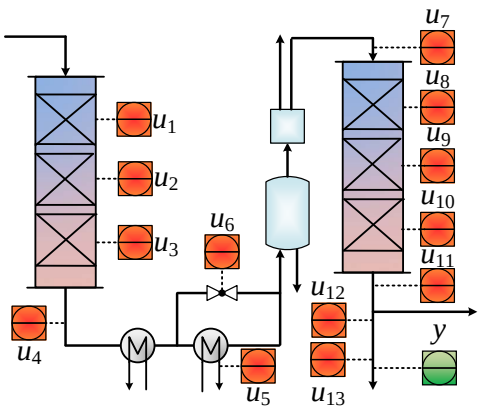


图 2.6 水煤气变换单元的流程示意图

难以对出口的一氧化碳浓度进行实时测量，因此，需要借助过程中的辅助变量，建立对应的数据驱动模型进行软测量模型的构建。为此，本文以来源于某合成氨工厂的真实生产数据为基础，以工艺过程知识和机理分析为指导，精选了如图2.5中红色标记的关键过程测点（详细信息见表2.2）作为软测量模型的辅助变量，最终构建了包含 7200 组样本的水煤气变换单元数据集。

表 2.3 水煤气变换单元的变量描述表

变量符号	变量描述
u_1	高温水煤气变换反应器温度测点 1
u_2	高温水煤气变换反应器温度测点 2
u_3	高温水煤气变换反应器温度测点 3
u_4	高温水煤气变换反应器出口温度
u_5	冷却水出口温度
u_6	弛放气温度
u_7	低温水煤气变换反应器输入温度
u_8	低温水煤气变换反应器温度测点 1
u_9	低温水煤气变换反应器温度测点 2
u_{10}	低温水煤气变换反应器温度测点 3
u_{11}	低温水煤气变换反应器出口温度
u_{12}	低温水煤气变换反应器出口压力
u_{13}	产品气压力
y	一氧化碳浓度

2.4 本章小结

本章首先系统阐述了概率隐变量模型的理论框架及其训练算法。在此基础上，本章基于概率密度函数优化与最优控制理论引入了泛函优化方法体系。本章在最后的部分从采集背景和变量构成这两个层面介绍了实验验证环节所采用的软测量数据集。

3 概率密度函数优化诱导的隐变量驱动数据补全方法

摘要： 在软测量建模过程中，设备常处于卡边操作环境，这使得测量仪表的工况往往面临苛刻条件，从而导致数据缺失问题普遍存在。为应对这一挑战，本章基于泛函优化理论，对利用概率隐变量模型进行缺失数据补全的过程进行了系统分析与改进。从空间尺度的泛函优化视角出发，阐明了概率隐变量模型在补全缺失数据时面临的推理结果易发散以及缺失机制漂移等问题，通过结合最小移动方案问题与再生核希尔伯特空间，对模型推理环节进行重新设计以提升对缺失数据的补全效果。在此基础上，进一步提出了与缺失机制无关的模型训练策略，并证明了在该策略下模型推理结果的等价性。综合上述工作，给出了基于概率隐变量模型的缺失数据补全方法并从理论上论证了该算法收敛性。实验结果表明所提出的方法在缺失数据补全准确性、敏感性分析、消融实验以及收敛性方面均具有良好的表现与优势。

关键词： 隐变量模型 缺失值补全 沃瑟斯坦空间 最小移动方案问题 再生核希尔伯特空间

3.1 引言

在工业过程软测量建模领域，数据完整性问题始终是制约模型性能的关键瓶颈。究其根源，工业传感器在复杂工况下面临的多模态干扰（包括热应力冲击、强电磁干扰、电力系统波动及通信信道不稳定等）导致采集数据普遍存在缺失现象。以典型化工过程为例，精馏塔换热器温度传感器在高压蒸汽过温工况下易出现漂移失效，而关键动设备（如离心泵、往复式压缩机）的振动传感器则常因机械共振引发数据包丢失。这些缺失数据不仅破坏时序连续性，更会显著降低软测量模型的预测精度。因此，构建有效的缺失数据补全机制已成为软测量建模流程中不可或缺的前处理环节。

概率隐变量模型，由于其独特的建模方式和推断能力，近年来在缺失数据补全领域展现出广泛的应用潜力。这类模型通过将缺失数据建模为隐变量推断问题，利用隐变量捕捉数据的潜在结构，并通过 EM 算法实现高效的缺失数据补全。但在实际应用中，现有的基于概率隐变量模型的缺失数据补全方法仍面临两方面的核心挑战。首先，数据补

全问题本质上属于判别性任务，而在传统概率隐变量框架中，为求解隐变量的后验分布，EM 算法通常引入“熵泛函”作为正则项。但这一正则项的加入往往导致推理分布向着熵最大化方向演化，从而引发补全结果的发散问题，尤其是在面对高缺失率场景时表现尤为突出。其次，隐变量模型需要显式地建模缺失数据和观测数据之间的条件分布，但由于实际工业过程中的传感器故障存在时空随机性，缺失模式具有高度的不确定性。若所建模型的条件分布无法准确匹配这种缺失模式，最终可能导致补全效果显著下降。

针对上述问题，本章提出了一种基于泛函优化理论的缺失数据补全方法。具体而言，从空间尺度的泛函优化视角入手，全面阐明了概率隐变量模型在补全缺失数据时所面临的推理发散性问题和缺失机制漂移问题的深层次原因，结合最小移动方案优化策略与再生核希尔伯特空间理论，对模型的推理过程进行了重新设计，从而有效提升了对缺失数据的补全效果。通过上述理论推导，进一步提出了一种与缺失机制无关的模型训练策略，通过减少模型对特定缺失模式的依赖性，使其能够在不同缺失机制下实现一致、鲁棒的补全效果，并从理论上证明了补全结果的分布等价性。为验证所提方法的有效性，从模拟数据补全的有效性、真实数据补全的准确性、模型软测量、下游软测量任务建模精度以及补全方法的收敛性四个维度开展了系统性实验验证，实验结果充分证实了所提出方法的优越性。

3.2 相关工作回顾与科学问题分析

在过去几十年中，概率隐变量模型因其在捕捉数据内在结构、处理不确定性以及生成高质量数据方面的独特优势，成为工业过程缺失数据补全问题中的重要工具。概率隐变量模型通常将缺失数据视为一种特殊的待推断隐变量，数据补全问题则被表述为对隐变量的推断问题。

在这一框架下，传统的期望最大化（EM）算法被扩展为对隐变量、缺失数据和模型参数三者分别进行推断、补全和优化的联合算法，从而实现缺失数据的有效推断[29,126]。例如，Zhu 等人[127]将缺失数据补全视为概率分布的矩（moment）的推断问题，并提出了分布式贝叶斯网络对缺失值进行重构，在大规模工业场景中验证了该方法的有效性。孔祥印[29]提出了显式和隐式考虑缺失变量的变量分解概率主成分分析方法，并将其扩展到深度轻隐变量模型框架中，在工业过程故障诊断的案例研究中展示了其可行性。这些

研究表明，概率隐变量模型在工业数据补全中的理论框架具有一定的通用性和适用性。进入深度学习时代后，正如第2章所提到的，由于深度网络的不可逆特性，观测数据的补全难以像统计学习时代那样直接视为隐变量进行推断。为此，研究者们尝试引入对比学习或对抗训练等框架来解决这一问题。例如，姚邹静等人[128]引入了生成对抗网络针对软测量建模任务面临的缺失值问题进行个性化补全，并在转炉炼钢的终点元素含量预测中铝的预测问题中验证了所提出方法的有效性。Liu 等人[129]通过在隐空间对缺失数据进行粗粒度特征提取与重构，并结合卷积网络的升采样模块完成细粒度特征补全，从而实现缺失数据的恢复。在此基础上，Liu 等人[130]进一步引入扩散模型以规避隐变量模型推断中的归一化常数计算，并通过随机掩盖观测数据从而构建监督信号进而实现对缺失数据的补全。另一方面，Yuan 等人[131]提出了基于生成对抗网络的方法，通过对初始补全数据与真实数据的分布差异进行迭代优化，从而逐步提高补全数据的质量。这些方法在一定程度上丰富了隐变量驱动的缺失数据补全技术体系。尽管上述方法取得了显著进展，但在工业过程数据补全中仍存在以下亟待解决的关键科学问题：

（1） **补全结果的发散性：**从推断的角度来看，隐变量模型的目标是对隐空间的“分布”进行推断，而非具体的“值”。为了对齐分布，通常需要引入一定的发散性/多样性（diversity），但这种发散性与缺失数据补全所追求的高精度（high accuracy）目标并不完全一致，可能导致推断结果的不稳定性和不确定性。

（2） **训练和推理目标的不一致性：**无论是基于统计时代的概率隐变量模型，还是深度学习时代的隐变量模型，其训练目标通常是最大化观测数据的联合分布（极大似然估计）。然而，缺失数据补全本质上是一个基于观测数据推断缺失数据的判别式任务，这使得训练目标与推理目标之间存在不一致性。此外，基于神经网络的概率隐变量模型由于生成式模型的不可逆性，通常需要通过引入掩码矩阵来掩蔽部分数据以构造标签，促使模型进行重构。如果掩码矩阵与数据的缺失模式不对齐，这种不一致性问题将进一步加剧。

3.3 基于隐变量模型的数据补全方法

3.3.1 缺失数据补全的任务阐述

在2.1节的基础上，设理想的、无缺失项的数据集记为：

$$\boldsymbol{X}^{(\text{ideal})} \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}},$$

其中 N 表示样本数量，D 表示特征（变量）维度。由于工业过程测量设备常在极端条件下运行，传感器容易发生故障，从而导致获取的测量数据中经常出现缺失项，记为：

$$\boldsymbol{X}^{(\text{obs})} = \boldsymbol{X}^{(\text{ideal})} \odot \boldsymbol{M} + \texttt{NaN} \odot \left(\mathbb{1}_{\mathrm{N}\times\mathrm{D}} - \boldsymbol{M}\right),$$

其中“$\odot$”表示哈达玛积（Hadmard product），NaN 是英文“not a number”的首字母缩写，$\mathbb{1}_{\mathrm{N}\times\mathrm{D}}$ 为全 1 矩阵，$\boldsymbol{M} \in \{0,1\}^{\mathrm{N}\times\mathrm{D}}$ 则是遮罩矩阵，用于标记某一元素是否观测到（1 表示已观测，0 表示缺失），而缺失率 p_{miss} 则定义为 $p_{\text{miss}} \coloneqq 1 - \frac{\sum_i^{\mathrm{N}} \sum_{j=1}^{\mathrm{D}} \boldsymbol{M}_{i,j}}{\mathrm{N}\times\mathrm{D}}$。缺失数据补全的核心任务表述为根据现有的观测数据 $\boldsymbol{X}^{(\text{obs})}$（“obs”是观测“observation”的英文首字母），对其缺失条目进行补全，从而得到一个“完整”的数据矩阵，记为：

$$\widehat{\boldsymbol{X}} = \boldsymbol{X}^{(\text{obs})} \odot \boldsymbol{M} + \boldsymbol{X}^{(\text{imp})} \odot \left(\mathbb{1}_{\mathrm{N}\times\mathrm{D}} - \boldsymbol{M}\right),$$

其中 $\boldsymbol{X}^{(\text{imp})}$（“imp”是补全“imputation”的英文首字母）是由缺失数据补全方法产生的补全值矩阵。

根据 Rubin 的论文[132]，缺失数据的产生机制通常可分为以下三种类型。

- **完全随机缺失（Missing Completely at Random，MCAR）：**在 MCAR 场景下，数据是否缺失与任何观测到或未观测到的变量均无关，即缺失完全由随机过程产生，无任何可识别的系统性偏差。

- **可忽略随机缺失（Missing at Random，MAR）：**在 MAR 场景下，缺失的概率仅依赖于已经观测到的变量，而与尚未观测到的变量无关。此时，虽然数据缺失和某些特定的已观测变量可能存在依赖关系，但此类依赖通常不会造成显式的偏差。

- **非随机缺失（Missing Not at Random，MNAR）：**在 MNAR 场景下，数据的缺失与某些未观测变量或缺失值本身相关，这时的缺失机制更为复杂。如果不引入额外的先验假设或领域知识，往往难以对 MNAR 场景下的缺失进行定量建模[133]。

在 MCAR 和 MAR 场景下，往往无需显式建模缺失数据分布，即可利用适当的补全方法修正缺失对模型和分析结果的影响；然而在 MNAR 场景下，由于缺失机制与未观测信息挂钩，如果没有额外的假设或对缺失机制的限制，插补结果容易受到较大偏差[133-134]。针对 MNAR 引发的一系列更为复杂的问题，需要进一步引入更加复杂的统计学方法[135]加以处理。因此，本章在进行理论推导的过程中不引入有关缺失机制的讨论。

3.3.2　基于概率隐变量模型的缺失数据补全方法

将需要补全的缺失数据记为 $\boldsymbol{X}^{(\text{miss})}$。在概率隐变量模型的框架下，将 $\boldsymbol{X}^{(\text{miss})}$ 视为潜在(隐)变量，通过为观测数据 $\boldsymbol{X}^{(\text{obs})}$ 和缺失数据 $\boldsymbol{X}^{(\text{miss})}$ 构建联合分布 $p_\theta(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})})$，并使用 EM 算法对 $\boldsymbol{X}^{(\text{miss})}$ 进行推断，便可实现对缺失数据的补全。具体而言，将迭代轮次记为 τ，基于概率隐变量模型的缺失数据补全可以分解为如下类似 VBEM 算法的交替优化步骤：

- **E 步（期望步）：**根据下式计算完整数据对数似然的期望：

$$\begin{aligned} &\mathcal{Q}_{\tau+1}(\boldsymbol{X}^{(\text{miss})}) \\ =&\arg\max_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})} \ \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\log p_\theta(\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{obs})})] \\ =&\arg\min_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})} \ \mathbb{D}_{\text{KL}}[\mathcal{Q}(\boldsymbol{X}^{(\text{miss})}) \| \mathcal{P}(\boldsymbol{X}^{(\text{miss})} | \boldsymbol{X}^{(\text{obs})})] \end{aligned} \tag{3-1}$$

- **M 步（最大化步）：**更新模型参数，以最大化 E 步中计算得到的期望：

$$\theta_{\tau+1} = \arg\max_{\theta} \ \mathbb{E}_{\mathcal{Q}_{\tau+1}(\boldsymbol{X}^{(\text{miss})})}[\log p_\theta(\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{obs})})] \tag{3-2}$$

通过交替执行 E 步和 M 步，EM 算法即能得到缺失数据的补全结果。

3.4　概率密度函数优化诱导的隐变量驱动数据补全框架

3.4.1　研究动机分析

本小节首先引入如下定理说明可测集 Ω 支撑的最大熵分布以支撑后续的研究动机分析：

定理 3.1： 令 $\Omega \subseteq \mathbb{R}^{\mathrm{D}}$ 为一个可测集，且其勒贝格测度（Lebesgue measure）$|\Omega|$ 有限，$0 < |\Omega| < \infty$。在 Ω 上考虑随机变量 x，其分布由概率密度函数 $\mathcal{Q}(x)$ 给定，且定义其熵泛函为 $\mathbb{H}[\mathcal{Q}(x)] \coloneqq -\int_{\Omega} \mathcal{Q}(x) \log \mathcal{Q}(x)\,\mathrm{d}x$。则在所有满足 $\mathcal{Q}(x) \in \{\mathcal{Q}(x) | \mathcal{Q}(x) \geq 0, \int_{\Omega} \mathcal{Q}(x)\,\mathrm{d}x = 1\}$ 的概率密度函数中，使得 $\mathbb{H}[\mathcal{Q}(x)]$ 最大的分布为：

$$\mathcal{Q}^{*}(x) = \begin{cases} \dfrac{1}{|\Omega|}, & x \in \Omega, \\ 0, & x \notin \Omega. \end{cases} \tag{3-3}$$

证明：为了保证概率密度函数归一化约束 $\int_{\Omega} \mathcal{Q}(x)\,\mathrm{d}x = 1$ 的前提下最大化微分熵 $\mathbb{H}[\mathcal{Q}(x)]$，尝试引入拉格朗日乘子（Lagrangian multiplier）$\lambda \in \mathbb{R}$ 构造下列拉格朗日泛函（Lagrangian functional）：

$$\mathcal{F}[\mathcal{Q}(x)] = -\int_{\Omega} \mathcal{Q}(x)\ \log\big(\mathcal{Q}(x)\big)\,\mathrm{d}x\ +\ \lambda\Big(\int_{\Omega} \mathcal{Q}(x)\,\mathrm{d}x - 1\Big), \tag{3-4}$$

并对密度函数 $\mathcal{Q}(x)$ 求解一阶变分（其中 δ 为一阶变分符号），令其为零可以得到：

$$\frac{\delta \mathcal{F}[\mathcal{Q}(x)]}{\delta \mathcal{Q}(x)}\ =\ -\log \mathcal{Q}(x)\ -\ 1\ +\ \lambda\ =\ 0. \tag{3-5}$$

验证二阶变分，可以得到：

$$\frac{\delta^{2} \mathcal{F}[\mathcal{Q}(x)]}{\delta \mathcal{Q}^{2}(x)}\ =\ -\frac{1}{\mathcal{Q}(x)}\ \leq\ 0, \tag{3-6}$$

根据勒让德条件[136]（Legendre condition）可知式(3-5)得到的极值点为式(3-4)所定义的泛函的最大值点。由此可得：

$$\log \mathcal{Q}(x) = \lambda - 1 \Rightarrow \mathcal{Q}(x) = e^{\lambda - 1}, \tag{3-7}$$

上述等式表明在支撑集 Ω 上的概率密度函数 $\mathcal{Q}(x)$ 为常数。

由于 $\mathcal{Q}(x)$ 需满足 $\int_{\Omega} \mathcal{Q}(x)\,\mathrm{d}x = 1$，因此将式(3-7)右端代入 $\int_{\Omega} \mathcal{Q}(x)\,\mathrm{d}x = 1$ 可得：

$$e^{\lambda - 1} \int_{\Omega} \mathrm{d}x\ =\ 1\ \ \Longrightarrow\ \ e^{\lambda - 1} = \frac{1}{|\Omega|}.$$

因此：

$$\mathcal{Q}^{*}(x) = \frac{1}{|\Omega|}, \quad (x \in \Omega).$$

则式(3-3)得证。　　**证毕**

在此基础上，从极大似然的角度考虑数据补全任务，从极大似然的角度而言，数据补全任务属于判别性任务，因此理想的数据补全损失函数应该设计为：

$$\arg\max_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})} \mathcal{L}^{\text{imp}} \coloneqq \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})]. \tag{3-8}$$

在这个代价泛函的基础上，基于概率隐变量模型的缺失数据补全方法存在以下两个方面的不足之处：

- **补全结果的发散性：**对比式(3-1)和式(3-8)，可以发现，在进行数据补全时，基于概率隐变量模型的数据补全方法在 E 步进行数据补全时优化的代价泛函和 $\mathcal{L}^{\text{imp}}$ 的关系可以通过下式给出：

$$-\mathbb{D}_{\text{KL}}[\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})||\mathcal{P}(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})] = \mathcal{L}^{\text{imp}} + \mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})]. \tag{3-9}$$

  换言之，在 E 步进行数据补全的时候，模型会隐式地加入“熵泛函”，将补全值向着“熵泛函”最大化的方向进行优化。根据定理3.1，可以发现基于 EM 算法得到的补全结果会趋向于支撑 Ω 上的均匀分布，而均匀分布的概率密度函数在支撑 Ω 上是一个常数，换言之在支撑 Ω 上的任何一个点都有可能成为补全值。因此，基于式(3-8)所构建的补全流程会存在补全结果的发散问题，最终不利于模型对数据进行精确补全。

- **训练目标的不一致性：**数据补全的目标函数是条件概率密度函数 $p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$ 而不是联合概率密度函数 $p_\theta(\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{obs})})$，在参数 θ 训练的过程中通常最大化的是联合概率密度函数 $p_\theta(\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{obs})})$。这一差异最终会导致模型在推理的时候会造成推理和训练目标的不一致性，最终降低数据补全的精度。

在上述分析的基础上，接下来的重点在于探讨如何突破这两大不足，通过采用泛函优化的思路来重新设计推理流程以及隐变量训练目标，以构建更有效的缺失数据补全方法。

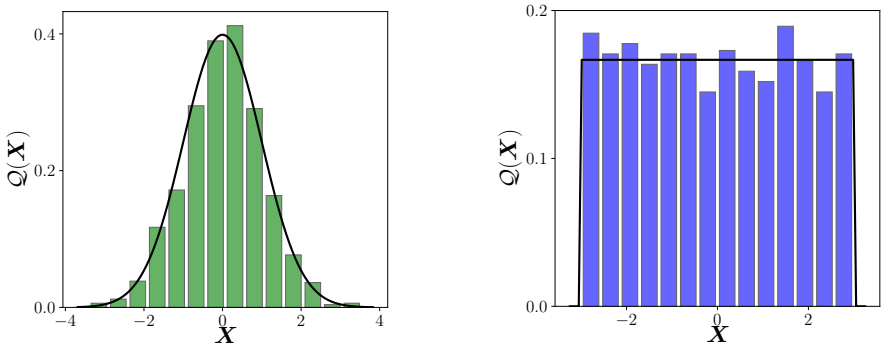
3.4.2 利用负熵泛函抑制发散的数据补全方法

在数据补全任务中，发散性问题会显著降低补全值的可靠性，导致补全结果难以反映真实数据分布。根据定理 3.1，熵泛函作为正则项时，模型倾向于最大化熵，从而导

致补全结果在支撑 Ω 上趋于均匀分布，这种现象加剧了补全值的发散性。因此，为了抑制发散性问题，可以通过在熵泛函前添加负号，将其转化为“负熵泛函”，从而重新设计 E 步的损失泛函如下：

$$\begin{aligned}
&\mathcal{F}^{\mathrm{NER}} \\
=&\mathcal{L}^{\mathrm{imp}} - \mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})] \\
=&\int_{\mathbb{R}^{\mathrm{D}}} \mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})[\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})]\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})} + \int_{\mathbb{R}^{\mathrm{D}}} \mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})\,\log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})\,\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}.
\end{aligned} \tag{3-10}$$

其中“NER”是负熵正则（negative entropy regularization）的英文首字母缩写。



(a) $\boldsymbol{X} \sim \mathcal{N}(0,1)$ (b) $\boldsymbol{X} \sim \mathcal{U}[-3,3]$

图 3.1 $\mathcal{Q}(\boldsymbol{X})$ 在不同概率密度函数下的归一化频率分布直方图对比

在 E 步最大化 $\mathcal{F}^{\mathrm{NER}}$ 时，负熵泛函项 $-\mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})]$ 会隐式地将熵向最小化的方向优化，从而让补全值更加“集中”在某个特定值上。这一过程可以通过图3.1直观理解：

- **图3.1(a)：** 当 $\mathcal{Q}(\boldsymbol{X}) \sim \mathcal{U}[-3,3]$ 时，分布的熵泛函为 $\mathbb{H}[\mathcal{Q}(\boldsymbol{X})] = \log[3-(-3)] \approx 1.79$，表示补全值分布较为均匀，发散性较高。
- **图3.1(b)：** 而当 $\mathcal{Q}(\boldsymbol{X}) \sim \mathcal{N}(0,1)$ 时，分布的熵泛函为 $\mathbb{H}[\mathcal{Q}(\boldsymbol{X})] = \frac{1}{2}\log(2\pi e) \approx 1.42$，显示补全值更加集中在某个特定区域，从而显著减少了发散性。

综上所述，通过引入负熵泛函，E 步的优化过程从最大化熵转变为最小化熵，从而有效减少了补全值的发散性。

尽管引入负熵泛函能够有效抑制补全值的发散性，使得补全结果更加集中于某一个特定的值，从而显著提高模型的稳定性和数据补全的精度，但这一改进也带来了新的

挑战。具体而言，由于负熵项的引入，损失泛函 $\mathcal{F}^{\mathrm{NER}}$ 的优化过程不再具备传统 EM 算法通过 KL 散度推导最优解的解析性结构。这一问题的核心在于，负熵泛函正则项会使得 E 步对应的优化命题的解析解形式被破坏，无法再直接写出分布 $\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})$ 的闭式解（closed form solution），使得在补全过程中无法直接利用传统的贝叶斯推断框架简化优化过程。为此，必须重新设计优化策略以构造新的数据补全流程。

基于2.2.1节的理论基础，通过将2.2.1节式(2-11)中的欧几里得距离 $\|z-z_\tau\|_2^2$ 替换为 2-沃瑟斯坦距离 $\mathcal{W}_2(\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{X}),\mathcal{Q}(\boldsymbol{X}))$，并将目标函数 $\boldsymbol{F}(z)$ 替换为代价泛函 $\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]$，可建立如下最小移动方案[137-138]（minimizing movement scheme）优化问题：

$$\inf_{\boldsymbol{T}(\boldsymbol{X})} \ \mathcal{F}[\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{X})]-\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]+\frac{1}{2\varepsilon}\mathcal{W}(\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{X}),\mathcal{Q}(\boldsymbol{X})), \tag{3-11}$$

在假设映射 $\boldsymbol{T}(\boldsymbol{X})$ 具有参数化形式 $\boldsymbol{T}(\boldsymbol{X})=\boldsymbol{X}+\varepsilon\phi(\boldsymbol{X})$（其中 ε 为无穷小量，$\phi(\boldsymbol{X}):\mathbb{R}^{\mathrm{D}}\to\mathbb{R}^{\mathrm{D}}$ 表示微扰方向）的前提下，通过泛函优化方法，可以建立如下定理来刻画最优微扰方向 $\phi^*(\boldsymbol{X})$ 的解析表达式：

定理 3.2： 对于定义在概率测度空间上的最小移动方案优化问题，在 $\boldsymbol{T}(\boldsymbol{X})$ 满足下式的条件下：

$$\boldsymbol{T}(\boldsymbol{X})=\boldsymbol{X}+\varepsilon\phi(\boldsymbol{X}), \tag{3-12}$$

其中 ε 为无穷小量，$\phi(\boldsymbol{X}):\mathbb{R}^{\mathrm{D}}\to\mathbb{R}^{\mathrm{D}}$ 表示微扰方向，最优微扰方向 $\phi^*(\boldsymbol{X})$ 由代价泛函 $\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]$ 的一阶变分决定：

$$\phi^*(\boldsymbol{X})=-\nabla_{\boldsymbol{X}}\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}. \tag{3-13}$$

证明： 在沃瑟斯坦空间中，概率密度函数 $\mathcal{Q}(\boldsymbol{X})$ 随着时间 τ 的演化过程应该满足如下的偏微分方程：

$$\frac{\partial\mathcal{Q}_\tau(\boldsymbol{X})}{\partial\tau}=-\nabla_{\boldsymbol{X}}\cdot[\mathcal{Q}_\tau(\boldsymbol{X})\phi(\boldsymbol{X})]\Rightarrow\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{X})=\mathcal{Q}(\boldsymbol{X})-\varepsilon\nabla_{\boldsymbol{X}}\cdot\mathcal{Q}(\boldsymbol{X})\phi(\boldsymbol{X})+\mathbf{H.O.T.}(\varepsilon^2), \tag{3-14}$$

其中 **H.O.T.** 是高阶项（higher order term）的英文首字母缩写。

根据2.2.1节中 $\mathcal{W}_2$ 距离的定义，可以得到如下结果：

$$\mathcal{W}_2^2(\mathcal{Q}(\boldsymbol{X}),\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{X}))=\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\boldsymbol{X}-\boldsymbol{T}^*(\boldsymbol{X})\|_2^2\mathrm{d}\boldsymbol{X}=\varepsilon^2\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\phi^*(\boldsymbol{X})\|_2^2\mathrm{d}\boldsymbol{X} \tag{3-15}$$

其中 $\boldsymbol{T}^*(\boldsymbol{X})$ 是最优的映射方向，而 $\phi^*(\boldsymbol{X})$ 为最优的微扰方向。因此，最小移动方案问题可以重写为下式：

$$\begin{aligned}
&\inf_{\phi(\boldsymbol{X})}\ \mathcal{F}[\mathcal{Q}(\boldsymbol{X})]-\varepsilon\int_{\mathbb{R}^{\mathrm{D}}}\nabla_{\boldsymbol{X}}\cdot[\mathcal{Q}(\boldsymbol{X})\phi(\boldsymbol{X})\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}]\mathrm{d}\boldsymbol{X}-\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]\\
&\qquad+\frac{\varepsilon}{2}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\phi(\boldsymbol{X})\|_2^2\mathrm{d}x,\\
\Rightarrow&\inf_{\phi(\boldsymbol{X})}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\cdot\phi^{\top}(\boldsymbol{X})\nabla_x\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}\mathrm{d}\boldsymbol{X}+\frac{1}{2}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\phi(\boldsymbol{X})\|_2^2\mathrm{d}\boldsymbol{X},\\
\overset{\text{(i)}}{\Rightarrow}&\underbrace{\inf_{\phi(\boldsymbol{X})}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\phi(\boldsymbol{X})+\nabla_{\boldsymbol{X}}\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}\|_2^2\mathrm{d}\boldsymbol{X}}_{\geq 0}
\end{aligned}\tag{3-16}$$

其中步骤“(i)”基于下列不等式：

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\cdot\phi^{\top}(\boldsymbol{X})\nabla_{\boldsymbol{X}}\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(x)}\mathrm{d}\boldsymbol{X}+\frac{1}{2}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\phi(\boldsymbol{X})\|_2^2\mathrm{d}\boldsymbol{X}\\
\leq&\frac{1}{2}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\nabla_{\boldsymbol{X}}\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}\|_2^2\mathrm{d}\boldsymbol{X}+\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\cdot\phi^{\top}(\boldsymbol{X})\nabla_{\boldsymbol{X}}\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}\mathrm{d}\boldsymbol{X}\\
&+\frac{1}{2}\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(\boldsymbol{X})\|\phi(\boldsymbol{X})\|_2^2\mathrm{d}\boldsymbol{X}.
\end{aligned}\tag{3-17}$$

因此，可知最优的微扰方向 $\phi^*(\boldsymbol{X})$ 需要满足下列等式：

$$\phi^*(\boldsymbol{X})+\nabla_{\boldsymbol{X}}\frac{\delta\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]}{\delta\mathcal{Q}(\boldsymbol{X})}=0,\tag{3-18}$$

进而式(3-13)得证。　　**证毕**

根据定理3.2，可以得到提升代价泛函 $\mathcal{F}^{\mathrm{NER}}$ 的映射方向 $\boldsymbol{T}(\boldsymbol{X}^{(\mathrm{miss})})$：

$$\begin{aligned}
&\boldsymbol{T}(\boldsymbol{X}^{(\mathrm{miss})})\\
=&\boldsymbol{X}^{(\mathrm{miss})}+\varepsilon\nabla_{\boldsymbol{X}^{(\mathrm{miss})}}\frac{\delta\mathcal{F}^{\mathrm{NER}}[\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})]}{\delta\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})}\\
=&\boldsymbol{X}^{(\mathrm{miss})}+\varepsilon[\nabla_{\boldsymbol{X}^{(\mathrm{miss})}}\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})+\nabla_{\boldsymbol{X}^{(\mathrm{miss})}}\log\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})].
\end{aligned}\tag{3-19}$$

至此，本小节给出了提升代价泛函 $\mathcal{L}^{\mathrm{NE}}$ 的最优函数映射。需要指出的是，实现式(3-19)需要显式地估计概率密度函数 $\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})$ 以计算 $\nabla_{\boldsymbol{X}^{(\mathrm{miss})}}\log\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})$，这导致了在以 Python 为代表的计算机语言上实现式(3-19)会面临较大的困难。

因此，接下来考虑将映射 $\boldsymbol{T}(\boldsymbol{X}^{(\mathrm{miss})})$ 在再生核希尔伯特空间（reproducing kernel Hilbert space，RKHS）内，寻求 $\boldsymbol{T}(\boldsymbol{X}^{(\mathrm{miss})})$ 的解析表达式，简化补全流程的实现。为此

引入 $\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) = \boldsymbol{X}^{(\text{miss})} + \varepsilon\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})$ 对 $\boldsymbol{T}(\boldsymbol{X}^{(\text{miss})})$ 进行逼近，并导出如下优化命题：

$$\underset{\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})}{\arg\min} \quad \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\|\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) - \boldsymbol{T}(\boldsymbol{X}^{(\text{miss})})\|_2^2]. \tag{3-20}$$

在此基础上，$\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})$ 的解析表达式可以通过下列定理给出：

定理 3.3：当微扰方向 $\phi(\boldsymbol{X}^{(\text{miss})})$ 处于RKHS（记为 $\mathscr{H}$）时，并且其核函数 $K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})})$ 满足边界条件 $\lim_{\|\boldsymbol{X}^{(\text{miss})}\|\to\infty} \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})}) = 0$ 时，式(3-20)所定义的优化命题的最优解为：

$$\begin{aligned}\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) =& \boldsymbol{X}^{(\text{miss})} - \varepsilon\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})\prime})}[\nabla_{\boldsymbol{X}^{(\text{miss})\prime}} K(\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{miss})\prime})] \\ &+ \varepsilon\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})\prime})}\{[\nabla_{\boldsymbol{X}^{(\text{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{miss})\prime}|\boldsymbol{X}^{(\text{obs})})]^\top K(\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{miss})\prime})\}.\end{aligned} \tag{3-21}$$

证明：式(3-20)可以重构为：

$$\begin{aligned}&\underset{\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})}{\arg\min} \quad \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\|\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) - \boldsymbol{T}(\boldsymbol{X}^{(\text{miss})})\|_2^2] \\ \Rightarrow &\underset{\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})}{\arg\min} \quad \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\|\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) - \phi^*(\boldsymbol{X}^{(\text{miss})})\|_2^2] \\ \Rightarrow^{\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})}_{\arg\min} &\ \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}\{\|\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) \\ &\qquad - [\nabla_{\boldsymbol{X}^{(\text{miss})}} \log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})}) + \nabla_{\boldsymbol{X}^{(\text{miss})}} \log \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})]\|_2^2\} \\ &\underset{\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})}{\arg\min} \quad \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\frac{1}{2}\|\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})\|_2^2] \\ \Rightarrow &\qquad - \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}\{\phi_{\text{RKHS}}^\top(\boldsymbol{X}^{(\text{miss})})[\nabla_{\boldsymbol{X}^{(\text{miss})}} \log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})}) \\ &\qquad\qquad + \nabla_{\boldsymbol{X}^{(\text{miss})}} \log \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})]\}\end{aligned} \tag{3-22}$$

另一方面，对于核函数 $K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})})$，可以定义其形式如 $\xi(\boldsymbol{X}^{(\text{miss})})$ 的特征映射（feature map）从而将核函数分解为 $K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})}) = \langle\xi(\boldsymbol{X}^{(\text{miss})\prime}), \xi(\boldsymbol{X}^{(\text{miss})})\rangle_{\mathscr{H}}$。在此基础上，进一步将核函数的谱分解 $K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})}) = \sum_{i=1}^{\infty} \Lambda_i \Xi_i(\boldsymbol{X}^{(\text{miss})\prime})\Xi_i(\boldsymbol{X}^{(\text{miss})})$（其中 Λ_i 和 Ξ_i 分别是特征值和归一化正交基）应用在微扰方向 $\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})$ 上，可以得到 $\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) = \sum_{i=1}^{\infty} \widehat{\psi}_i\sqrt{\Lambda_i}\Xi_i(\boldsymbol{X}^{(\text{miss})})$，其中 $\widehat{\psi}_i$ 为特征重要性权重满足，而特征正交基满足 $\sum_{i=1}^{\infty} \|\widehat{\psi}_i\|_2^2 < \infty$。因此式(3-22)可以重构为下列优化命题：

$$\begin{aligned}\underset{\widehat{\psi}_i\in\mathscr{H}}{\arg\max}\ & -\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\sum_{i=1}^{\infty}\sqrt{\Lambda_i}\nabla_{\boldsymbol{X}^{(\text{miss})}}\log p_\theta(\boldsymbol{X}^{(\text{miss})},\boldsymbol{X}^{(\text{obs})})\widehat{\psi}_i\Xi_i(\boldsymbol{X}^{(\text{miss})})]\\ & +\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\nabla_{\boldsymbol{X}^{(\text{miss})}}\cdot\sum_{i=1}^{\infty}\widehat{\psi}_i\sqrt{\Lambda_i}\Xi_i(\boldsymbol{X}^{(\text{miss})})]-\frac{1}{2}\sum_{i=1}^{\infty}\|\widehat{\psi}_i\|_{\mathscr{H}}^2.\end{aligned}\tag{3-23}$$

在此基础上，最优的特征重要性权重 $\widehat{\psi}_i^*$ 通过式(3-23)对 $\widehat{\psi}_i$ 求偏导，并使之等于 0 进行给出，其具体表达式如下：

$$\widehat{\psi}_i^*=-\sqrt{\Lambda_i}\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}[\nabla_{\boldsymbol{X}^{(\text{miss})}}\Xi(\boldsymbol{X}^{(\text{miss})})+\nabla_{\boldsymbol{X}^{(\text{miss})}}\log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})\Xi(\boldsymbol{X}^{(\text{miss})})].\tag{3-24}$$

将式(5-19)代入 $\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})=\sum_{i=1}^{\infty}\widehat{\psi}_i\sqrt{\Lambda_i}\Xi_i(\boldsymbol{X}^{(\text{miss})})$，可以得到：

$$\begin{aligned}\phi_{\text{RKHS}}^*(\boldsymbol{X}^{(\text{miss})})=&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})\prime})}\{[\nabla_{\boldsymbol{X}^{(\text{miss})\prime}}\log p_\theta(\boldsymbol{X}^{(\text{miss})\prime}|\boldsymbol{X}^{(\text{obs})})]^\top K(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{miss})\prime})\}\\ &-\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})\prime})}[\nabla_{\boldsymbol{X}^{(\text{miss})\prime}}K(\boldsymbol{X}^{(\text{miss})},\boldsymbol{X}^{(\text{miss})\prime})].\end{aligned}\tag{3-25}$$

则式(3-21)得证。　　**证毕**

至此，本小节完成了抑制补全结果发散性的泛函的重构并给出了其在 RKHS 下的解析表达式。需要指出的是，上述补全过程需要获得似然分布 $\log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$ 的得分函数（score function）$\nabla_{\boldsymbol{X}^{(\text{miss})}}\log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$，而在实际工业过程中，缺失值出现的位置并不是固定的，这就导致难以构建合适的函数以逼近条件概率密度。为此，接下来的内容将解决如何修改定理3.3所给出的补全策略，构建适合联合概率密度函数的补全流程，以解决训练目标的一致性问题。

3.4.3　利用联合分布解决一致性问题的建模策略

要使用式(3-21)中的 $\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})})$ 来近似式(3-12)和(3-13)中的映射函数 $\boldsymbol{T}(\boldsymbol{X}^{(\text{miss})})$，关键在于精确估计条件概率密度 $p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$。然而，工业过程中传感器故障的时空随机性导致缺失模式具有高度不确定性，这使得条件概率密度的直接建模面临显著挑战。具体而言，条件分布 $p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$ 的估计精度在很大程度上取决于模型对缺失矩阵 $\boldsymbol{M}$ 的构建方式。在进行模型推理时，理想情况下，推理阶段的缺失矩阵应该与训练阶段保持相同的统计性质。然而，由于缺失值的时空不确定性，推理和训练阶段的缺失模式往往难以完全一致，从而导致模型推理与训练之间的分布存在偏差。为了解决这一问题，本小节将进一步探讨如何用联合概率密度 $p_\theta(\boldsymbol{X}^{(\text{miss})},\boldsymbol{X}^{(\text{obs})})$ 替代式 (3-21)中的

条件概率密度函数 $p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$，以有效规避由“训练目标不一致性”引发的问题。为此，定义符号 $\boldsymbol{X}^{(\text{joint})} \coloneqq (\boldsymbol{X}^{(\text{miss})}, \boldsymbol{X}^{(\text{obs})})$。

根据式(3-21)，可以写出在联合概率密度建模策略下的一个类似的微扰方向为：

$$\begin{aligned}\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{miss})}) = \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})\prime})}\{-\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} K(\boldsymbol{X}^{(\text{joint})}, \boldsymbol{X}^{(\text{joint})\prime}) \\ + \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})\prime})}[\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime}]^\top K(\boldsymbol{X}^{(\text{joint})}, \boldsymbol{X}^{(\text{joint})\prime})\},\end{aligned} \tag{3-26}$$

其中 $\nabla_{\boldsymbol{X}^{(\text{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime})$ 的表达式如下：

$$\nabla_{\boldsymbol{X}^{(\text{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime}) = \nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime}) \odot (\mathbb{1}_{\text{N}\times\text{D}} - \boldsymbol{M}) + 0 \times \boldsymbol{M}. \tag{3-27}$$

为了满足边值条件 $\lim_{\|\boldsymbol{X}^{(\text{miss})}\|\to\infty} \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})}) = 0$，本章采用如下式所定义的径向基函数（radial basis function，RBF）作为核函数 $K(\boldsymbol{X}^{(\text{miss})\prime}, \boldsymbol{X}^{(\text{miss})})$ 的表达式：

$$K(\boldsymbol{X}^{(\text{joint})}, \boldsymbol{X}^{(\text{joint})\prime}) = \exp(-\frac{\|\boldsymbol{X}^{(\text{joint})} - \boldsymbol{X}^{(\text{joint})\prime}\|^2}{\upsilon}), \tag{3-28}$$

其中 υ 为带宽。需要指出的是，$\boldsymbol{X}^{(\text{joint})\prime}$ 和 $\boldsymbol{X}^{(\text{joint})}$ 的取值是相同的，$\boldsymbol{X}^{(\text{joint})\prime}$ 上的撇号 $'$ 仅用于区分梯度算符 ∇ 的作用变量，而梯度算符可以直接使用基于自动微分框架的深度学习后端如 PyTorch[139]和 JAX[140]进行实现。在此基础上，$\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} K(\boldsymbol{X}^{(\text{joint})}, \boldsymbol{X}^{(\text{joint})\prime})$ 的表达式可以给出如式(3-26)所示的形式：

$$\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} K(\boldsymbol{X}^{(\text{joint})}, \boldsymbol{X}^{(\text{joint})\prime}) = \nabla_{\boldsymbol{X}^{(\text{joint})\prime}} K(\boldsymbol{X}^{(\text{joint})}, \boldsymbol{X}^{(\text{joint})\prime}) \odot (\mathbb{1}_{\text{N}\times\text{D}} - \boldsymbol{M}) + 0 \times \boldsymbol{M}. \tag{3-29}$$

在此基础上，可以定义映射 $\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})})$

$$\boldsymbol{T}_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})}) \coloneqq (\boldsymbol{X}^{(\text{joint})}) = \boldsymbol{X}^{(\text{miss})} + \varepsilon\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})}) \tag{3-30}$$

及其对应的泛函：

$$\mathcal{F}^{\text{joint-NER}} \coloneqq \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}[\log p_\theta(\boldsymbol{X}^{(\text{joint})})] - \mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})]. \tag{3-31}$$

根据上述内容，可以给出如下定理说明式(3-31)所定义的泛函 $\mathcal{F}^{\text{joint-NER}}$ 与式(3-10)所定义的泛函 $\mathcal{F}^{\text{NER}}$ 之间的关系：

定理 3.4： 在变分分布 $\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})$ 满足平均场分解假设，即

$$\mathcal{Q}(\boldsymbol{X}^{(\text{joint})}) = \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})p(\boldsymbol{X}^{(\text{obs})}), \tag{3-32}$$

成立时，泛函 $\mathcal{F}^{\text{joint-NER}}$ 与泛函 $\mathcal{F}^{\text{NER}}$ 具有如下关系：

$$\mathcal{F}^{\text{joint-NER}}[\mathcal{Q}] = \mathcal{F}^{\text{NER}}[\mathcal{Q}] - \mathcal{C}, \tag{3-33}$$

其中 $\mathcal{C}$ 为一个非负常数。

在证明该定理前，有必要讨论式(3-32)所给出的平均场假设的合理性。在说明合理性前，需要说明如下客观条件：

（1） 在整个补全过程中，无论对 $\boldsymbol{X}^{(\text{miss})}$ 进行任何修改，$\boldsymbol{X}^{(\text{obs})}$ 都保持不变。

（2） 在客观条件（1）的基础上，从 Liu 等人的论文[141-142]观点而言，$\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})$ 保持不变是合理的，因此 $\mathcal{Q}(\boldsymbol{X}^{(\text{obs})}|\boldsymbol{X}^{(\text{miss})}) = \mathcal{Q}(\boldsymbol{X}^{(\text{obs})})$。这一等式反映了在利用映射 $\boldsymbol{T}(\boldsymbol{X}^{(\text{miss})})/\boldsymbol{T}(\boldsymbol{X}^{(\text{joint})})$ 对缺失数据进行补全过程中 $\boldsymbol{X}^{(\text{obs})}$ 与 $\boldsymbol{X}^{(\text{miss})}$ 之间的独立性。

在此基础上，考察在沃瑟斯坦空间上概率密度函数 $\mathcal{Q}(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})})$ 随着时间 τ 的演化结果可以得到如下的等式关系：

$$\begin{aligned}
&\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})})}{\partial \tau} \\
=&\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})}{\partial \tau} \\
=&\underbrace{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})}{\partial \tau}}_{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})=\frac{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})}|\boldsymbol{X}^{(\text{miss})})\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})}} + \underbrace{\mathcal{Q}(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{obs})})}{\partial \tau}}_{0},
\end{aligned} \tag{3-34}$$

其中第一个下括号是贝叶斯公式，第二个下括号是根据客观条件（2）做出的结论。因此，式(3-34)可以进一步化为：

$$\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})})}{\partial \tau} = \underbrace{\frac{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})}{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})}\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{obs})}|\boldsymbol{X}^{(\text{miss})})\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\partial \tau}}_{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})}|\boldsymbol{X}^{(\text{miss})})=\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})} = \mathcal{Q}(\boldsymbol{X}^{(\text{obs})})\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\partial \tau} \tag{3-35}$$

类似地，当变分分布 $\mathcal{Q}(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})})$ 用 $\mathcal{Q}(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})}) = \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})\mathcal{Q}(\boldsymbol{X}^{(\text{obs})})$ 进行因子化时，可以得到类似的结论：

$$\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{obs})}, \boldsymbol{X}^{(\text{miss})})}{\partial \tau} = \frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{obs})})\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\partial \tau} = \mathcal{Q}(\boldsymbol{X}^{(\text{obs})})\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\partial \tau}. \tag{3-36}$$

根据上述分析，式(3-32)的合理性得到了说明。

在说明式(3-32)的合理性的基础上，接下来将对定3.4给出证明：

证明：证明过程将分为两个部分：微扰方向导出和泛函等价性证明，微扰方向导出的目的是为了证明 $\mathcal{F}^{\text{joint-NER}}$ 能否导出 $\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})})$，而泛函等价性证明则是为了证明 $\mathcal{F}^{\text{joint-NER}}$ 和 $\mathcal{F}^{\text{NER}}$ 之间的关系。

- **微扰方向导出：**在最开始，需要证明泛函 $\mathcal{F}^{\text{joint-NER}}$ 能导出微扰方向 $\phi(\boldsymbol{X}^{(\text{joint})})$。根据引理2.1和式(3-34)，下列等式成立：

$$\begin{aligned}
&\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\partial \tau} = -\nabla_{\boldsymbol{X}^{(\text{miss})}} \cdot [\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})\phi(\boldsymbol{X}^{(\text{miss})})] \\
\Rightarrow &\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}{\partial \tau} \times p(\boldsymbol{X}^{(\text{obs})}) = -\nabla_{\boldsymbol{X}^{(\text{miss})}} \cdot [\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})\phi(\boldsymbol{X}^{(\text{miss})})] \times p(\boldsymbol{X}^{(\text{obs})}) \\
\overset{\text{(i)}}{\Rightarrow} &\frac{\partial \mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}{\partial \tau} = -\nabla_{\boldsymbol{X}^{(\text{miss})}} \cdot [\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})\phi(\boldsymbol{X}^{(\text{joint})})],
\end{aligned} \tag{3-37}$$

当微扰方向 $\phi(\boldsymbol{X}^{(\text{joint})})$ 处于RKHS（记为 $\mathscr{H}^{\text{D}}$），并且其核函数 $K(\boldsymbol{X}^{(\text{joint})\prime}, \boldsymbol{X}^{(\text{joint})})$ 满足边界条件 $\lim_{\|\boldsymbol{X}^{(\text{joint})}\|\to\infty} \mathcal{Q}(\boldsymbol{X}^{(\text{joint})})K(\boldsymbol{X}^{(\text{joint})\prime}, \boldsymbol{X}^{(\text{joint})}) = 0$ 时，定义形式如 $\xi(\boldsymbol{X}^{(\text{joint})})$ 的特征映射将核函数分解为 $K(\boldsymbol{X}^{(\text{joint})\prime}, \boldsymbol{X}^{(\text{joint})}) = \langle \xi(\boldsymbol{X}^{(\text{joint})\prime}), \xi(\boldsymbol{X}^{(\text{joint})}) \rangle_{\mathscr{H}^{\text{D}}}$。紧接着，将核函数的谱分解 $K(\boldsymbol{X}^{(\text{joint})\prime}, \boldsymbol{X}^{(\text{joint})}) = \sum_{i=1}^{\infty} \Lambda_i \Xi_i(\boldsymbol{X}^{(\text{joint})\prime}) \Xi_i(\boldsymbol{X}^{(\text{joint})})$（其中 Λ_i 和 Ξ_i 分别是特征值和归一化正交基）应用在微扰方向 $\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})})$ 上，可以得到 $\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})}) = \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\Lambda_i} \Xi_i(\boldsymbol{X}^{(\text{joint})})$，其中 $\widehat{\psi}_i$ 为特征重要性权重满足，而特征正交基满足 $\sum_{i=1}^{\infty} \|\widehat{\psi}_i\|_2^2 < \infty$。因此可以构造下列优化命题：

$$\begin{aligned}
&\arg\max_{\phi(\boldsymbol{X}^{(\text{joint})}) \in \mathcal{H}^{\text{D}}} \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}[\phi^{\top}(\boldsymbol{X}^{(\text{joint})}) \nabla_{\boldsymbol{X}^{(\text{miss})}} \log p_{\theta}(\boldsymbol{X}^{(\text{joint})})] \\
&\qquad\qquad - \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}[\nabla_{\boldsymbol{X}^{(\text{miss})}} \cdot \phi(\boldsymbol{X}^{(\text{joint})})] - \frac{1}{2}\|\phi(\boldsymbol{X}^{(\text{joint})})\|_{\mathcal{H}^{\text{D}}}^2, \\
\Rightarrow &\arg\max_{\phi(\boldsymbol{X}^{(\text{joint})}) \in \mathcal{H}^{\text{D}}} \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})\prime})}[\sum_{i=1}^{\infty} \sqrt{\xi_i} \nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p_{\theta}(\boldsymbol{X}^{(\text{joint})\prime})^{\top} \widehat{\psi}_i \Xi_i(\boldsymbol{X}^{(\text{joint})\prime})] \\
&\qquad\qquad - \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})\prime})}[\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \cdot \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\xi_i} \Xi_i(\boldsymbol{X}^{(\text{joint})\prime})] - \frac{1}{2}\sum_{i=1}^{\infty} \|\widehat{\psi}_i\|_2^2,
\end{aligned} \tag{3-38}$$

在此基础上，最优的特征重要性权重 $\widehat{\psi}_i^*$ 通过式(3-38)对 $\widehat{\psi}_i$ 求偏导，并使之等于0进行给出，其具体表达式如下：

$$\widehat{\psi}_i^* = -\sqrt{\Lambda_i} \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}[\nabla_{\boldsymbol{X}^{(\text{miss})}} \Xi(\boldsymbol{X}^{(\text{joint})}) + \nabla_{\boldsymbol{X}^{(\text{miss})}} \log p_{\theta}(\boldsymbol{X}^{(\text{joint})}) \Xi(\boldsymbol{X}^{(\text{joint})})]. \tag{3-39}$$

将式(3-39)代入 $\phi_{\mathrm{RKHS}}(\boldsymbol{X}^{(\mathrm{joint})}) = \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\Lambda_i} \Xi_i(\boldsymbol{X}^{(\mathrm{joint})})$，可以得到：

$$\begin{aligned}\phi^*_{\mathrm{RKHS}}(\boldsymbol{X}^{(\mathrm{joint})}) =& \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}\{[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})\prime})]^\top K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})\} \\ & - \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})].\end{aligned} \tag{3-40}$$

则泛函 $\mathcal{F}^{\text{joint-NER}}$ 诱导映射 $\boldsymbol{T}(\boldsymbol{X}^{(\mathrm{joint})})$ 的合理性得证。

- **泛函等价性证明：**首先给出 $\mathcal{F}^{\text{joint-NER}}$ 的表达式：

$$\underbrace{\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})}[\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})]}_{:=\text{term 1}} + \{\underbrace{-\mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})]}_{:=\text{term 2}}\}. \tag{3-41}$$

针对“term 1”有下式所示的推导：

$$\begin{aligned}&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})}[\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})] \\ \geq& \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})}[\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}] + \underbrace{\mathbb{E}_{p(\boldsymbol{X}^{(\mathrm{obs})})}[\log p(\boldsymbol{X}^{(\mathrm{obs})})]}_{\text{负熵（负常数）}} \\ =& \iint_{\mathbb{R}^{\mathrm{D}}} p(\boldsymbol{X}^{(\mathrm{obs})})\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})}) \log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}\mathrm{d}\boldsymbol{X}^{(\mathrm{obs})} \\ &+ \underbrace{\mathbb{E}_{p(\boldsymbol{X}^{(\mathrm{obs})})}[\log p(\boldsymbol{X}^{(\mathrm{obs})})]}_{\text{负熵（负常数）}} \\ =& \iint_{\mathbb{R}^{\mathrm{D}}} p(\boldsymbol{X}^{(\mathrm{obs})})\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})}) \log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})})\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}\mathrm{d}\boldsymbol{X}^{(\mathrm{obs})} \\ &+ \underbrace{\iint_{\mathbb{R}^{\mathrm{D}}} \mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})p(\boldsymbol{X}^{(\mathrm{obs})}) \log p(\boldsymbol{X}^{(\mathrm{obs})})\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}\mathrm{d}\boldsymbol{X}^{(\mathrm{obs})}}_{\text{负熵（负常数）}} \\ =& \iint_{\mathbb{R}^{\mathrm{D}}} \underbrace{p(\boldsymbol{X}^{(\mathrm{obs})})\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})},\boldsymbol{X}^{(\mathrm{obs})})}\underbrace{[\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}|\boldsymbol{X}^{(\mathrm{obs})}) + \log p(\boldsymbol{X}^{(\mathrm{obs})})]}_{\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})},\boldsymbol{X}^{(\mathrm{obs})})}\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}\mathrm{d}\boldsymbol{X}^{(\mathrm{obs})} \\ =& \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})},\boldsymbol{X}^{(\mathrm{obs})})}[\log p_\theta(\boldsymbol{X}^{(\mathrm{miss})}, \boldsymbol{X}^{(\mathrm{obs})})].\end{aligned} \tag{3-42}$$

类似地，针对“term 2”有下式所示的推导：

$$
\begin{aligned}
&-\mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})] \\
\geq &-\mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})]+\underbrace{\mathbb{E}_{p(\boldsymbol{X}^{(\text{obs})})}[\log p(\boldsymbol{X}^{(\text{obs})})]}_{\text{负熵（负常数）}} \\
=&\iint_{\mathbb{R}^{\mathrm{D}}} p(\boldsymbol{X}^{(\text{obs})})\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})\log\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})\mathrm{d}\boldsymbol{X}^{(\text{miss})}\mathrm{d}\boldsymbol{X}^{(\text{obs})} \\
&+\underbrace{\iint_{\mathbb{R}^{\mathrm{D}}} p(\boldsymbol{X}^{(\text{obs})})\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})\log p(\boldsymbol{X}^{(\text{obs})})\mathrm{d}\boldsymbol{X}^{(\text{miss})}\mathrm{d}\boldsymbol{X}^{(\text{obs})}}_{\text{负熵（负常数）}} \\
=&\iint_{\mathbb{R}^{\mathrm{D}}} \underbrace{p(\boldsymbol{X}^{(\text{obs})})\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}_{\mathcal{Q}(\boldsymbol{X}^{(\text{obs})},\boldsymbol{X}^{(\text{miss})})}\underbrace{[\log\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})+\log p(\boldsymbol{X}^{(\text{obs})})]}_{\log\mathcal{Q}(\boldsymbol{X}^{(\text{obs})},\boldsymbol{X}^{(\text{miss})})}\mathrm{d}\boldsymbol{X}^{(\text{miss})}\mathrm{d}\boldsymbol{X}^{(\text{obs})} \\
=&-\mathbb{H}[\mathcal{Q}(\boldsymbol{X}^{(\text{obs})},\boldsymbol{X}^{(\text{miss})})].
\end{aligned}
\tag{3-43}
$$

结合式(3-42)和式(3-43)，可以得到下列等式：

$$
\mathcal{F}^{\text{NER}}-\mathcal{C}=\mathcal{F}^{\text{joint-NER}}, \tag{3-44}
$$

其中 $\mathcal{C}\geq 0$，则定理3.4得证。

证毕

在定理3.4的基础上，可以得到如下的定理；

定理 3.5： 基于定理3.4的结论，映射 $\boldsymbol{T}(\cdot)$ 在缺失数据空间与联合数据空间上具有一致性，即成立：

$$
\boldsymbol{T}(\boldsymbol{X}^{(\text{miss})})=\boldsymbol{T}(\boldsymbol{X}^{(\text{joint})}). \tag{3-45}
$$

证明： 由于 $\boldsymbol{T}(\boldsymbol{X})=\boldsymbol{X}+\varepsilon\phi_{\text{RKHS}}(\boldsymbol{X})$，而 $\phi_{\text{RKHS}}(\boldsymbol{X})$ 取决于泛函 $\nabla_{\boldsymbol{X}}\mathcal{F}[\mathcal{Q}(\boldsymbol{X})]$，因此，仅需验证 $\nabla_{\boldsymbol{X}^{(\text{miss})}}\dfrac{\delta\mathcal{F}^{\text{NER}}}{\delta\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}$ 与 $\nabla_{\boldsymbol{X}^{(\text{miss})}}\dfrac{\delta\mathcal{F}^{\text{joint-NER}}}{\delta\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}$ 是否相等即可。为此，可以给出下式所示的推导：

$$
\begin{aligned}
&\mathcal{F}^{\text{NER}}=\mathcal{F}^{\text{joint-NER}}+\mathcal{C} \\
\Rightarrow&\nabla_{\boldsymbol{X}^{(\text{miss})}}\frac{\delta\mathcal{F}^{\text{NER}}}{\delta\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}=\nabla_{\boldsymbol{X}^{(\text{miss})}}\frac{\delta\mathcal{F}^{\text{joint-NER}}+\text{const}}{\delta\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})} \\
\Rightarrow&\nabla_{\boldsymbol{X}^{(\text{miss})}}\frac{\delta\mathcal{F}^{\text{NER}}}{\delta\mathcal{Q}(\boldsymbol{X}^{(\text{miss})})}=\nabla_{\boldsymbol{X}^{(\text{miss})}}\frac{\delta\mathcal{F}^{\text{joint-NER}}}{\delta\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}.
\end{aligned}
\tag{3-46}
$$

则式(3-45)得证。 **证毕**

至此，本小节通过证明条件概率密度泛函诱导的映射与联合概率密度泛函诱导的映射的等价性证明了在最小移动方案框架下通过建模联合概率密度函数即可等价地建模条件概率密度函数，进而规避模型在训练和推理过程的不一致性问题。

3.4.4 补全算法总结和收敛性证明

尽管第3.4.2小节和第3.4.3小节分别通过负熵泛函诱导的最小移动方案框架和联合建模的策略解决了补全结果的发散性问题和训练目标的不一致问题，但是并没有给出完整的数据补全流程。为此，本小节将在第3.4.2小节和第3.4.3小节的基础上给出完整的数据补全算法流程和相关的收敛性证明以为实际工业过程缺失数据补全提供指导。

值得注意的是，联合空间映射 $\boldsymbol{T}(\boldsymbol{X}^{(\text{joint})})$ 的核心计算瓶颈在于 $\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime})$ 的表达式尚未确定。为此，本小节将首先给出 $\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime})$ 的训练方法并在基础上给出完整的缺失值补全算法流程。

本小节采用了去噪得分匹配（denoising score matching，DSM）方法[143-144]对得分函数进行训练，其中 θ 记为神经网络参数，而 $\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p_\theta(\boldsymbol{X}^{(\text{joint})\prime})$ 表示用以 θ 为参数的神经网络对得分函数 $\nabla_{\boldsymbol{X}^{(\text{joint})\prime}} \log p(\boldsymbol{X}^{(\text{joint})\prime})$ 进行参数化。在此基础上，DSM 通过引入高斯噪声 ϵ 对观测样本 $\boldsymbol{X}^{(\text{joint})}$ 进行扰动，生成噪声样本 $\widehat{\boldsymbol{X}}^{(\text{joint})} = \boldsymbol{X}^{(\text{joint})} + \epsilon, \epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathcal{I})$，并以此构造学习目标以最小化模型预测得分与真实得分之间的差异：

$$\mathcal{L}^{\text{DSM}} := \frac{1}{2}\mathbb{E}_{q_\sigma(\hat{\boldsymbol{X}}^{(\text{joint})}|\boldsymbol{X}^{(\text{joint})})}[\|\nabla_{\hat{\boldsymbol{X}}^{(\text{joint})}} \log p_\theta(\widehat{\boldsymbol{X}}^{(\text{joint})}) - \nabla_{\hat{\boldsymbol{X}}^{(\text{joint})}} \log q_\sigma(\widehat{\boldsymbol{X}}^{(\text{joint})}|\boldsymbol{X}^{(\text{joint})})\|_2^2]. \quad (3\text{-}47)$$

在上述公式中，$q_\sigma(\widehat{\boldsymbol{X}}^{(\text{joint})}|\boldsymbol{X}^{(\text{joint})})$ 表示高斯扰动密度分布，其得分函数解析表达如下：

$$\nabla_{\hat{\boldsymbol{X}}^{(\text{joint})}} \log q_\sigma(\widehat{\boldsymbol{X}}^{(\text{joint})}|\boldsymbol{X}^{(\text{joint})}) = -\frac{\widehat{\boldsymbol{X}}^{(\text{joint})} - \boldsymbol{X}^{(\text{joint})}}{\sigma^2}. \quad (3\text{-}48)$$

在此基础上，参数 θ 的训练流程在算法3.1中进行总结，其中 HU_{score} 是隐层单元（hidden unit）的英文首字母缩写。

在完成得分函数 $\nabla_{\boldsymbol{X}^{(\text{joint})}} \log p_\theta(\boldsymbol{X}^{(\text{joint})})$ 的训练后，本小节进一步总结了基于隐变量模型的数据缺失值补全方法，并将其具体实现归纳为算法3.2所示。这一方法通过在代价泛函中引入负熵泛函正则项，并结合 RKHS 中的最小移动方案框架，以沃瑟斯坦距离为理论基础进行导出。基于其核心思想和技术特点，本章将该方法命名为基于核负熵正则化沃瑟斯坦距离的填补（Kernelized Negative Entropy-regularized Wasserstein distance-

算法 3.1 基于 DSM 的得分函数训练算法伪代码

输入： 联合数据：$\boldsymbol{X}^{(\text{joint})}$、神经网络学习率：$\eta$、训练轮次：$\mathcal{E}$ 和神经网络隐藏单元数：HU_{score}。
参数： 神经网络参数 θ。
输出： 训练好的神经网络 $\nabla_{\boldsymbol{X}^{(\text{joint})}} \log p_\theta(\boldsymbol{X}^{(\text{joint})})$。

1: **for** $e = 1$ to $\mathcal{E}$ **do**
2: $\quad \widehat{\boldsymbol{X}}^{(\text{joint})} \leftarrow \boldsymbol{X}^{(\text{joint})} + \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathcal{I})$ > 数据加噪
3: $\quad \nabla_{\widehat{\boldsymbol{X}}^{(\text{joint})}} \log q_\sigma(\widehat{\boldsymbol{X}}^{(\text{joint})} | \boldsymbol{X}^{(\text{joint})}) \leftarrow -\dfrac{\widehat{\boldsymbol{X}}^{(\text{joint})} - \boldsymbol{X}^{(\text{joint})}}{\sigma^2}$
4: $\quad \mathcal{L}^{\text{DSM}} \leftarrow$ 式(3-47)
5: $\quad \theta_{e+1} \leftarrow \theta_e - \eta \nabla_\theta \mathcal{L}^{\text{DSM}}|_{\theta=\theta_e}$ > 以学习率 η 进行参数更新
6: **end for**
7: **return** $\nabla_{\boldsymbol{X}^{(\text{joint})}} \log p_\theta(\boldsymbol{X}^{(\text{joint})})$。

based Imputation，KnewImp）算法；而图3.2给出了 KnewImp 算法的示意图。根据图3.2，本章并列出以下要点对算法的几个主要阶段进行详细说明：

（1） **数据初始化（算法3.2第 1 行）：**首先，对输入的观测数据矩阵 $\boldsymbol{X}^{(\text{obs})}$ 进行均值补全，生成初始化的预补全数据集 $\boldsymbol{X}^{(\text{imp})}$。

（2） **DSM 训练与泛函 $\mathcal{L}^{\text{joint-NER}}$ 优化（算法3.2第 2-9 行）：**在每次迭代中，算法交替进行以下两个核心步骤：

- **DSM 训练（算法3.2第 3-5 行）：**此阶段被称为“训练”步。在此阶段，首先构造联合数据矩阵 $\boldsymbol{X}^{(\text{joint})}$，其由当前的缺失值估计 $\boldsymbol{X}^{(\text{miss})}$ 和观测值 $\boldsymbol{X}^{(\text{obs})}$ 组成。然后，基于 $\boldsymbol{X}^{(\text{joint})}$ 以算法3.1为基础进行训练，以获得 $\nabla_{\boldsymbol{X}^{(\text{joint})}} \log p_\theta(\boldsymbol{X}^{(\text{joint})})$。
- **泛函 $\mathcal{L}^{\text{joint-NER}}$ 优化（算法3.2第 7-11 行）：**此阶段被称为“补全”步。在此阶段，首先基于第 τ 次优化的 $\boldsymbol{X}_\tau^{(\text{miss})}$ 构造联合数据矩阵 $\boldsymbol{X}^{(\text{joint})}$。然后，基于 $\boldsymbol{X}^{(\text{joint})}$ 确定微扰方向 $\phi(\boldsymbol{X}^{(\text{joint})})$ 和映射方向 $\boldsymbol{T}(\boldsymbol{X}^{(\text{joint})})$，并更新得到第 $\tau+1$ 次优化的 $\boldsymbol{X}_{\tau+1}^{(\text{miss})}$。

（3） **最终补全结果生成（算法3.2第 14 行）：**在完成 $\mathcal{T}$ 次迭代后，算法将观测值 $\boldsymbol{X}^{(\text{obs})}$ 与最终补全的缺失值 $\boldsymbol{X}^{(\text{imp})}$ 进行合并生成完整的补全数据集 $\widehat{\boldsymbol{X}}$，其中观测值 $\boldsymbol{X}^{(\text{obs})}$ 保持不变。

在算法3.2的基础上，本小节将在接下来的内容对其收敛性进行理论分析和证明。参照传统 EM 算法的收敛性定义[110]，本文的收敛性定义如下：

算法 3.2 KnewImp 算法伪代码

输入： 观测数据 $\boldsymbol{X}^{(\text{obs})}$、遮罩矩阵 $\boldsymbol{M}$、循环次数 $\mathscr{T}$、优化迭代轮次 T、网络训练轮次 $\mathcal{E}$、网络学习率 η、神经网络隐层单元 HU_{score}、带宽 v 和微扰大小 ε。
参数： 神经网络参数 θ。
输出： 补全数据 $\hat{\boldsymbol{X}}$。

$\boldsymbol{X}^{(\text{imp})} \leftarrow \text{Initialize}(\boldsymbol{X}^{(\text{obs})}) \odot (\mathbb{1}_{\text{N}\times\text{D}} - \boldsymbol{M})$　　> 预补全数据以进行初始化
for $t \leftarrow 0$ **to** $\mathscr{T} - 1$ **do**
　$\boldsymbol{X}^{(\text{miss})} \leftarrow \boldsymbol{X}^{(\text{imp})}$
　$\boldsymbol{X}^{(\text{joint})} \leftarrow \boldsymbol{X}^{(\text{miss})} \odot (\mathbb{1}_{\text{N}\times\text{D}} - \boldsymbol{M}) + \boldsymbol{X}^{(\text{obs})} \odot \boldsymbol{M}$
　$\nabla_{\boldsymbol{X}^{(\text{joint})}} \log p_\theta(\boldsymbol{X}^{(\text{joint})}) \leftarrow$ 算法3.1　　> “训练”步
　$\boldsymbol{X}_0^{(\text{miss})} \leftarrow \boldsymbol{X}^{(\text{imp})}$
　for $\tau \leftarrow 0$ **to** $\text{T} - 1$ **do**
　　$\boldsymbol{X}^{(\text{joint})} \leftarrow \boldsymbol{X}_\tau^{(\text{miss})} \odot (\mathbb{1}_{\text{N}\times\text{D}} - \boldsymbol{M}) + \boldsymbol{X}^{(\text{obs})} \odot \boldsymbol{M}$
　　$\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})}) \leftarrow$ 式(3-26)
　　$\boldsymbol{X}_{\tau+1}^{(\text{miss})} \leftarrow \boldsymbol{T}(\boldsymbol{X}^{(\text{joint})}) = \boldsymbol{X}^{(\text{miss})} + \varepsilon\phi_{\text{RKHS}}(\boldsymbol{X}^{(\text{joint})})$　　> “补全”步
　end for
　$\boldsymbol{X}^{(\text{imp})} \leftarrow \boldsymbol{X}_{\text{T}}^{(\text{miss})}$
end for
$\widehat{\boldsymbol{X}} \leftarrow \boldsymbol{X}^{(\text{obs})} \odot \boldsymbol{M} + \boldsymbol{X}^{(\text{imp})} \odot (\mathbb{1}_{\text{N}\times\text{D}} - \boldsymbol{M})$
return $\widehat{\boldsymbol{X}}$

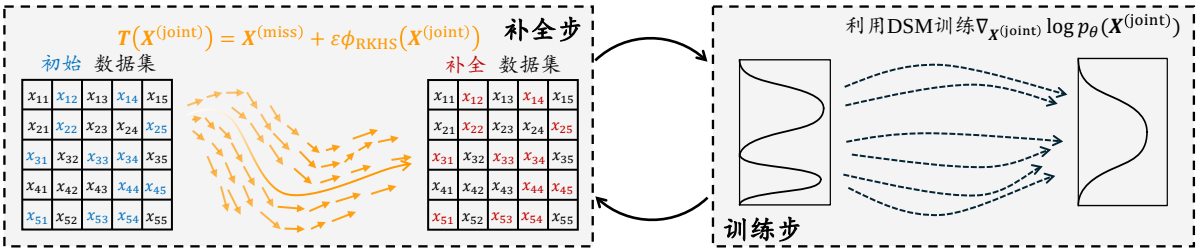


图 3.2 KnewImp 算法示意图

定义 1： 一个数列 $\{a_1, a_2, \ldots, a_\mathcal{E}\}$ 被认为是收敛的，当且仅当存在一个实数 $\mathcal{R}$，使得对于任意给定的 $\gamma > 0$，都存在一个正整数 N，使得对于所有 $n \geq N$，数列的项 a_n 始终满足不等式 $|a_n - \mathcal{R}| < \gamma$。

该收敛性的证明过程可以通过单调收敛定理（文献[145]中的定理 2.14）进行实现。换言之，如果某个数列是单调递增或递减，并且该数列受到某个常数的上界或下界限制，则该数列收敛。在该定义的基础上，接下来的内容将分别针对 KnewImp 算法的“补全”步与“训练”步的收敛性进行证明：

定理 3.6： 设 $\{\mathcal{F}_\tau^{\text{joint-NER}}\}_{\tau=1}^{\text{T}}$ 为 KnewImp 算法“补全”步生成的迭代序列。若微扰大小 ε 充分小，则存在 $\mathcal{F}^{\text{joint-NER}*}$，使得对于任意给定的 $\gamma > 0$, 存在正整数 N 使得当 $\tau > N$ 时有 $\|\mathcal{F}_\tau^{\text{joint-NER}} - \mathcal{F}^{\text{joint-NER}*}\| < \gamma$ 成立。

证明：首先将微扰方向重构为下式：

$$
\begin{aligned}
&\phi_{\mathrm{RKHS}}(\boldsymbol{X}^{(\mathrm{joint})}) \\
=&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}[-\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})] \\
&\quad + \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}\{[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})\prime})]^\top K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})\} \\
\overset{\text{(i)}}{=}&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}\{[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})]^\top K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})\} \\
&\quad + \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}\{[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})\prime})]^\top K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})\} \\
=&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}\{[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime}) + \nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})\prime})]^\top K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})\}
\end{aligned}
\tag{3-49}
$$

其中步骤“(i)”基于分部积分（integration by parts）法导出。

$$
\begin{aligned}
&\frac{\mathrm{d}\mathcal{F}^{\text{joint-NER}}}{\mathrm{d}\tau} \\
=&\int_{\mathbb{R}^{\mathrm{D}}} -\{\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \cdot [\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) u(\boldsymbol{X}^{(\mathrm{joint})})]\}[\log p_\theta(\boldsymbol{X}^{(\mathrm{joint})}) + \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) + 1]\mathrm{d}\boldsymbol{X}^{(\mathrm{joint})} \\
\overset{\text{(ii)}}{=}&\int_{\mathbb{R}^{\mathrm{D}}} [\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) \phi(\boldsymbol{X}^{(\mathrm{joint})})]^\top \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} [\log p_\theta(\boldsymbol{X}^{(\mathrm{joint})}) + \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) + 1]\mathrm{d}\boldsymbol{X}^{(\mathrm{joint})} \\
=&\int_{\mathbb{R}^{\mathrm{D}}} [\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) \phi(\boldsymbol{X}^{(\mathrm{joint})})]^\top \{\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} [\log p_(\boldsymbol{X}^{(\mathrm{joint})}) + \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})})]\}\mathrm{d}\boldsymbol{X}^{(\mathrm{joint})} \\
=&\int_{\mathbb{R}^{\mathrm{D}}} [\phi(\boldsymbol{X}^{(\mathrm{joint})})]^\top [\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})}) \\
&\qquad + \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})}) \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})})]\mathrm{d}\boldsymbol{X}^{(\mathrm{joint})} \\
=&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})})}\{\phi^\top(\boldsymbol{X}^{(\mathrm{joint})})[\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})}) + \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})})]\},
\end{aligned}
\tag{3-50}
$$

其中步骤“(ii)”由分部积分法导出。将式(3-49)代入式(3-49)可以得到如下等式：

$$
\begin{aligned}
&\frac{\mathrm{d}\mathcal{F}_{\text{joint-NER}}}{\mathrm{d}\tau} \\
=&\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})})} \mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime})}\{[\nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})\prime}) + \nabla_{\boldsymbol{X}^{(\mathrm{miss})\prime}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})\prime})]^\top \\
&\qquad \times K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})[\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log p_\theta(\boldsymbol{X}^{(\mathrm{joint})}) + \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log \mathcal{Q}(\boldsymbol{X}^{(\mathrm{joint})})]\} \\
\overset{\text{(iii)}}{\geq}&0
\end{aligned}
\tag{3-51}
$$

其中步骤“(iii)”由核函数 $K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime})$ 的半正定性属性，即 $K(\boldsymbol{X}^{(\mathrm{joint})}, \boldsymbol{X}^{(\mathrm{joint})\prime}) \succeq 0$ 导出。因此，根据上述不等式，可以得出 $\mathcal{F}^{\text{joint-NER}}$ 随着 τ 的演化单调递增的结论，而 ε 越小则说明 $\mathcal{F}^{\text{joint-NER}}$ 离散得到的迭代序列越接近式(3-51)所定义的 ODE。

在此基础上，将证明 $\mathcal{F}^{\text{joint-NER}}$ 的有界性。注意到，$\mathcal{F}^{\text{joint-NER}}$ 满足该不等式：

$$\begin{aligned}&\mathcal{F}^{\text{joint-NER}}\\ \leq&\mathcal{F}^{\text{joint-NER}}-2\mathbb{E}_{\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})}[\log\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})]\\ =&-\mathbb{D}_{\text{KL}}\left[\mathcal{Q}(\boldsymbol{X}^{(\text{joint})})\|p_\theta(\boldsymbol{X}^{(\text{joint})})\right]\\ \leq&0.\end{aligned}\tag{3-52}$$

式(3-52)说明，泛函 $\mathcal{F}^{\text{joint-NER}}$ 的一个上界是 0，则根据单调有界定理，利用 $\boldsymbol{T}(\boldsymbol{X}^{(\text{joint})})$ 优化泛函 $\mathcal{F}^{\text{joint-NER}}$ 而产生的迭代序列收敛性得证。 **证毕**

定理 3.7： 设 $\{\mathcal{L}_\tau^{\text{DSM}}\}_{\tau=1}^{\text{T}}$ 为 KnewImp 算法“训练”步生成的迭代序列。若微扰大小 ε 充分小，则存在 $\mathcal{L}^{\text{DSM}*}$，使得对于任意给定的 $\gamma>0$, 存在正整数 N 使得当 $\tau>N$ 时有 $\|\mathcal{L}_\tau^{\text{DSM}}-\mathcal{L}^{\text{DSM}*}\|<\gamma$ 成立。

证明： 在最开始，回顾基于梯度下降法更新的 DSM 的训练过程：

$$\theta_{\tau+1}=\theta_\tau-\eta\times\nabla_\theta\mathcal{L}^{\text{DSM}}|_{\theta=\theta_\tau}.\tag{3-53}$$

在此基础上可以得到下式所示的 ODE：

$$\begin{aligned}&\frac{\theta_{\tau+1}-\theta_\tau}{\eta}=-\nabla_\theta\mathcal{L}^{\text{DSM}}|_{\theta=\theta_\tau}\\ \Rightarrow&\lim_{\eta\to0}\frac{\theta_{\tau+1}-\theta_\tau}{\eta}=-\nabla_\theta\mathcal{L}^{\text{DSM}}|_{\theta=\theta_\tau}\\ \Rightarrow&\frac{\mathrm{d}\theta}{\mathrm{d}\tau}=-\nabla_\theta\mathcal{L}^{\text{DSM}}.\end{aligned}\tag{3-54}$$

与此同时，注意到：

$$\frac{\mathrm{d}\mathcal{L}_{\text{DSM}}}{\mathrm{d}\tau}=\left\langle\nabla_\theta\mathcal{L}^{\text{DSM}},\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right\rangle.\tag{3-55}$$

将式(3-54)代入式(3-55)可以得到下列结果：

$$\frac{\mathrm{d}\mathcal{L}^{\text{DSM}}}{\mathrm{d}\tau}=-\left\langle\nabla_\theta\mathcal{L}^{\text{DSM}},\nabla_\theta\mathcal{L}^{\text{DSM}}\right\rangle\leq0,\tag{3-56}$$

上述不等式表面 $\mathcal{L}_{\text{DSM}}$ 随着 τ 的演化是单调递减的，而学习率 η 越小，则表明产生的迭代序列越接近 $\dfrac{\mathrm{d}\mathcal{L}^{\text{DSM}}}{\mathrm{d}\tau}$。

最后，回顾式(3-47)可知下列不等式：

$$\mathcal{L}_{\text{DSM}}\geq0.\tag{3-57}$$

则泛函 $\mathcal{L}_{\text{DSM}}$ 的一个下界是 0，则根据单调有界定理，利用梯度下降方法在 DSM 框架下优化泛函 $\mathcal{L}_{\text{DSM}}$ 而产生的迭代序列收敛性得证。 **证毕**

3.5 实验验证

在本节，将对所提出的 KnewImp 算法的有效性进行实验验证。本节的实验将从如下五个研究问题进行展开：

- **问题 1（有效性）：**KnewImp 算法能否对满足不同类型的分布数据进行补全？
- **问题 2（表现性）：**KnewImp 算法相比其他数据补全方法在工业过程数据集的缺失数据补全任务上表现如何？
- **问题 3（增益机理）：**是什么让 KnewImp 算法在工业过程数据集的缺失数据补全任务上取得了如此好的表现效果？
- **问题 4（敏感性）：**在超参数产生变动的情况下，KnewImp 算法的性能会产生什么样的变化？
- **问题 5（收敛性）：**KnewImp 算法能否收敛？

鉴于本章的研究重点为软测量建模前置步骤——缺失数据补全问题，因此区别于式(2-37a)至式(2-37d)给出的指标，本章选取了数据补全领域常用的评价指标——掩码平均绝对误差（masked mean absolute error，mMAE）和掩码 2-沃瑟斯坦距离（masked 2-Wasserstein distance，mWASS）作为评估标准，其基本定义分别见式(3-58a)和式(3-58b)：

$$\text{mMAE} \coloneqq \frac{\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{D}}[|\boldsymbol{X}_{i,j}^{(\text{ideal})}-\hat{\boldsymbol{X}}_{i,j}|\odot(\mathbb{1}_{\mathrm{N}\times\mathrm{D}}-\boldsymbol{M})_{i,j}]}{\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{D}}(\mathbb{1}_{\mathrm{N}\times\mathrm{D}}-\boldsymbol{M})_{i,j}}, \tag{3-58a}$$

$$\text{mWASS} \coloneqq \mathcal{W}_2^2[\frac{1}{\boldsymbol{m}_1}\sum_{i=1}^{\boldsymbol{m}_1}\delta_{[\hat{\boldsymbol{X}}_{\boldsymbol{M}_1}]_{i,:}},\frac{1}{\boldsymbol{m}_1}\sum_{i=1}^{\boldsymbol{M}_1}\delta_{[\boldsymbol{X}_{\boldsymbol{M}_1}^{(\text{ideal})}]_{i,:}}], \tag{3-58b}$$

其中，$\boldsymbol{M}_1 \coloneqq \{i:\exists j,\boldsymbol{M}_{i,j}=0\}$ 表示在 $\boldsymbol{M}_{i,j}$ 中至少有一个缺失值的集合，$\boldsymbol{m}_1$ 为至少存在一个缺失值的数据点的数量，而 $\delta_{\boldsymbol{X}}$ 表示集中于 $\boldsymbol{X}$ 的狄拉克测度（Dirac measure）。

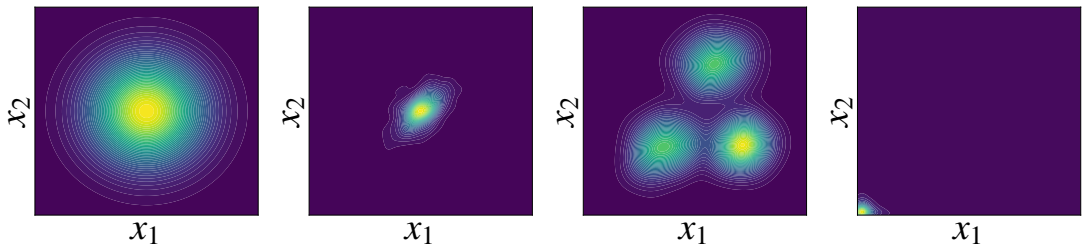
3.5.1 二维模拟案例分析

本小节旨在研究**问题 1**：“KnewImp 算法能否对满足不同类型的分布数据进行补全”。为系统评估算法在工业过程数据不同分布特性下的适应性，本研究设计了基于四

种典型二维分布的数据补全实验。考虑到工业过程中操作模式变化可能导致数据呈现不同的分布特征，实验特别选取了以下具有代表性的分布类型（每种分布生成 1000 组实验数据）：

- **标准高斯分布：** $\boldsymbol{X}^{(\text{ideal})} \sim \mathcal{N}(0, \mathcal{I})$；
- **学生-t 分布（重尾分布）：** $boldsymbolX^{(\text{ideal})} \sim St(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$；
- **混合高斯分布（多峰分布）：** $\boldsymbol{X}^{(\text{ideal})} \sim \frac{1}{3}\times\mathcal{N}([1,2], \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix})+\frac{1}{3}\times\mathcal{N}([-1,-2], \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix})+\frac{1}{3}\times\mathcal{N}([2,-2], \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix})$；
- **斜偏高斯分布（斜偏分布）：** $\boldsymbol{X}^{(\text{ideal})} = \exp(\epsilon), \epsilon \sim \mathcal{N}(0, \mathcal{I})$。

上述四种分布的概率密度函数等高线图分别如图3.3(a)至图3.3(d)所示。



(a) 标准高斯分布　(b) 学生-t 分布　(c) 混合高斯分布　(d) 斜偏高斯分布

图 3.3 概率密度函数等高线图

表3.1展示了 KnewImp 算法在 30% 缺失率下对四种典型分布数据（标准高斯、学生-t、混合高斯、斜偏高斯）的补全性能。从整体趋势来看，KnewImp 在斜偏高斯分布中展现出最优异的补全性能，其 mMAE 值稳定在 0.42 左右（MAR：0.422±0.253、MCAR：0.417±0.140、MNAR：0.421±0.111），显著低于其他分布类型。特别是在 MCAR 和 MNAR 机制下，斜偏高斯分布的 mWASS 值分别达到 0.210±0.026 和 0.202±0.006，表明算法对非线性变换后的分布具有出色的适应性。这一结果验证了基于得分函数建模的策略能够有效捕捉非对称分布的特征。

在重尾分布（学生-t 分布）场景中，算法表现出稳定的鲁棒性。相较于标准高斯分布，学生-t 分布在 MAR 机制下的 mMAE 降低了 4.2%（0.769 下降至 0.737），但 mWASS 值略有上升（0.481 上升至 0.513），这反映了算法在保持中心趋势估计精度的同时，对尾部异常值具有更好的包容性。值得注意的是，混合高斯分布在所有缺失机制下都表现出最大的标准差（如 MAR 机制 mMAE 标准差达 0.097），说明多峰分布确实为数据补全带来了更大挑战，但算法仍能维持 0.763-0.824 的 mMAE 区间，展现了良好的泛化能力。

上述现象进一步说明了 KnewImp 算法在不依赖数据归一化假设的前提下，能够通过建模数据得分函数，准确捕捉数据的全局和局部特性，有效应对复杂数据分布的多样性；而上述实验结果从实践上回答了**问题 1**，明确证明了 KnewImp 算法在厚尾、多峰、偏态等特性分布数据中的实用性，展现了其在实际工业过程数据处理中的应用潜力。

表 3.1 KnewImp 算法在 $p_{\text{miss}} = 30\%$ 时在服从不同分布的模拟数据集上的补全精确度结果

场景	分布类型	mMAE	mWASS
MAR	标准高斯	$0.769_{\pm 0.030}$	$0.481_{\pm 0.026}$
	学生-t	$0.737_{\pm 0.053}$	$0.513_{\pm 0.048}$
	混合高斯	$0.763_{\pm 0.097}$	$0.419_{\pm 0.104}$
	斜偏高斯	$0.422_{\pm 0.253}$	$0.492_{\pm 0.025}$
MCAR	标准高斯	$0.769_{\pm 0.013}$	$0.287_{\pm 0.014}$
	学生-t	$0.698_{\pm 0.030}$	$0.307_{\pm 0.014}$
	混合高斯	$0.824_{\pm 0.017}$	$0.391_{\pm 0.023}$
	斜偏高斯	$0.417_{\pm 0.140}$	$0.210_{\pm 0.026}$
MNAR	标准高斯	$0.778_{\pm 0.034}$	$0.309_{\pm 0.030}$
	学生-t	$0.715_{\pm 0.028}$	$0.323_{\pm 0.019}$
	混合高斯	$0.807_{\pm 0.042}$	$0.380_{\pm 0.050}$
	斜偏高斯	$0.421_{\pm 0.111}$	$0.202_{\pm 0.006}$

3.5.2 工业过程数据集补全精度对比实验

本小节旨在研究**问题 2**：“KnewImp 算法相比其他数据补全方法在工业过程数据集的缺失数据补全任务上表现如何?”。为系统评估 KnewImp 算法在工业过程缺失数据补全任务下的性能，本小节将 KnewImp 算法与相关数据缺失值补全方法在第2.3.2.1中的脱丁烷精馏塔数据集上的缺失数据补全任务上进行了对比。为此，本小节选用了下列算法作为基线方法进行对比：

- **扩散模型类方法：**基于条件得分的扩散模型-表格版[146-147]（Conditional Score-based Diffusion Models_Tabular，CSDI_T）和缺失扩散模型[148]（MissDiff）。

- **隐变量类方法：**缺失数据重要性加权自编码器[149]（missing data importance-weighted autoencoder，MIWAE）和缺失数据的输入、细化和因果学习模型[150]（Missing data Imputation Refinement And Causal LEarning，MIRACLE）。

- **分布对齐类方法：**辛克宏补全[134]（Sinkhorn imputation，Sink）、转换分布对齐[151]（transformed distribution matching，TDM）和局部相似度保持补全[152]（local similarity preserved transport for direct imputation，LSPT-D）。

实验的相关超参数在附录B中给出，而相关实验结果在表3.2进行展示。

表 3.2 脱丁烷精馏塔数据集上的补全精度对比结果

场景	模型	$p_{\text{miss}}=0.1$		$p_{\text{miss}}=0.2$		$p_{\text{miss}}=0.3$		$p_{\text{miss}}=0.4$		$p_{\text{miss}}=0.5$	
		mMAE	mWASS	mMAE	mWASS	mMAE	mWASS	mMAE	mWASS	mMAE	mWASS
MAR	CSDI_T	0.765†	2.621†	0.749†	2.402†	0.789†	2.437†	0.720†	2.364†	0.746†	2.304
	MissDiff	0.757†	1.966†	0.737†	2.145†	0.780†	2.046†	0.707†	2.123†	0.738†	2.217
	MIRACLE	0.422†	0.197	0.409†	0.218	0.485	0.278	0.480	0.317	0.513	**0.380**
	MIWAE	0.468†	0.213	0.428†	0.218†	0.499†	0.280†	0.479	**0.303**	**0.489**	0.386
	Sink	0.690†	0.345†	0.677†	0.402†	0.726†	0.493†	0.691†	0.549†	0.710†	0.690
	TDM	0.677†	0.363†	0.679†	0.433†	0.736†	0.536†	0.701†	0.597†	0.730†	0.755
	LSPT-D	0.512†	0.258†	0.516†	0.309†	0.584†	0.398†	0.576†	0.463†	0.609†	0.603
	KnewImp	**0.397**	**0.191**	**0.361**	**0.186**	**0.440**	**0.265**	**0.459**	0.334	0.508	0.462
超越基线方法数		8	8	8	8	8	8	8	6	7	6
MCAR	CSDI_T	0.730†	2.572†	0.738†	2.555†	0.738†	2.658†	0.735†	2.678†	0.739†	2.702
	MissDiff	0.727†	1.852†	0.735†	1.732†	0.735†	1.469†	0.732†	1.503†	0.735†	1.596
	MIRACLE	0.430†	0.185†	0.479†	0.236	0.528†	**0.344**	0.583†	0.468†	0.637†	0.669
	MIWAE	0.462†	0.209†	0.501†	0.262†	0.525†	0.353	**0.559**†	**0.468**	0.603†	**0.640**
	Sink	0.655†	0.318†	0.675†	0.401†	0.682†	0.538†	0.690†	0.699†	0.702†	0.923
	TDM	0.648†	0.345†	0.675†	0.438†	0.689†	0.591†	0.698†	0.764†	0.712†	1.012
	LSPT-D	0.491†	0.241†	0.538†	0.321†	0.568†	0.451†	0.597†	0.610†	0.627†	0.838
	KnewImp	**0.373**	**0.169**	**0.428**	**0.234**	**0.469**	0.353	**0.510**	0.496	**0.554**	0.719
超越基线方法数		8	8	8	8	8	6	8	6	8	6
MNAR	CSDI_T	0.740†	2.531†	0.744†	2.514†	0.732†	2.446†	0.736†	2.623†	0.737†	2.645
	MissDiff	0.732†	1.578†	0.739†	1.965†	0.729†	1.789†	0.725†	1.569†	0.730†	1.639
	MIRACLE	0.459†	0.224†	0.499†	0.268	0.543†	0.360	0.616†	0.491	0.689†	0.733
	MIWAE	0.482†	0.238†	0.504†	0.272†	0.527†	**0.352**	0.563†	**0.431**†	0.589†	**0.572**
	Sink	0.673†	0.351†	0.697†	0.450†	0.686†	0.562†	0.710†	0.730†	0.711†	0.928
	TDM	0.659†	0.373†	0.691†	0.484†	0.691†	0.606†	0.715†	0.790†	0.721†	1.004
	LSPT-D	0.510†	0.272†	0.561†	0.365†	0.573†	0.472†	0.616†	0.637†	0.644†	0.848
	KnewImp	**0.390**	**0.194**	**0.429**	**0.251**	**0.466**	0.360	**0.509**	0.496	**0.560**	0.693
超越基线方法数		8	8	8	8	8	8	8	6	8	7

注：匕首符号 † 表示在成对样本 t-检验中，KnewImp 方法相比其他基线方法在统计学上具有显著性差异（$p<0.05$）。**粗体**标注结果为各指标的最优表现，而下划线标注结果则为各指标的次优表现。

表3.2展示了以下实验现象：

（1） 基于隐变量模型的方法，如 MIWAE 和 MIRACLE，在不同缺失场景和缺失率下普遍表现出良好的数据补全性能，往往能取得次优甚至最优的结果。

（2） 相比于基于隐变量模型的方法以及旨在对齐数据分布的方法，基于扩散模型的补全方法（如 CSDI_T 和 MissDiff）在本研究的评估中并未展现出显著的优越性。尤其是在用于衡量分布相似性差异的指标 mWASS 上，扩散模型方法的数值通常比多数基线方法高出一个数量级。

（3） 本章提出的 KnewImp 算法在多数实验场景下均表现出优异的性能，并在统计学上呈现出显著性差异，验证了其有效性。

实验结果中的现象（1）充分证明了隐变量模型在缺失数据补全任务中的有效性和可靠性，这归因于其强大的数据内在结构建模能力。而现象（2）则揭示了一个重要发现：虽然扩散模型通过巧妙规避隐变量模型中归一化常数计算的难题，在图像生成等生成任务中取得了显著成功，但当应用于缺失数据补全这一本质上具有判别性质的任务时，其固有的局限性变得明显。具体而言，扩散模型本身追求生成多样性的优化目标，加之其对缺失数据采用的遮盖训练策略，共同导致了推理结果的发散性与不一致性问题，显著削弱了补全精度。这一现象从反面佐证了3.4.2节提出的基于负熵诱导的补全框架以及3.4.3节构建的联合分布诱导补全方法的必要性和价值，进一步支持了本研究的理论贡献。现象（3）凸显了本章提出的 KnewImp 算法的显著优势和理论框架的有效性。具体而言，KnewImp 在多数实验场景（包括不同缺失模式以及不同缺失率）下均展现出卓越性能，不仅在精度指标上优于现有方法，这种全方位的性能优势经过严格的统计显著性检验（$p < 0.05$），充分验证了 KnewImp 算法在补全工业过程缺失数据时的稳健性和适应性。总而言之，上述实验结果充分验证了 KnewImp 算法在实际工业数据补全任务中的卓越性能，从实践角度有力地回答了**问题 2**，进一步巩固了 KnewImp 算法的理论价值与应用潜力。

3.5.3 消融实验

本小节关注于**问题 3**：“是什么让 KnewImp 算法在工业过程数据集的缺失数据补全任务上取得了如此好的表现效果?”为此，本小节考虑进行消融实验，以研究 KnewImp 算法取得如此优越性的原因。为此，本小节通过消融实验来研究 KnewImp 算法的两个

关键组成部分，以阐明 KnewImp 算法的有效性。具体而言，本小节对以下两个关键模块进行消融实验：

- **负熵泛函正则项：**在此模块的消融实验中，负熵泛函被从 $\mathcal{F}^{\text{joint-NER}}$ 中移除。
- **联合分布建模策略：**在此模块的消融实验中，与常规的扩散模型类补全方法[130,146]类似，模型通过掩蔽部分数据以构建标签进而学习 $\nabla_{\boldsymbol{X}^{(\text{miss})}} \log p_\theta(\boldsymbol{X}^{(\text{miss})}|\boldsymbol{X}^{(\text{obs})})$。

实验结果在表3.3中进行了展示。

表 3.3 脱丁烷精馏塔数据集上的 KnewImp 算法的消融实验对比结果

场景	消融模块		$p_{\text{miss}}=0.1$		$p_{\text{miss}}=0.3$		$p_{\text{miss}}=0.5$	
	负熵泛函	联合分布	mMAE(↑)	mWASS(↑)	mMAE(↑)	mWASS(↑)	mMAE(↑)	mWASS(↑)
MAR	✗	✗	65.0%	1.3E3%	77.9%	1.2E3%	51.2%	6.1E2%
	✗	✓	19.0%	11.1%	12.7%	8.5%	26.9%	30.5%
	✓	✗	64.6%	1.3E3%	90.1%	1.3E3%	56.1%	6.8E2%
	✓	✓	-	-	-	-	-	-
MCAR	✗	✗	81.2%	1.9E3%	57.6%	9.0E2%	32.7%	3.9E2%
	✗	✓	25.2%	21.0%	11.3%	5.3%	10.7%	7.1%
	✓	✗	81.3%	1.9E3%	57.7%	9.0E2%	32.6%	3.9E2%
	✓	✓	-	-	-	-	-	-
MNAR	✗	✗	75.8%	1.5E3%	55.2%	8.6E2%	32.2%	4.1E2%
	✗	✓	21.7%	21.0%	12.3%	9.3%	12.3%	11.8%
	✓	✗	65.7%	1.4E3%	55.2%	8.6E2%	32.3%	4.1E2%
	✓	✓	-	-	-	-	-	-

表3.3展示了以下实验现象：

（1） 移除“负熵泛函”模块后，模型的推理结果较为病态（ill-posedness），表现为模态发散严重和模型性能显著下降。例如，在 MAR 和 $p_{\text{miss}}=0.1$ 条件下，mWASS 恶化到 1.3E3%，mMAE 也降到了 65.0%，对比完整模型的优化效果尤为明显。这表明“负熵泛函”模块在稳定推理分布及削减数据间模态漂移中扮演了关键角色。

（2） 移除“联合分布”建模策略时，各指标均显著劣化，尤其在分布相似性（mWASS）上的退化程度显示了建模全局联合分布对补全任务不可或缺的重要性。例如，在 MCAR 和 $p_{\text{miss}}=0.3$ 条件下，mMAE 高达 57.7%，mWASS 达到 9.0E2%，远高于完整模型，可见仅依赖条件分布无法有效校正数据间潜在依赖关系。

（3） 当同时移除“负熵泛函”和“联合分布”建模策略时，模型性能降至最低。在所有缺失机制下，mMAE 和 mWASS 均表现出数量级的恶化（如 MNAR 场景下，$p_{\text{miss}} = 0.3$ 时，mWASS 达到 8.6E2%）。相比之下，完整模型所展现的显著性能提升表明，“负熵泛函”和“联合分布”建模策略并非相互独立，而是通过协同作用共同确保补全结果的准确性和分布一致性。

（4） 无论是 MAR、MCAR 还是 MNAR 场景，完整模型在不同缺失机制下的性能改善模式一致性较高。这说明 KnewImp 依赖的关键模块具有普适性，可有效应对工业场景中复杂多变的缺失模式。

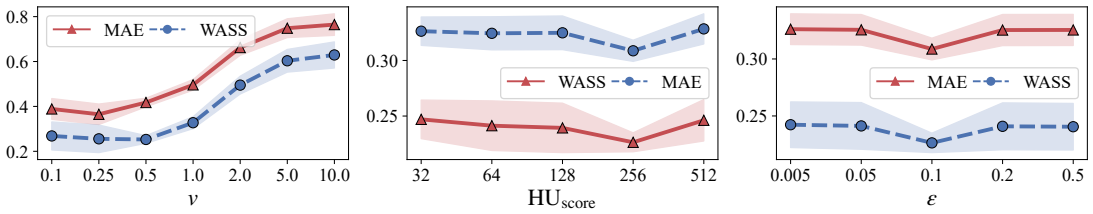
（5） 随缺失率从 $p_{\text{miss}} = 0.1$ 增加到 $p_{\text{miss}} = 0.5$，模型的性能差异进一步放大，尤其在高缺失率条件下（如 $p_{\text{miss}} = 0.5$），依赖 NER 与 Joint 的正确协同显得更加重要。例如，在 MNAR 条件下，缺失率较大时 mMAE 和 mWASS 出现了显著增长，而完整模型的鲁棒性依然能保持良好水平，这表明 KnewImp 在应对高度缺失的工业数据中表现尤为出色。

通过消融实验可以发现，负熵泛函正则项和联合分布建模策略是 KnewImp 算法性能的核心支柱。实验结果表明，缺乏负熵泛函正则项时，模型推理分布易出现病态，导致显著的发散性和分布偏移（如 MAR 场景下 mWASS 激增至 1.3E3%），验证了负熵泛函正则项在稳定推理分布方面的关键作用；而缺乏联合分布建模策略时，模型补全精度和分布一致性显著劣化（如 MNAR 场景下 mMAE 达到 55.2%），表明直接建模条件分布通常会面临推理和训练阶段缺失数据分布的不一致问题。此外，负熵泛函正则项和联合分布建模策略的协同效应尤为显著，二者同时被移除时（第一行），性能降至最低（如 MCAR 场景下 mWASS 达 3.9E2%），而完整模型展现出跨缺失机制和缺失率的稳定高性能，进一步验证了二者在提升推理结果稳定性和对齐训练-推理阶段模型一致性中的互补性。这一发现从实践角度回答了**问题 3**，并佐证了3.4.2节和3.4.3节中提出理论框架的有效性。

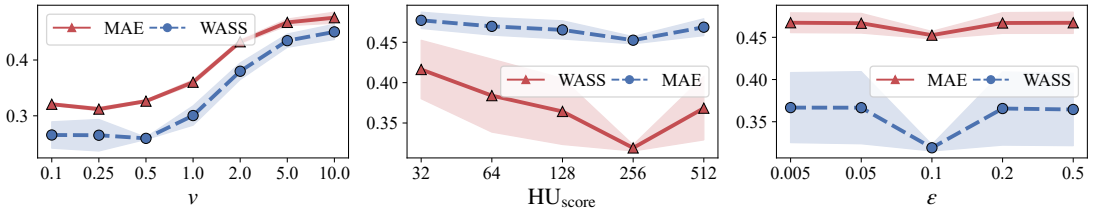
3.5.4 敏感性分析

本小节关注于**问题 4**：“在超参数产生变动的情况下，KnewImp 算法的性能会产生什么样的变化?”为此，本小节考虑进行消融实验，以研究 KnewImp 算法取得如此优越

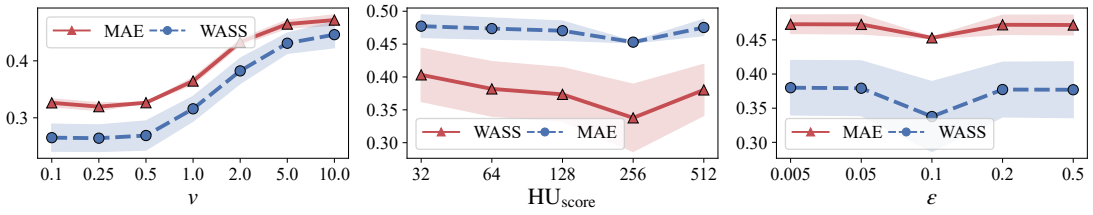
性的原因。为此，本小节考虑调整 KnewImp 算法的带宽 v、隐藏层单元数目 $\mathrm{HU}_{\mathrm{score}}$ 和微扰大小 ε 对 $p_{\mathrm{miss}}=0.3$ 时的补全精确度影响，从而从实践上回答这个问题。图3.4(a)、图3.4(b)和图3.4(c)呈现了随着上述超参数变化，KnewImp 算法在脱丁烷精馏塔数据集上的补全精度对比结果，阴影区域表示 ±1 标准差置信范围。



(a) MAR 场景下的敏感性分析结果



(b) MCAR 场景下的敏感性分析结果



(c) MNAR 场景下的敏感性分析结果

图 3.4 KnewImp 方法在 $p_{\mathrm{miss}}=0.3$ 时的敏感性分析结果

图3.4展示了以下实验现象：

（1） 随着带宽 v 的增加，补全精度显著下降。这是由于过大的带宽导致速度场过度平滑，使变分分布 $\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})})$ 的探索空间过于扩展，削弱了分布的集中性，从而降低了补全性能。

（2） 当隐藏层单元数从 128 增加到 512 时，补全精度呈现先上升后下降的趋势。这表明，网络规模过小可能限制了模型对数据集的充分拟合，而网络规模过大则容易引发过拟合问题，二者均对补全效果产生不利影响。

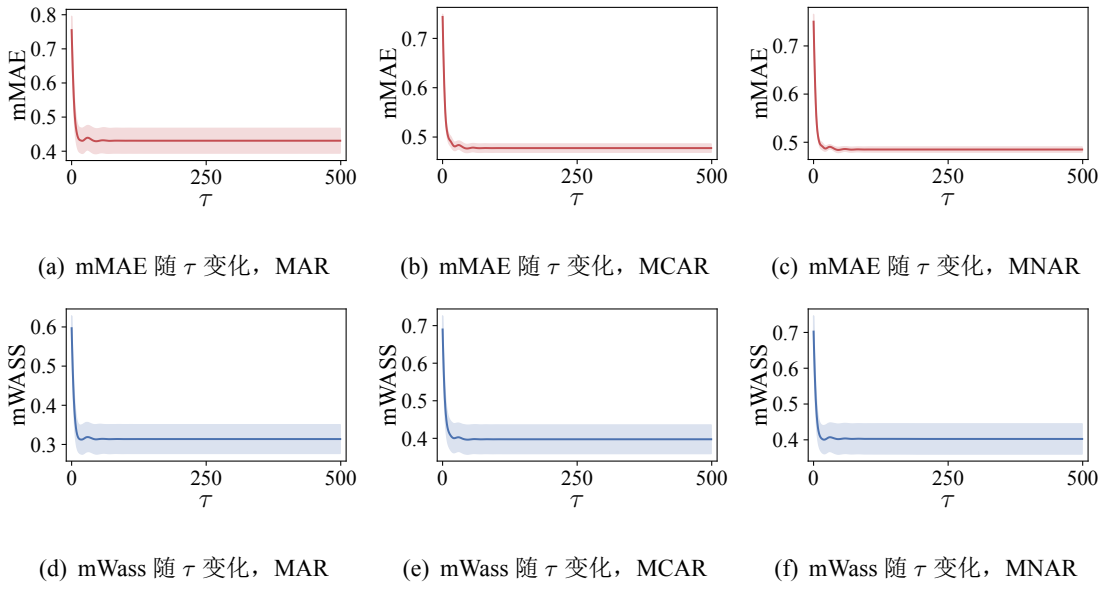
（3） 微扰大小 ε 对补全性能的影响呈现非单调趋势：较小的微扰大小（如 $\varepsilon=0.005$）

尽管提高了泛函优化的精度，但因收敛时间较长可能降低整体性能；而较大的微扰大小（如 $\varepsilon = 0.05$）则因步长过大导致优化误差增加，显著降低补全精度。这表明，步长的选择需要在泛函优化的数值精度和计算效率之间取得平衡。

综上所述，在将 KnewImp 算法应用于缺失数据补全任务时，带宽 v 应保持较小，以确保分布的集中性；隐藏层单元数 $\mathrm{HU}_{\mathrm{score}}$ 应设置为适中规模，以在数据拟合和模型泛化之间取得平衡；微扰大小 ε 则需适中，以在泛函优化的精度和计算效率之间权衡。总而言之，上述敏感性分析的实验结果结果从实践角度回答了**问题 4**，明确了超参数如何影响模型性能，并为 KnewImp 算法在工业场景中的应用提供了重要参考。

3.5.5　收敛性分析

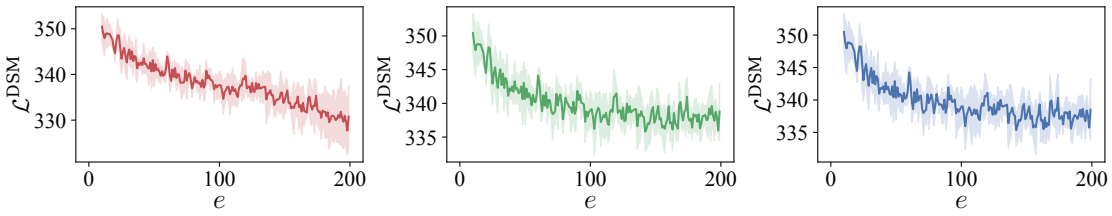
最后，本小节将讨论**问题 5**：“KnewImp 算法能否收敛？”需要指出的是，由于 $\mathcal{Q}(\boldsymbol{X}^{(\mathrm{miss})}$ 在计算过程中无法被直接估计，因此难以显式计算 $\mathcal{F}^{\mathrm{NER}}$。尽管如此，仍可以通过观察 mMAE 和 mWASS 在迭代过程中的变化，来从实践上验证 KnewImp 算法“补全”步的收敛性。为此，本小节在图3.5和图3.6中分别展示了 KnewImp 算法应用在脱丁烷精馏塔数据集时的“补全”mMAE 与 mWASS 随迭代轮数变化的曲线和“训练”步 $\mathcal{L}^{\mathrm{DSM}}$ 随迭代轮数变化的曲线，其中阴影部分表示 ± 1 倍标准差范围。



(a) mMAE 随 τ 变化，MAR　(b) mMAE 随 τ 变化，MCAR　(c) mMAE 随 τ 变化，MNAR

(d) mWass 随 τ 变化，MAR　(e) mWass 随 τ 变化，MCAR　(f) mWass 随 τ 变化，MNAR

图 3.5 KnewImp 方法在 $p_{\mathrm{miss}} = 0.3$ 时的“补全”步收敛性分析结果



(a) $\mathcal{L}^{\mathrm{DSM}}$ 随 e 变化，MAR　　(b) $\mathcal{L}^{\mathrm{DSM}}$ 随 e 变化，MCAR　　(c) $\mathcal{L}^{\mathrm{DSM}}$ 随 e 变化，MNAR

图 3.6 KnewImp 方法在 $p_{\mathrm{miss}} = 0.3$ 时的“训练”步收敛性分析结果

通过观察图3.5，可以发现在“补全”步中，mMAE 和 mWASS 随迭代轮数 τ 的变化趋势表明，KnewImp 算法在不同缺失机制（MAR、MCAR、MNAR）下均表现出快速且稳定的下降。在前 100 轮迭代中，mMAE 和 mWASS 显著降低并趋于平稳，表明模型的补全精度逐步提升，最终达到稳定状态。这一现象表明，KnewImp 算法能够高效收敛到具有较低误差的补全分布，从而确保算法的全局稳定性。此外，通过观察图3.6，可以观察到，$\mathcal{L}^{\mathrm{DSM}}$ 的值随着训练迭代轮数的增加稳步下降，并最终趋于收敛。在不同缺失机制（MAR、MCAR、MNAR）下，$\mathcal{L}^{\mathrm{DSM}}$ 曲线均展现出类似的收敛特性，这一现象验证了训练过程的稳定性和有效性。值得注意的是，无论是在“补全”步还是“训练”步，KnewImp 算法均在较少的迭代轮数内实现了快速收敛。总而言之，上述实验现象从实践上回答了**问题 5**，并进一步证明了定理3.6和定理3.7的合理性。

3.6　本章小结

本章从泛函优化的角度出发针对隐变量驱动的数据补全方法中出现的“补全结果的发散性”以及“训练与推理目标不一致性”问题提出了系统性解决方案。具体而言，本章从泛函优化的层面分析了隐变量模型在补全过程中出现“补全结果的发散性”问题的原因，并在此基础上引入了“负熵泛函正则项”，设计出了保证“补全结果集中性”的泛函 $\mathcal{F}^{(\mathrm{NER})}$，并利用 RKHS 导出了该泛函的优化策略。进一步地，本章证明了条件分布相关的泛函 $\mathcal{F}^{(\mathrm{NER})}$ 所诱导的补全过程可以通过联合分布相关的泛函 $\mathcal{F}^{(\text{joint-NER})}$ 诱导得到，从而避免了直接构建条件概率模型的问题。基于这些理论，本章总结相关策略并提出了 KnewImp 算法，同时从理论上证明了该算法的收敛性。本章在最后部分通过模拟数据集上的补全有效性、实际工业数据集上的补全准确性、超参数的敏感性分析、消融实验及补全过程的收敛性分析等方面实验对 KnewImp 算法的优越性进行了验证。

4 基于无穷时域最优控制的稳态隐变量模型构建方法

摘要： 本章基于无穷时域最优控制理论，重构了稳态概率隐变量模型的训练过程。首先，剖析了传统 EM 算法在概率隐变量模型训练中的局限性，并引入变分分布量子化方法，将隐变量分布推断问题转化为无穷时域最优控制问题。在此基础上，通过求解并借助再生核希尔伯特空间实现该最优控制问题，提出了新的概率隐变量分布推断方法和具有收敛性保证的 EM 算法。最后，在实验部分从隐变量分布推断准确性、软测量建模精度等维度验证了所提出方法的有效性。

关键词： 最优控制 稳态隐变量模型 再生核希尔伯特空间 软测量建模

4.1 引言

概率隐变量模型及其变体因其能够显式地描述观测数据的不确定性，在工业过程软测量模型的构建中受到广泛关注[14,153]。这类模型通过引入服从简单分布的隐变量，将原本复杂的工业过程数据建模问题转化为工业过程数据与隐变量的联合分布建模，显著降低了建模难度。然而，正如2.1节所述，在利用隐变量简化建模的同时，也不可避免地带来了模型表达能力受限的问题。具体而言，隐变量模型的损失函数通常包含 KL 散度项，而为了简化其计算并降低训练复杂度，通常需要将隐变量的先验分布和模型分布限定在某个预定义的归一化函数族中。这种规范化（specification）策略虽然提高了计算效率，但不可避免地限制了模型对复杂数据的表达能力，最终导致模型在下游软测量任务中的性能下降。

综上所述，如何规避隐变量建模中的模型规范化策略，重构新的隐变量模型训练算法，并推导出易于实现且具有收敛性保证的解决方案，具有重要的研究价值。为此，本章引入无穷时域最优控制理论对隐变量模型的训练过程进行系统性重构，提出了全新的隐变量推断策略，并严格推导了其实现方式。在此基础上，进一步提出了新的隐变量模型训练算法，并对其收敛性进行了严格证明。为验证所提方法的有效性，从隐变量推断的有效性、推理精度、下游软测量任务建模精度以及模型收敛性四个维度开展了系统性实验验证，实验结果充分证实了所提出方法的优越性。

4.2 相关工作回顾与科学问题分析

过去几十年来，概率隐变量模型因其能够有效处理测量噪声和数据不确定性，在工业过程软测量建模领域得到了广泛应用[14,154-155]。早期概率隐变量模型主要基于线性投影假设，即观测数据是隐变量通过线性变换得到，这使得可以通过矩阵求逆技术实现隐变量的直接推断[42,110]。然而，线性假设本质上限制了模型的表达能力，特别是在处理具有高度非线性和动态特性的工业软测量数据时表现尤为明显。为突破这一局限，研究者逐步将多层感知器[51]、循环神经网络[65]和变压器网络[156]等先进神经网络架构引入概率隐变量模型。这些进展通过利用特定数据集的归纳偏置，显著增强了模型的表示能力，使其能够更灵活地建模非线性和动态数据结构。尽管这些变形取得了显著进展，正则化相关的挑战仍然是制约概率隐变量模型发展的关键瓶颈。传统方法通常需要将隐变量约束在预定义的归一化分布族中，以降低计算复杂度并方便正则化项的引入[157]。虽然这种约束简化了优化过程，但根据“免费午餐”定理[100]，这种约束往往会损害模型灵活性并限制近似精度，从而影响下游任务的性能。

近年来，随机微分方程（stochastic differential equation）理论[158-159]和最优传输（optimal transportation）理论[160-163]的最新发展推动了扩散模型[164-165]的出现，这些模型采用了创新的正则化方法。与仅关注隐变量不同，扩散模型对从隐变量到观测数据的生成过程进行正则化，从而将正则化域扩展到路径空间[166-167]（path space）。这一范式转变在图像生成[168]和数据补全[146,169]等领域取得了显著成功。然而，尽管生成过程正则化提高了建模的灵活性，其在概率隐变量模型中的应用仍待深入探索。特别是如何将这一策略应用于概率隐变量模型，并开发具有收敛性保证的新型训练算法，尚未得到充分解决。

基于上述分析，本章旨在解决以下两个关键科学问题：

（1） **隐变量模型规范化的放松：**能否为概率隐变量模型严格地设计放松策略，以提高其在下游任务中的性能？

（2） **放松策略在概率隐变量模型训练中的实现：**能否开发具有收敛性保证的新型隐变量模型训练算法，将上述放松策略纳入概率隐变量模型训练过程中？

4.3 基于无穷时域最优控制的稳态隐变量推断框架

4.3.1 问题阐述

本章考虑以下问题:给定 D 维过程观测数据集 $\{x_i|x_i=[u_i,y_i]\in\mathbb{R}^{\mathrm{D}\times 1},u_i\in\mathbb{R}^{\mathrm{D_{PV}}\times 1},y_i\in\mathbb{R}^{\mathrm{D_{QV}}\times 1},i=1,\ldots,\mathrm{N}\}$，需要构建一个以 θ 为参数的概率隐变量模型 $p_\theta(x|z)$，以描述数据 $\{x_i\}_{i=1}^{\mathrm{N}}$ 从隐变量 $z\in\mathbb{R}^{\mathrm{D_{LV}}}$ 的生成过程 $\mathcal{P}(x|z)$，并设计相应的训练流程。在此过程中，假设生成过程的概率密度函数在 $\mathbb{R}^{\mathrm{D_{LV}}}$ 上是光滑的，即导数 $\nabla_x\log p_\theta(x|z)$ 与 $\nabla_z\log\mathcal{P}(z)$ 存在。

基于上述问题与假设，本章的研究目标可总结如下：

（1） 所提出的概率隐变量模型 $p_\theta(x|z)$ 应尽可能减少对观测数据生成过程 $\mathcal{P}(x|z)$ 的约束，例如不假设生成过程的可逆性。

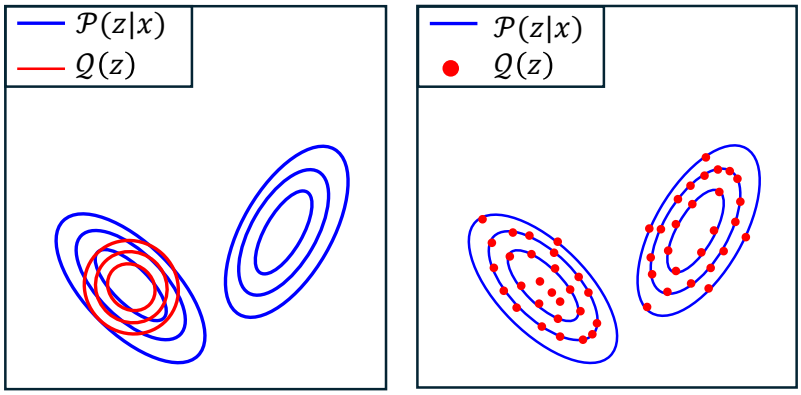
（2） 所提出的训练流程应显式避免对近似分布 $\mathcal{Q}(z)$ 处于归一化函数族的假设。

4.3.2 研究动机分析

考虑式(2-5)，当 $\mathcal{Q}(z)\propto\mathcal{P}(z|x)$ 时，$\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(z)\|\mathcal{P}(z|x)\right]=0$。因此，为了简化 $\mathcal{P}(z|x)$ 的计算流程，先验分布 $\mathcal{P}(z)$ 和条件概率密度函数 $p_\theta(x|z)$ 必须经过精心设计，以确保后验分布 $\mathcal{P}(z|x)$ 与先验分布 $\mathcal{P}(z)$ 属于同一族分布，从而得到闭式解（closed-form solution），这一设计思路被称为“共轭先验”（conjugate prior）。尽管采用“共轭先验”的方法简化了变分分布 $\mathcal{Q}(z)$ 的计算，但这可能限制了 $\mathcal{Q}(z)$ 的灵活性，从而进一步影响模型在下游工业任务上的表现。

类似的问题同样出现在基于“摊销变分推断”的深度概率隐变量模型的训练过程中。考虑式(2-8)，可以发现，为了便于计算 KL 散度项 $\mathbb{D}_{\mathrm{KL}}[q_\varphi(z|x)\|\mathcal{P}(z)]$，同样需要对 $q_\varphi(z|x)$ 进行分布族的限制，使得 $q_\varphi(z)$ 和 $\mathcal{P}(z)$ 同属于一个分布族或者 $q_\varphi(z)$ 和 $\mathcal{P}(z)$ 的支撑存在重叠（overlapped）区域，以简化 KL 散度项的计算。因此，尽管采用深度神经网络提高了 $q_\varphi(z|x)$ 的表示复杂变分分布的能力，但是仍旧没有摆脱对变分分布 $\mathcal{Q}(z)$ 的限制，从而不可避免地会对模型在下游工业任务上的表现产生影响。

具体地说,基于共轭先验的隐变量建模策略或者基于“摊销变分推断”的建模策略将 $\mathcal{Q}(z)\propto\mathcal{P}(x|z)\mathcal{P}(z)$ 限制在一个预定义的标准化分布族 $\mathbb{F}$ 中，即 $\mathcal{Q}(z)\in\mathbb{F}$，如果隐变量的



(a) $\mathcal{Q}(z) \in \mathbb{F}, \mathbb{F} = \{\mathcal{N}(\mu, \Sigma)|\mu \in \mathbb{R}^{\mathrm{D_{LV}}}, \Sigma \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}, \Sigma \succ 0\}$ (b) $\mathcal{Q}(z) = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} \delta(z - z_i)$

图 4.1 $\mathcal{Q}(z)$ 对 $\mathcal{P}(z)$ 逼近的示意图

真实分布 $\mathcal{P}(x|z)\mathcal{P}(z) \notin \mathbb{F}$，则模型的性能会不可避免地产生退化。例如，考虑图4.1(a)所示的案例，当限制 $\mathcal{Q}(z)$ 属于一个单峰高斯分布族，即 $\mathcal{Q}(z) \in \mathbb{F} \coloneqq \{\mathcal{N}(\mu, \Sigma)|\mu \in \mathbb{R}^{\mathrm{D_{LV}}}, \Sigma \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}, \Sigma \succ 0\}$，而数据的后验分布体现为双峰的高斯分布时，$\mathcal{Q}(z)$ 难以对这种双峰高斯分布进行有效逼近，从而产生较大的逼近误差，最终不可避免地导致概率隐变量模型在软测量建模场景中的性能下降。

如上所述，将变分分布 $\mathcal{Q}(z)$ 预先定义为属于一个标准化分布族可能会限制模型的灵活性，从而阻碍下游任务的表现。为了解决这个问题，可以考虑通过一组粒子的粒子 $\{z_i|z_i \in \mathbb{R}^{\mathrm{D_{LV}}}, i = 1, \ldots, \mathrm{M}\}$ 来近似概率密度函数 $\mathcal{Q}(z)$（其中 M 表示粒子数目），实现对变分分布进行“量子化”（quantization），从而提高变分分布的灵活性。此时，变分分布 $\mathcal{Q}(z)$ 可以表示为下式：

$$\mathcal{Q}(z) = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} \delta(z - z_i), \tag{4-1}$$

其中正体希腊字母 δ 表示表示狄拉克德尔塔函数（Dirac delta function），满足如下性质：

$$\delta(z - z_i) = \begin{cases} +\infty & z = z_i \\ 0 & z \neq z_i \end{cases}, \tag{4-2}$$

并且狄拉克德尔塔函数在 $\mathbb{R}^{\mathrm{D_{LV}}}$ 的积分满足如下条件：

$$\int_{\mathbb{R}^{\mathrm{D_{LV}}}} \delta(z - z_i)\,\mathrm{d}z = 1. \tag{4-3}$$

因此，式(4-1)满足如下的非负性约束：

$$\mathcal{Q}(z)=\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\delta(z-z_i)\geq 0, \tag{4-4}$$

和归一化约束：

$$\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}(z)\,\mathrm{d}z=\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\delta(z-z_i)\,\mathrm{d}z=1. \tag{4-5}$$

可以发现不论如何改变 $z_i|_{i=1}^{\mathrm{M}}$ 的位置，$\mathcal{Q}(z)$ 始终是一个良定义的概率测度。基于这种表示方式，可以通过改变空间 $\mathbb{R}^{\mathrm{D_{LV}}}$ 中粒子 $\{z_i|z_i\in\mathbb{R}^{\mathrm{D_{LV}}},i=1,\ldots,\mathrm{M}\}$ 的坐标对变分分布 $\mathcal{Q}(z)$ 进行“塑造”(steering)，使得变分分布 $\mathcal{Q}(z)$ 其能够如图4.1(b)所示对 $\mathbb{R}^{\mathrm{D_{LV}}}$ 上的各种具有复杂形态的分布进行有效逼近。换言之，对变分分布进行“量子化”后，无需将变分分布限定在特定的归一化概率密度函数族内，从而放松了概率隐变量模型中的”规范化”要求，为后续训练中采用该放松策略提供了可能性。

4.3.3　最优控制问题的构造

虽然使用 $z_i|_{i=1}^{\mathrm{M}}$ 来表示 $\mathcal{Q}(z)$ 增强了模型的灵活性并提高了近似精度，但也存在不足之处。具体而言，如图4.2左侧所示，若样本 $z_i|_{i=1}^{\mathrm{M}}$ 的空间分布与目标 $\mathcal{P}(z|x)$ 的概率密度等高线不匹配会导致较大的近似误差；为此，根据第 2.2.1 节的内容，本章考虑以引入微扰 $\phi(z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^{\mathrm{D_{LV}}}$ 来调整 $z_i|_{i=1}^{\mathrm{M}}$ 的位置，图4.2中间到右侧的部分所示。在此基础上，由 $z_i|_{i=1}^{\mathrm{M}}$ 定义的分布 $\mathcal{Q}(z)$ 的形态得以逐渐改变，从而使 $z_i|_{i=1}^{\mathrm{M}}$ 逐步逼近 $\mathcal{P}(z|x)$[1]。

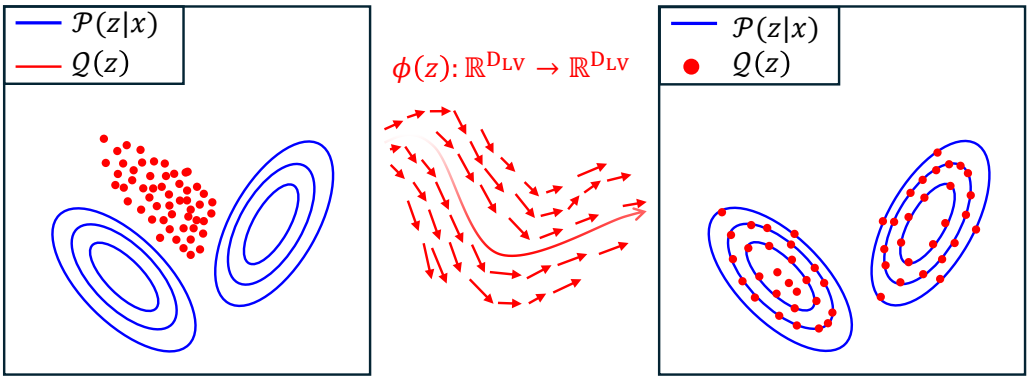


图 4.2 变分分布 $\mathcal{Q}(z)\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\delta(z-z_i)$ 在微扰 $\phi(z)$ 的作用下对 $\mathcal{P}(z|x)$ 进行逼近的示意图

[1]为简化符号表示并提高可读性，在本章后续内容的推导中，发现 $z_{i,\tau}$ 的演化过程不显式地依赖于 τ；因此，在后续内容中将 $z_{i,\tau}$ 简记为 z_i

根据上述分析，如何保证使得 $\int_{\mathbb{R}^{\mathrm{D}}}\mathcal{Q}(z)\mathrm{d}z=1$ 恒成立的沃瑟斯坦空间中设计微扰方向 $\phi(z)$，使得在微扰过程中 $\mathcal{Q}_\tau(z)$ 与 $\mathcal{P}(z|x)$ 之间的 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)]$ 逐渐减小，进而最终实现对 $\mathcal{P}(z|x)$ 的逼近过程是本节需要进行研究的重点问题。由于本节的核心内容是方向 $\phi(z)$ 的设计，因此首先在这里考虑 $\tau\to 0$ 的情况，以简化后续的分析流程。特别地，当 $\varepsilon\to 0$ 时候，$z|_{i=1}^{\mathrm{M}}$ 的演化满足如下的 ODE：

$$z_{\boldsymbol{T}}=z+\varepsilon\phi(z)\Rightarrow\frac{\mathrm{d}z}{\mathrm{d}\tau}=\lim_{\varepsilon\to 0}\frac{z_{\boldsymbol{T}}-z}{\varepsilon}=\phi(z). \tag{4-6}$$

与此同时，对 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)]$ 进行分析，将其重写为下式所示的形式：

$$\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(z)\|\mathcal{P}(z|x)\right]=\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{Q}(z)\right]-\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{P}(z|x)\right], \tag{4-7}$$

注意到 $\log\mathcal{P}(z|x)=\log p_\theta(x|z)\mathcal{P}(z)-\log\mathcal{P}(x)$ 与概率密度函数 $\mathcal{Q}(z)$ 无关，因此对 $-\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{P}(z|x)\right]$ 应用蒙特卡洛近似（Monte Carlo approximation）进行展开：

$$-\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{P}(z|x)\right]\approx-\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\left[\log\mathcal{P}(z|x)|_{z=z_i}\right]. \tag{4-8}$$

根据式(2-9)和式(2-10)，导出如下式所示的梯度下降形式对 $-\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{P}(z|x)\right]$ 进行优化：

$$z_{i+1}=z_i+\varepsilon\nabla_z\log\mathcal{P}(z|x)|_{z=z_i},\ \text{for}\ i=1,2,\ldots,\mathrm{M}. \tag{4-9}$$

对式(4-9)取 $\tau\to 0$ 的极限，可以得到如下的动力系统：

$$\frac{\mathrm{d}z}{\mathrm{d}\tau}=\nabla_z\log\mathcal{P}(z|x)|_{z=z_i},\ \text{for}\ i=1,2,\ldots,\mathrm{M}. \tag{4-10}$$

根据上面内容的分析，可以猜测微扰方向 $\phi(z)$ 应该满如下的形式：

$$\phi(z)=\nabla_z\log\mathcal{P}(z|x)+u(z), \tag{4-11}$$

其中 $\nabla_z\log\mathcal{P}(z|x)$ 对应 $-\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{P}(z|x)\right]$ 的优化，而控制项 $u(z)$ 对应 $\mathbb{E}_{\mathcal{Q}(z)}\left[\log\mathcal{Q}(z)\right]$ 的优化。在此基础上，根据引理2.1，可以构造下式所示的最优控制命题：

$$\arg\min_{u(z)}\quad\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}_{\mathrm{T}}(z)\|\mathcal{P}(z|x)\right]+\frac{1}{2}\int_0^{\mathrm{T}}u^\top(z)u(z)\mathrm{d}\tau, \tag{4-12}$$

$$\text{s.t.}\quad\frac{\mathrm{d}\mathcal{Q}_\tau(z)}{\mathrm{d}\tau}=-\mathcal{Q}_\tau(z)\nabla_z\cdot\left[\nabla_z\log\mathcal{P}(z|x)+u(z)\right], \tag{4-13}$$

其中目标泛函(4-12)中引入了二次项 $\frac{1}{2}\int_0^{\mathrm{T}} u^\top(z)u(z)\mathrm{d}\tau$ 以正则控制策略 $u(z)$，防止最优控制求解过程中出现病态问题，即最优控制解 $u^*(z)$ 不唯一问题。而约束条件(4-13)是式(2-23)根据下式进行变形的结果：

$$\frac{\mathrm{d}\mathcal{Q}_\tau(z)}{\mathrm{d}\tau} = \frac{\partial \mathcal{Q}_\tau(z)}{\partial \tau} + \left[\nabla_z \mathcal{Q}_\tau(z)\right]^\top \left[\phi(z)\right]. \tag{4-14}$$

尽管式(4-12)和(4-13)构建了求解 $\phi(z)$ 的最优控制命题，但是这部分的求解仍旧存在下列问题：

- **KL 散度的显式计算：**可以发现式(4-12)所定义的代价泛函包含了 KL 散度的计算，而 KL 散度的计算良定性的前提是分布的支撑必须是“重叠”（overlap）的。但是，正如图4.2的左侧所示，在 $\tau = 0$ 时刻的变分分布 $\mathcal{Q}_0(z)$ 很可能支撑不能完全重叠，使得 KL 散度的计算不是良定的，最终导致分布 $\mathcal{P}(z|x)$ 的推断失败。

- **概率密度函数 $\mathcal{Q}(z)$ 的显式估计：**按照文献[170-171]，可以通过设计形式为 $-\mathcal{K}(z)\widehat{\mathcal{Q}}_\tau(z)$ 的控制策略（其中 $\mathcal{K}(z)$ 和 $\widehat{\mathcal{Q}}_\tau(z)$ 分别表示控制增益和对 $\mathcal{Q}_\tau(z)$ 的估计）。然而，对 $z_i|_{i=1}^{\mathrm{M}}$ 的密度 $\mathcal{Q}_\tau(z)$ 进行实时状态估计面临着重大挑战，具体而言，$\mathcal{Q}_\tau(z)$ 满足式(4-13)所定义的非线性偏微分方程，而这个非线性 PDE 没有“解析解”（analytical solution）[172]，并且在其求“数值解”（numerical solution）的过程涉及到复杂的时间和空间离散化。如果直接进行偏微分方程求解，对 $\mathcal{Q}(z)$ 进行估计从而，设计控制律，最终会导致推断的时间和空间复杂度的增加，产生较大的计算成本。

因此，接下来的内容将聚焦在怎么解决这两个问题上，并在以解决过程为基础设计新的隐变量推断算法。

针对 KL 散度的显式计算问题，可以考虑将最优控制时域拉长至无穷，即 $\mathrm{T} \to \infty$，使系统达到 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\infty(z)\|\mathcal{P}(z|x)] = 0$ 稳态（$\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\infty(z)\|\mathcal{P}(z|x)] = 0$ 的合理性将在本章的后续内容中进行证明），从而直接规避 KL 散度的计算过程。以此为基础，式(4-12)给出的最优控制的代价泛函可以重构为如下的形式：

$$\underset{u(z)}{\arg\min} \qquad \frac{1}{2}\int_0^\infty u^\top(z)u(z)\mathrm{d}\tau. \tag{4-15}$$

至此，本节已经建立了在规避显式 KL 散度计算的前提下，通过求解无穷时域最优控制问题来推断隐变量分布 $\mathcal{P}(z|x)$ 的方法。此外，通过观察上述转化，可以得到如下注释：

注释 1： 从演化过程的角度来看，式(4-15)和式(4-13)中定义的最优控制问题将一个从给定的归一化概率密度函数族 $\mathbb{F}$ 中推断 $\mathcal{Q}(z)$ 的变分推断问题，重构为一个从无穷时域路径空间（infinite-horizon path space）$\mathscr{C}\left([0,\infty),\mathbb{R}^{\mathrm{D_{LV}}}\right)$ 中推断最优控制策略 $u^{*}(z)$ 的最优控制问题，这种松弛通过假设空间的延展显著地提高了隐变量分布推断过程的灵活性。

注释 2： 从终端条件的角度来看，式(4-15)和式(4-13)中定义的最优控制问题将 $\mathcal{Q}_{\mathrm{T}}(z)$ 与 $\mathcal{P}(z|x)$ 所属的概率密度函数族进行解耦，具体而言，该最优控制求解框架放松了终端时刻对 $\mathcal{Q}_{\mathrm{T}}(z)$ 的分布假设，转而利用一个随着时间 τ 的演化的 ODE 对 $\mathcal{Q}_{\tau}(z)$ 进行微扰，使 $\mathcal{Q}_{\tau}(z)$ 逐步逼近 $\mathcal{P}(z|x)$。这一策略将在 T 时刻的 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{\mathrm{T}}(z)\|\mathcal{P}(z|x)]$ 正则重新分配到整个 $\mathcal{Q}_{\tau}(z)$ 的演化过程，最终减少了对终端时刻变分分布 $\mathcal{Q}_{\mathrm{T}}(z)$ 的限制，提高了隐变量分布推断结果的灵活性。

4.4 隐变量推断及模型训练算法的设计与实现

4.4.1 最优控制问题的求解和平衡态分析

为了在规避显式估计概率密度函数 $\mathcal{Q}(z)$ 的同时，实现式(4-15)和式(4-13)中定义的最优控制问题在计算机语言（如 Python）中的实现，本小节将对上述最优控制问题进行求解，并根据求解结果验证 $\lim_{\mathrm{T}\to\infty}\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}_{\mathrm{T}}(z)\|\mathcal{P}(z)\right]=0$ 的合理性。在求解之前，首先引入如下假设条件，以简化后续最优控制策略的推导过程：

假设 4.1： 微扰方向 $\phi(z)$ 和概率密度函数 $\mathcal{Q}(z)$ 的乘积在 $\|z\|$ 趋近于无穷大时消失，即

$$\lim_{\|z\|\to\infty}\phi(z)\mathcal{Q}(z)=0. \tag{4-16}$$

假设 4.2： 微扰方向 $\phi(z)$ 在 $\|z\|$ 趋近于无穷大时消失，即 $\lim_{\|z\|\to\infty}\phi(z)=0$。

假设 4.3： 概率密度函数 $\mathcal{Q}(z)$ 和 $\mathcal{P}(z|x)$ 有界（bounded）。

基于上述假设，可以得到以下引理：

引理 4.1： 当假设4.1和假设4.2成立时，下列等式成立：

$$\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\phi^{\top}(z)\nabla_{z}\mathcal{Q}(z)\mathrm{d}z=-\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}(z)\nabla_{z}\cdot\phi(z)\mathrm{d}z. \tag{4-17}$$

证明：利用高斯散度定理[173]，$\int_{\mathbb{R}^{D_{LV}}} \nabla_z \cdot \phi(z)\mathcal{Q}(z)\mathrm{d}z$ 可以被重构为下式：

$$\int_{\mathbb{R}^{D_{LV}}} \nabla_z \cdot \phi(z)\mathcal{Q}(z)\mathrm{d}z = \oint_{\partial z} \phi(z)\mathcal{Q}(z) \cdot \vec{n}(z)\mathrm{d}S(z), \tag{4-18}$$

其中 $\vec{n}(z)$ 和 $\mathrm{d}S(z)$ 分别为单位法向量（unit normal vector）和表面元（surface element）。由于假设4.1成立，并且 $\mathcal{Q}(z)$ 的支撑是 $\mathbb{R}^{D_{LV}}$，则将式(4-16)回代至式(4-18)右端，可以导出如下结论：

$$\oint_{\partial z} \phi(z)\mathcal{Q}(z) \cdot \vec{n}(z)\mathrm{d}S(z) = 0. \tag{4-19}$$

将式(4-19)带入式(4-18)右端，并对式(4-18)左端应用分部积分（integration by parts），可以得到下列结果：

$$\int_{\mathbb{R}^{D_{LV}}} \phi^{\top}(z)\nabla_z \mathcal{Q}(z)\mathrm{d}z + \int_{\mathbb{R}^{D_{LV}}} \mathcal{Q}(z)\nabla_z \cdot \phi(z)\mathrm{d}z = 0.$$

从而引理得证。 **证毕**

根据上述引理，可以提出如下的定理求解式(4-15)和式(4-13)给出的无穷时域最优控制命题：

定理 4.1： 式(4-15)和式(4-13)所定义的无穷时域最优控制命题，其最优控制律可以通过下式进行给出：

$$u^*(z) = -\nabla_z \log \mathcal{Q}_\tau(z). \tag{4-20}$$

证明：根据第2.2.2章的介绍，利用庞特里亚金极大值原理[107]，可以首先列出式(4-15)和式(4-13)所定义的无穷时域最优控制命题的哈密顿函数：

$$\boldsymbol{H}(z) = \frac{1}{2}u^{\top}(z)u(z) - \lambda\mathcal{Q}_\tau(z)\nabla_z \cdot \left[\nabla_z \log \mathcal{P}(z|x) + u(z)\right], \tag{4-21}$$

其中伴随态 $\lambda \in \mathbb{R}$。

根据最优控制的一阶条件，可以得出如下的充分条件：

$$\begin{aligned}
&\frac{\partial \boldsymbol{H}(z)}{\partial u(z)} \\
=&u(z) - \frac{\partial \lambda\mathcal{Q}_\tau(z)\nabla_z \cdot u(z)}{\partial u(z)} \\
\overset{\text{(i)}}{=}&u(z) - \frac{\partial \lambda u(z)\nabla_z \mathcal{Q}_\tau(z)}{\partial u(z)} \\
=&u(z) - \lambda\nabla_z \mathcal{Q}_\tau(z) \\
=&0,
\end{aligned} \tag{4-22}$$

其中步骤“(i)”应用了引理4.1。在此基础上，对式(4-22)验证勒让德条件[136]可以得到如下结果：

$$\frac{\partial^2 \boldsymbol{H}(z)}{\partial u^2(z)} = \mathcal{I} \succ 0. \tag{4-23}$$

因此，式(4-22)给出的最优控制问题的解为最优解。因此，根据式(4-22)导出最优控制律 $u^*(z)$ 如下所示：

$$u^*(z) = -\lambda \nabla_z \mathcal{Q}_\tau(z). \tag{4-24}$$

此外注意到，伴随态 λ 应该满足如下的 ODE：

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\tau} = -\frac{\partial \boldsymbol{H}(z)}{\partial \mathcal{Q}_\tau(z)} = 0. \tag{4-25}$$

因此，将式(4-25)带入式(4-21)，可以得到如下的条件：

$$\lambda \nabla_z \cdot [\nabla_z \log \mathcal{P}(z|x) - \lambda \nabla_z \mathcal{Q}_\tau(z)] = 0. \tag{4-26}$$

而根据伴随态的在 T 时刻的边值条件，可以得到如下的结果：

$$\nabla_z \frac{\delta \mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}_\tau(z) \| \mathcal{P}(z|x)\right]}{\delta \mathcal{Q}_\tau(z)} = \nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z) = 0. \tag{4-27}$$

在定理4.1的基础上，通过对比式(4-26)和式(4-27)，可以得到伴随态 λ 需要满足如下条件：

$$\lambda = \frac{1}{\mathcal{Q}_\tau(z)}. \tag{4-28}$$

将式(4-28)代入式(4-24)，可以得到如下的结果：

$$u^*(z) = -\frac{\nabla_z \mathcal{Q}_\tau(z)}{\mathcal{Q}_\tau(z)} = -\nabla_z \log \mathcal{Q}_\tau(z).$$

从而定理得证。 **证毕**

在这个基础上，微扰 $\phi(z)$ 的表达式可以通过下式进行给出：

$$\phi(z) = \nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z). \tag{4-29}$$

而由于微扰 $\phi(z)$ 的存在而导致的 $\mathcal{Q}_\tau(z)$ 的改变可以通过如下的 ODE 进行给出：

$$\frac{\mathrm{d}\mathcal{Q}_\tau(z)}{\mathrm{d}\tau} = -\mathcal{Q}_\tau(z) \nabla_z \cdot \left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z)\right]. \tag{4-30}$$

进一步地，根据式(4-30)，可以给出驱动 $z_i|_{i=1}^{\mathrm{M}}$ 的方程：

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = \nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z) \mid_{z=z_i}, \ i = 1, 2, \ldots, \mathrm{M}. \tag{4-31}$$

此外，根据式(4-31)给出的隐变量分布的推理流程，可以有如下的注释：

注释 3： 通过模拟式(4-31) 中定义的 ODE 来推断隐变量分布 $\mathcal{P}(z|x)$ 仅需得分函数 $\nabla_z \log \mathcal{P}(z|x)$，这显著简化了隐变量推断过程的实现难度。具体而言，评分函数 $\nabla_z \log \mathcal{P}(z|x)$ 可以以如下的形式进行拆解：

$$\nabla_z \log \mathcal{P}(z|x) = \nabla_z \log \mathcal{P}(z) + \underbrace{\nabla_z \log \mathcal{P}(x|z)}_{=\nabla_z \log p_\theta(x|z)} - \underbrace{\nabla_z \log \mathcal{P}(x)}_{=0} = \nabla_z \log \mathcal{P}(z) + \nabla_z \log p_\theta(x|z), \tag{4-32}$$

其中 $\nabla_z \log \mathcal{P}(z)$ 可以通过解析的方法获得，而 $\nabla_z \log p_\theta(x|z)$ 则可以利用基于自动微分框架的深度学习框架（如 PyTorch[139] 和 JAX[140]）进行计算。此外，得分函数的计算过程自然地规避了难以计算的观测数据 x 的概率密度函数 $\mathcal{P}(x)$，因此有效简化了变分推断的计算流程。

注意到，在式(4-12)式(4-15)的推导中，引入了假设条件 $\lim_{\mathrm{T}\to\infty} \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{\mathrm{T}}(z)\|\mathcal{P}(z|x)] = 0$。因此在得出最优控制的控制律后有必要进一步对该假设条件的成立条件进行证明。为此，在本节给出如下的定理通过说明该最优控制的平衡态（equilibrium state）以说明 $\lim_{\mathrm{T}\to\infty} \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_{\mathrm{T}}(z)\|\mathcal{P}(z|x)] = 0$ 的合理性：

定理 4.2： 当 $\tau \to \infty$ 时，微扰 $\phi(z) = \nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z)$ 诱导的而产生演化的概率密度函数 $\mathcal{Q}_\tau(z)$ 渐进收敛于 $\mathcal{P}(z|x)$，换言之，下式成立：

$$\lim_{\tau\to\infty} \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)] = 0. \tag{4-33}$$

证明： 正如在定理4.1的证明过程所示，当 $\tau \to \infty$ 时，概率密度函数 $\mathcal{Q}_\tau(z)$ 将不再发生改变，在这个基础上回顾式(2-23)所给出的 PDE，可以得到如下的分析结果：

$$\begin{aligned}
&\lim_{t\to\infty} \frac{\partial \mathcal{Q}_\tau(z)}{\partial \tau} \\
=& -\left[\nabla_z \mathcal{Q}_\infty(z)\right]^\top \left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\infty(z)\right] \\
&\quad - \mathcal{Q}_\infty(z) \nabla_z \cdot \left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\infty(z)\right] \\
\overset{\text{(i)}}{=}& -\nabla_z \cdot \{\underbrace{\mathcal{Q}_\infty(z)\left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\infty(z)\right]}_{:=\mathrm{Flux}(z)}\} \\
=& 0,
\end{aligned} \tag{4-34}$$

其中步骤“(i)”应用了分部积分法，而 $\mathrm{Flux}(z)$ 表示流函数。在这个基础上，可以发现，流函数 $\mathrm{Flux}(z)$ 随空间的变化率为 0。换言之，流函数 $\mathrm{Flux}(z)$ 是一个常数。

此外，由于假设4.2和4.3，可以发现，流函数 $\mathrm{Flux}(z)$ 满足下式所示的边值条件：

$$\lim_{\|z\|\to\infty} \mathrm{Flux}(z) = 0. \tag{4-35}$$

因此根据式(4-34)和(4-35)可以得出下式所示的结论：

$$\mathrm{Flux}(z) = 0. \tag{4-36}$$

将式(4-36)带入式(4-34)，得到如下的结果：

$$\mathcal{Q}_\infty(z)\left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\infty(z)\right] = 0, \tag{4-37}$$

根据概率密度函数的归一化要求，有 $\mathcal{Q}_\infty(z) \geq 0$，因此有下列等式：

$$\left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\infty(z)\right] = 0. \tag{4-38}$$

在此基础上，对式(4-38)两端同时进行不定积分，可以得到如下的结果：

$$\mathcal{Q}_\infty(z) = \mathcal{C}\mathcal{P}(z|x), \tag{4-39}$$

其中 $\mathcal{C}$ 是一个常数，满足 $\mathcal{C} > 0$。因此，可以进一步得到如下的结论：

$$\mathcal{Q}_\infty(z) \propto \mathcal{P}(z|x), \tag{4-40}$$

则定理得证。 **证毕**

4.4.2 最优控制问题的实现方法

尽管第4.4.1章建立了最优控制的求解策略 $u^*(z) = -\nabla_z \log \mathcal{Q}_\tau(z)$，但是该最优控制求解策略需要显示地估计 $\nabla_z \log \mathcal{Q}_\tau(z)$，而 $\mathcal{Q}_\tau(z)$ 的估计需要求解式(4-13)所定义的非线性 PDE，最终致使在计算机语言中（如 Python）计算 $\nabla_z \log \mathcal{Q}_\tau(z)$ 存在较大困难。因此，本节将导出易处理的（tractable）的最优控制律表达式以实现该最优控制问题的实现进而实现 $\mathcal{Q}(z)$ 对 $\mathcal{P}(z|x)$ 的逼近。

考虑引入拟设（ansatz）函数[2] $\psi(z) : \mathbb{R}^{\mathrm{D_{LV}}} \to \mathbb{R}^{\mathrm{D_{LV}}}$ 对最优控制策略 $\nabla_z \log \mathcal{Q}_\tau(z)$ 进行逼近；则式(4-31)可以重构为如下形式：

$$\frac{\mathrm{d}z_i}{\mathrm{d}\tau} = \left[\nabla_z \log \mathcal{P}(z|x) + \psi(z)\right]|_{z=z_i}, \text{ for } i = 1, 2, \ldots, \mathrm{M}. \tag{4-41}$$

[2] 术语“ansatz”源于量子力学[174]，指的是多体波函数（many-body wavefunction）的近似形式

为了保证这个拟设函数的对原最优控制的 ODE 的逼近精度足够准确，通过内积相似度准则构建下式所示的优化命题：

$$\arg\min_{\psi(z)} \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z)\{[\nabla_z \log \mathcal{P}(z|x)+\psi(z)]^\top [\nabla_z \log \mathcal{P}(z|x)-\nabla_z \log \mathcal{Q}_\tau(z)]\}\mathrm{d}z. \quad (4\text{-}42)$$

在这个基础上，给出下列命题以优化拟设函数 $\psi(z)$：

定理 4.3： 在满足下列条件的前提下：（1）当 $\|z\|$ 趋于无穷大时，拟设函数的值趋于 0，即 $\lim_{\|z\|\to\infty}\psi(z)=0$；（2）概率密度函数 $\mathcal{Q}_\tau(z)$ 有界；拟设函数 $\psi(z)$ 的优化目标可以重构为下式：

$$\arg\max_{\psi(z)} \ \mathbb{E}_{\mathcal{Q}_\tau(z)}\left[\psi^\top(z)\nabla_z \log \mathcal{P}(z|x)+\nabla_z\cdot\psi(z)\right]. \quad (4\text{-}43)$$

证明： 将式(4-42)进行展开，并忽略掉与 $\psi(z)$ 无关的项，可以得到如下的结果：

$$\begin{aligned}&\int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z)\psi^\top(z)\left[\nabla_z \log \mathcal{P}(z|x)-\nabla_z \log \mathcal{Q}_\tau(z)\right]\mathrm{d}z\\ =&\mathbb{E}_{\mathcal{Q}_\tau(z)}\left\{\psi^\top(z)\left[\nabla_z \log \mathcal{P}(z|x)-\nabla_z \log \mathcal{Q}_\tau(z)\right]\right\}\\ \overset{\text{(i)}}{=}&\mathbb{E}_{\mathcal{Q}_\tau(z)}\left[\psi^\top(z)\nabla_z \log \mathcal{P}(z|x)+\nabla_z\cdot\psi(z)\right]\end{aligned} \quad (4\text{-}44)$$

其中步骤“(i)”应用了下式所示的等式：

$$\begin{aligned}&\mathbb{E}_{\mathcal{Q}_\tau(z)}[\psi^\top(z)\nabla_z \log \mathcal{Q}_\tau(z)]\\ =&\int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z)\psi^\top(z)\nabla_z \log \mathcal{Q}_\tau(z)\mathrm{d}z\\ \overset{\text{(ii)}}{=}&-\int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z)\nabla_z\cdot\psi(z)\mathrm{d}z\\ =&\mathbb{E}_{\mathcal{Q}_\tau(z)}\left[\nabla_z\cdot\psi(z)\right],\end{aligned} \quad (4\text{-}45)$$

而步骤“(ii)”是基于定理4.1导出的结果。　　**证毕**

需要指出的是，式(4-43) 中定义的优化命题由于 $\psi(z)$ 的最优解不唯一会呈现出病态特性。为此，本小节提出将拟设函数限制在 RKHS 内对 $\psi(z)$ 进行求解，从而简化最优控制问题的实现。为此，本小节提出以下定理以描述当 $\psi(z)$ 被限制在 RKHS 内时，式(4-43)所定义的优化问题的最优解：

定理 4.4： 当拟设函数 $\psi(z)$ 被限制在 $\mathrm{D_{LV}}$ 维的 RKHS（记为 $\mathscr{H}^{\mathrm{D_{LV}}}$）内，并且 $\mathscr{H}^{\mathrm{D_{LV}}}$ 的核函数 $K(z',z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^{\mathrm{D_{LV}}}$ 满足边界条件 $\lim_{\|z\|\to\infty}K(z',z)=0$ 时，式(4-43)所定义的优化问题的最优解 $\psi^*_{\mathrm{RKHS}}(z)$ 可以通过下式给出：

$$\psi^*_{\mathrm{RKHS}}(z)=\mathbb{E}_{\mathcal{Q}_\tau(z')}\left[K^\top(z',z)\nabla_{z'}\log \mathcal{P}(z'|x)+\nabla_{z'}K(z',z)\right]. \quad (4\text{-}46)$$

证明：当 $\psi \in \mathscr{H}^{\mathrm{D_{LV}}}$ 时，求解拟设函数 $\psi(z)$ 的优化命题可以根据式(4-43)重构为下式：

$$\psi^*_{\mathrm{RKHS}}(z) = \underset{\psi\in\mathscr{H}^{\mathrm{D_{LV}}}}{\arg\max}\ \mathbb{E}_{\mathcal{Q}_\tau(z)}\left[\psi^\top(z)\nabla_z \log \mathcal{P}(z|x) + \nabla_z \cdot \psi(z)\right] - \frac{1}{2}\|\psi(z)\|^2_{\mathscr{H}^{\mathrm{D_{LV}}}}. \quad (4\text{-}47)$$

针对核函数 $K(z',z)$，可以定义其形式如 $\xi(z):\mathbb{R}^{\mathrm{D_{LV}}} \to \mathscr{H}$ 的特征映射从而将核函数分解为 $K(z',z) = \langle \xi(z'), \xi(z)\rangle_{\mathscr{H}^{\mathrm{D_{LV}}}}$。在此基础上，进一步将核函数的谱分解 $K(z',z) = \sum_{i=1}^{\infty} \Lambda_i \Xi_i(z')\Xi_i(z)$（其中 Λ_i 和 $\Xi_i:\mathbb{R}^{\mathrm{D_{LV}}} \to \mathbb{R}$ 分别是特征值和归一化正交基）应用在拟设函数 $\psi(z) \in \mathscr{H}^{\mathrm{D_{LV}}}$ 上，可以得到 $\psi(z) = \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\Lambda_i}\Xi_i(z)$。其中特征重要性权重满足 $\widehat{\psi}_i \in \mathbb{R}^{\mathrm{D_{LV}}}$，特征正交基满足 $\sum_{i=1}^{\infty}\|\widehat{\psi}_i\|_2^2 < \infty$。因此式(4-47)可以重构为下列优化命题：

$$\underset{\psi\in\mathscr{H}^{\mathrm{D_{LV}}}}{\arg\max}\ \mathbb{E}_{\mathcal{Q}_\tau(z)}[\sum_{i=1}^{\infty}\sqrt{\Lambda_i}\nabla_z^\top \log \mathcal{P}(z|x)\widehat{\psi}_i\Xi_i(z) + \nabla_z \cdot \sum_{i=1}^{\infty}\widehat{\psi}_i\sqrt{\Lambda_i}\Xi_i(z)] - \frac{1}{2}\sum_{i=1}^{\infty}\|\widehat{\psi}_i\|^2_{\mathscr{H}^{\mathrm{D_{LV}}}}. \quad (4\text{-}48)$$

在此基础上，最优的特征重要性权重 $\widehat{\psi}_i^*$ 通过式(4-48)对 $\widehat{\psi}_i$ 求偏导，并使之等于 0 进行给出，其具体表达式如下：

$$\widehat{\psi}_i^* = \sqrt{\Lambda_i}\mathbb{E}_{\mathcal{Q}_\tau(z')}\left[\nabla_{z'}\Xi(z') + \nabla_{z'}\log \mathcal{P}(z'|x)\Xi(z')\right]. \quad (4\text{-}49)$$

将式(4-49)代入 $\psi(z) = \sum_{i=1}^{\infty}\widehat{\psi}_i\sqrt{\Lambda_i}\Xi_i(z)$，可以得到：

$$\psi^*_{\mathrm{RKHS}}(z) = \mathbb{E}_{\mathcal{Q}_\tau(z')}\left[K^\top(z',z)\nabla_{z'}\log \mathcal{P}(z'|x) + \nabla_{z'}K(z',z)\right].$$

则式(4-46)得证。　　**证毕**

在定理4.4的基础上，式(4-41)给出的动力系统可以通过下式进行实现：

$$\frac{\mathrm{d}z_i}{\mathrm{d}\tau} = \left\{\nabla_z \log \mathcal{P}(z|x) + \mathbb{E}_{\mathcal{Q}_\tau(z')}\left[K^\top(z',z)\nabla_{z'}\log \mathcal{P}(z'|x) + \nabla_{z'}K(z',z)\right]\right\}|_{z=z_i}, \quad \text{for } i = 1,2,\ldots,\mathrm{M}. \quad (4\text{-}50)$$

对其以步长 ε 进行前向欧拉法（forward Euler's method）离散[175]，可以得到如下递推式：

$$z_{i,\tau+1} = z_{i,\tau} + \varepsilon\left\{\nabla_z \log \mathcal{P}(z|x) + \mathbb{E}_{\mathcal{Q}_\tau(z')}\left[K^\top(z',z)\nabla_{z'}\log \mathcal{P}(z'|x) + \nabla_{z'}K(z',z)\right]\right\}|_{z=z_{i,\tau}}, \quad \text{for } i = 1,2,\ldots,\mathrm{M}. \quad (4\text{-}51)$$

需要注意的是，在定理4.4的推导中引入了边值条件 $\lim_{\|z\|\to\infty}K(z',z) = 0$，为了满足这一边值条件，本章采用了与第3章类似的策略，将 RBF 函数作为核函数：

$$K(z,z') := \exp(-\frac{\|z-z'\|_2^2}{2v}),$$

其中 v 为带宽。根据 Liu 等人的论文[176]，在没有特别说明的情况下，本章将其设置为 $z_i|_{i=1}^{\mathrm{M}}$ 的中位数。通过观察式(4-51)可以得到如下的注释：

注释 4： 式(4-51)所给出的最优控制实现流程中中与概率密度函数 $\mathcal{Q}_\tau(z)$ 有关的运算仅仅出现在期望项 $\mathbb{E}_{\mathcal{Q}_\tau(z)}[\cdot]$ 中。由于该期望项可通过蒙特卡洛模拟近似，即 $\mathbb{E}_{\mathcal{Q}_\tau(z)}[f(z)] \approx \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} f(z_{i,\tau})$，其中 $z_{i,\tau} \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_\tau(z)$，因此该流程避免了显式求解式(2-27)定义的非线性 PDE，从而显著降低了式(4-15)和式(4-13)所定义的最优控制问题的实现难度。

值得注意的是，由于隐变量 z 的后验分布 $\mathcal{P}(z|x)$ 的推断通过模拟一个无穷时域最优控制问题实现，本文将提出的 $\mathcal{P}(z|x)$ 推理算法命名为 InfO 算法（Infinity-horizon optimal control-based variational inference algorithm）。

算法 4.1 InfO 算法伪代码

输入： 目标分布的概率密度函数：$\mathcal{P}(z|x)$、在 $\tau = 0$ 时刻用以构建起始变分分布 $\mathcal{Q}_0(z) = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} \delta(z - z_i)$ 的样本：$z_i|_{i=1}^{\mathrm{M}}$、终止时间：T，离散步长 ε。
输出： T 时刻的样本：$z_{i,\mathrm{T}}|_{i=1}^{\mathrm{M}}$。

```
for τ ← 0 to T − 1 do
    z_{i,τ+1} ← 式(4-51)                                  > 最优控制问题的模拟
end for
return z_{i,T}
```

4.4.3 隐变量推断算法与模型训练算法总结

尽管4.4.2节有效地将隐变量推断问题转化为最优控制问题，并给出了一个较为简捷的实现方式，但它并没有明确地给出新的概率隐变量模型和对应的训练算法。为此，本节首先在图4.3中给出基于 InfO 算法所设计的隐变量模型——InfO-概率隐变量模型（InfO Probabilistic Latent Variable Model，InfO-PLVM）的示意图。对比于图2.1所给出的常规概率隐变量模型结构，本章设计的 InfO-PLVM 在生成数据 x 的过程中引入了 ODE 对隐变量进行演化，这个 ODE 演化过程是“可逆”的，使得服从复杂分布的观测数据 x 可以从 $\tau = 0$ 时刻的服从简单分布的隐变量 z_0 逐渐演化生成，对隐变量的推理也可以通过讲这个 ODE 变换积分上下限进行直接推断。这一优势提高了模型对建模复杂工业过程的灵活性，为建模复杂的工业过程数据提供了有力的基础保障。

基于图4.3中的 InfO-PLVM 结构并结合第2.2.1章的密度演化方程和算法4.1推断得到的 $\mathcal{Q}_\mathrm{T}(z)$，可以根据式(2-9)和式(2-10)导出下式所示的模型 $p_\theta(x|z)$ 的参数 θ 的更新方程，

并将其作为 EM 算法 M 步的迭代策略：

$$\theta_{\tau+1}=\theta_{\tau}+\eta\times\nabla_{\theta}\mathbb{E}_{\mathcal{Q}(z)}[\log p_{\theta}(x|z)]|_{\mathcal{Q}(z)=\mathcal{Q}_{\mathrm{T}}(z)}, \tag{4-52}$$

其中 η 是更新步长，在深度学习的语境下又被称为学习率（learning rate）。与此同时，考虑到 $p_{\theta}(x|z)$ 是支撑在 $\mathbb{R}^{\mathrm{D_{LV}}}$ 的分布，因此式(4-52)可以重构为下式：

$$\theta_{\tau+1}=\theta_{\tau}+\frac{\xi}{\mathrm{M}}\times\sum_{i=1}^{\mathrm{M}}\left[-\nabla_{\theta}\|x-\widehat{x}_i\|_2^2\right], \tag{4-53}$$

其中 $\widehat{x}$ 是由 $p_{\theta}(x|z)$ 在给定 T 时刻的隐变量 $z_{i,\mathrm{T}}|_{i=1}^{\mathrm{M}}$ 时给出的预测值。

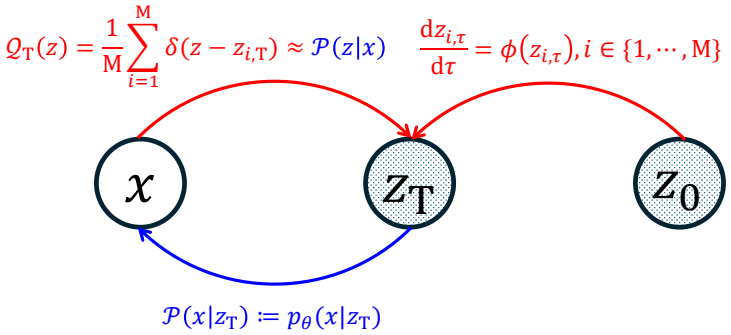


图 4.3 InfO-PLVM 的模型结构示意图

在此基础上，根据式(4-51)和(4-53)可以得到和 EM 算法类似的算法4.2，根据第4.4.2小节，本文将该算法命名为 InfO-EM 算法。

算法 4.2 InfO-EM 算法伪代码

输入： 先验分布的概率密度函数：$\mathcal{P}(z)$、概率隐变量模型 $p_{\theta}(x|z)$ 及其参数 θ、在 $\tau=0$ 时刻用以构建起始变分分布 $\mathcal{Q}_0(z)=\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\delta(z-z_i)$ 的样本：$z_i|_{i=1}^{\mathrm{M}}$、终止时间：T，离散步长：ε、学习率：η 和迭代次数：$\mathcal{E}$。
参数： 模型参数 θ。
输出： 优化后的模型参数 θ^*。

```
for e ← 1 to E do
    z^e_{i,T}|_{i=1}^{M} ← 算法 4.1        > E 步
    θ^e ← Eq. (4-53)                       > M 步
end for
return θ*
```

需要指出的是在 InfO-EM 算法的推导过程中，并没有引入对模型 $p_{\theta}(x|z)$ 的可逆性假设，因此所导出的 InfO-EM 算法可以拓展至不同的深度学习模型结构，从而具有广泛的适用性。接下来将对 InfO-EM 算法的收敛性给出证明。在论证 InfO-EM 算法的收敛性之前，需要引入如下两条引理作为证明所需的数学基础：

引理 4.2： 当样本 $z_i|_{i=1}^{\mathrm{M}}$ 受到微扰 $\phi(z):\mathbb{R}^{\mathrm{D_{LV}}}\to\mathbb{R}^{\mathrm{D_{LV}}}$ 而导致位置变化时，这一变化可用 ODE$\frac{\mathrm{d}z}{\mathrm{d}t}=\phi(z)$ 来描述。基于样本位置所定义的概率密度函数 $\mathcal{Q}_\tau(z)=\sum_{i=1}^{\mathrm{M}}\delta(z-z_i)$，则 $\mathcal{Q}_\tau(z)$ 与 $\mathcal{P}(z|x)$ 之间的 KL 散度随着时间的演化过程可以用如下的 ODE 进行描述：

$$\frac{\mathrm{d}\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)\right]}{\mathrm{d}\tau}=\mathbb{E}_{\mathcal{Q}_\tau(z)}\{[\phi^\top(z)][\nabla_z\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}]\}. \tag{4-54}$$

证明： 基于引理2.1，可以给出如下的 ODE 对 KL 散度随 τ 的演化进行描述

$$\begin{aligned}&\frac{\mathrm{d}\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)]}{\mathrm{d}\tau}\\=&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\frac{\partial\mathcal{Q}_\tau(z)}{\partial\tau}\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}+\mathcal{Q}_\tau(z)[\frac{1}{\mathcal{Q}_\tau(z)}\frac{\partial\mathcal{Q}_\tau(z)}{\partial\tau}]\mathrm{d}z\\=&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}[-\nabla_z\cdot(\mathcal{Q}_\tau(z)\phi(z))][\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}+1]\mathrm{d}z.\end{aligned} \tag{4-55}$$

而根据引理4.1，有如下结论：

$$\begin{aligned}&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\nabla_z\cdot\{[\mathcal{Q}_\tau(z)\phi(z)][\log\frac{\mathcal{Q}(z)}{\mathcal{P}(z|x)}+1]\}\mathrm{d}z\\=&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}[\mathcal{Q}_\tau(z)\phi(z)]^\top[\nabla_z\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}]\mathrm{d}z+\int_{\mathbb{R}^{\mathrm{D_{LV}}}}[\nabla_z\cdot(\mathcal{Q}_\tau(z)\phi(z))][\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}+1]\mathrm{d}z\\=&0.\end{aligned} \tag{4-56}$$

因此，式(4-55)的最后一行可以重构为下式：

$$\begin{aligned}&\frac{\mathrm{d}\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)]}{\mathrm{d}\tau}\\=&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}[\mathcal{Q}_\tau(z)\phi(z)]^\top[\nabla_z\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}]\mathrm{d}z\\=&\mathbb{E}_{\mathcal{Q}_\tau(z)}\{[\phi^\top(z)][\nabla_z\log\frac{\mathcal{Q}_\tau(z)}{\mathcal{P}(z|x)}]\}.\end{aligned} \tag{4-57}$$

从而式(4-55)得证。 **证毕**

引理 4.3： 当拟设函数 $\psi(z)$ 的假设空间没有任何限制时，由式(4-42)定义的求解拟设函数的优化命题的最优解满足下式:

$$\psi^*(z)=-\nabla_z\log\mathcal{Q}_\tau(z). \tag{4-58}$$

证明： 展开式(4-42)可以得到如下结果：

$$\begin{aligned}&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}_\tau(z)\left[\nabla_z\log\mathcal{P}(z|x)\right]^\top\left[\nabla_z\log\mathcal{P}(z|x)\right]\mathrm{d}z+\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}_\tau(z)\left[\nabla_z\log\mathcal{P}(z|x)\right]^\top\left[\psi(z)\right]\mathrm{d}z\\-&\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}_\tau(z)\left[\nabla_z\log\mathcal{Q}_\tau(z)\right]^\top\left[\nabla_z\log\mathcal{P}(z|x)\right]\mathrm{d}z-\int_{\mathbb{R}^{\mathrm{D_{LV}}}}\mathcal{Q}_\tau(z)\left[\nabla_z\log\mathcal{Q}_\tau(z)\right]^\top\left[\psi(z)\right]\mathrm{d}z.\end{aligned} \tag{4-59}$$

其中 $\mathbb{E}_{\mathcal{Q}_\tau(z)}\{[\nabla_z \log \mathcal{P}(z|x)]^\top [\nabla_z \log \mathcal{P}(z|x)]\}$ 和 $\mathbb{E}_{\mathcal{Q}_\tau(z)}\{[\nabla_z \log \mathcal{Q}_\tau(z)]^\top [\nabla_z \log \mathcal{P}(z|x)]\}$ 不依赖于 $\psi(z)$，因此式(4-42)可以简化为如下的形式：

$$\begin{aligned}&\arg\max_{\psi(z)} \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z) \left[\nabla_z \log \mathcal{P}(z|x)\right]^\top [\psi(z)]\, \mathrm{d}z - \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z) \left[\nabla_z \log \mathcal{Q}_\tau(z)\right]^\top [\psi(z)]\, \mathrm{d}z \\ \Rightarrow &\arg\max_{\psi(z)} \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}_\tau(z) \left[\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z)\right]^\top [\psi(z)]\, \mathrm{d}z.\end{aligned} \tag{4-60}$$

在此基础上，对 $\psi(z)$ 求一阶变分，并使之等于 0，可以得到如下结果：

$$\begin{aligned}&\frac{\delta}{\delta\psi(z)} \int_{\mathbb{R}^{\mathrm{D_{LV}}}} \nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z), \psi(z)\mathcal{Q}_\tau(z)\mathrm{d}z \\ =&\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z) \\ =&0,\end{aligned} \tag{4-61}$$

由于二阶变分：

$$\frac{\delta^2}{\delta\psi^2(z)} \int \nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z), \psi(z)\mathcal{Q}_\tau(z)\mathrm{d}z0, \tag{4-62}$$

根据勒让德条件[136]，式(4-61)给出的解为最优解。因此式(4-42)的最优解必须满足下式：

$$\nabla_z \log \mathcal{Q}_\tau(z) = \nabla_z \log \mathcal{P}(z|x). \tag{4-63}$$

通过对比式(4-63)和(4-42)可知，最优的 $\psi^*(z)$ 满足式(4-58)，则引理得证。 **证毕**

在上述引理和定义1 的基础上，本小节进一步提出以下定理以表述 InfO-EM 算法的收敛性：

定理 4.5： 设 $\{\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)]\}_{\tau=1}^{\mathcal{E}}$ 为 InfO-EM 算法生成的迭代序列。若离散步长 ε 和学习率 η 充分小，则存在 $\mathcal{C} \geq 0$，使得对于任意给定的 $\gamma > 0$，存在正整数 N 使得当 $\tau > N$ 时有 $\|\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)\|\mathcal{P}(z|x)] - \mathcal{C}\| < \gamma$ 成立。

证明： 收敛性的证明将从 E 步和 M 步分别展开。

- **E 步收敛性证明：** 根据引理4.3，可以得到如下的不等式：

$$\begin{aligned}&[\psi^*_{\mathrm{RKHS}}(z)]^\top [\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z)] \\ &\qquad \leq [-\nabla_z \log \mathcal{Q}_\tau(z)]^\top [\nabla_z \log \mathcal{P}(z|x) - \nabla_z \log \mathcal{Q}_\tau(z)].\end{aligned} \tag{4-64}$$

则在不等式两端同时加上 $[\nabla_z \log \mathcal{P}(z|x)]^\top [\nabla_z \log \mathcal{Q}_\tau(z) - \nabla_z \log \mathcal{P}(z|x)]$ 可以得到如下结果：

$$\begin{aligned}
&[\nabla_z \log \mathcal{P}(z|x)]^\top [\nabla_z \log \mathcal{Q}_\tau(z) - \nabla_z \log \mathcal{P}(z|x)] \\
&\qquad + [\psi^*_{\mathrm{RKHS}}(z)]^\top [\nabla_z \log \mathcal{Q}_\tau(z) - \nabla_z \log \mathcal{P}(z|x)] \\
\leq &[\nabla_z \log \mathcal{P}(z|x)]^\top [\nabla_z \log \mathcal{Q}_\tau(z) \nabla_z \log \mathcal{P}(z|x)] \\
&\qquad - [\nabla_z \log \mathcal{Q}_\tau(z)]^\top [\nabla_z \log \mathcal{Q}_\tau(z) - \nabla_z \log \mathcal{P}(z|x)] \\
= &- \|\nabla_z \log \mathcal{Q}_\tau(z) - \nabla_z \log \mathcal{P}(z|x)\|_2^2 \\
\leq &0.
\end{aligned} \tag{4-65}$$

通过对比式(4-65)与(4-54)可以发现当微扰方向为 $\phi(z) = \nabla_z \log \mathcal{P}(z|x) + \psi^*_{\mathrm{RKHS}}(z)$ 时，分布 $\mathcal{Q}_\tau(z)$ 与 $\mathcal{P}(z|x)$ 之间的 KL 散度会随着 τ 的演化单调递减。

此外，还注意到 KL 散度是一个非负的泛函，则下列不等式恒成立：

$$\mathbb{D}_{\mathrm{KL}} \left[\mathcal{Q}_\tau(z) \| \mathcal{P}(z|x)\right] \geq 0. \tag{4-66}$$

因此，根据式(4-65)和式(4-66)，可以得出式(4-50)所定义的动力系统收敛的结论。在这个分析的基础上，可以发现式(4-51)是对式(4-50)的前向欧拉法离散[175]，而离散步长 ε 越小，越接近式(4-50)所定义的动力系统，从而 E 步的收敛性得证。

- **M 步收敛性证明：**根据式(4-10)，对式(4-53)中的 η 取极限 $\eta \to 0$，可以得到如下的 ODE：

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = -\frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} \left[\nabla_\theta \|x - \widehat{x}_i\|_2^2\right]. \tag{4-67}$$

根据上式，考虑 $-\frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} \|x - \widehat{x}_i\|_2^2$ 随 τ 的演化轨迹，可以得到如下的结果：

$$\begin{aligned}
&-\frac{1}{\mathrm{M}} \frac{\mathrm{d} \sum_{i=1}^{\mathrm{M}} \|x - \widehat{x}_i\|_2^2}{\mathrm{d}\tau} \\
=&\{-\frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} \left[\nabla_\theta \|x - \widehat{x}_i\|_2^2\right]\}^\top \frac{\mathrm{d}\theta}{\mathrm{d}\tau} \\
=&\frac{1}{\mathrm{M}^2} [\sum_{i=1}^{\mathrm{M}} \nabla_\theta \|x - \widehat{x}_i\|_2^2]^\top [\sum_{i=1}^{\mathrm{M}} \nabla_\theta \|x - \widehat{x}_i\|_2^2] \\
\geq &0.
\end{aligned} \tag{4-68}$$

可以发现 $-\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\|x-\widehat{x}_i\|_2^2$ 随着 τ 的演化是单调递增的。

此外，注意到：

$$-\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\|x-\widehat{x}_i\|_2^2\leq 0 \tag{4-69}$$

在演化过程中恒成立。根据上述现象，可以得出式(4-68)所定义的动力系统收敛的结论。在这个分析的基础上，可以发现式(4-53)是对式(4-68)的前向欧拉法离散[175]，而离散步长 ε 越小，越接近式(4-68)所定义的动力系统，进而 M 步的收敛性得证。

综合上述分析，InfO-EM 算法的收敛性得证。 **证毕**

4.5 实验验证

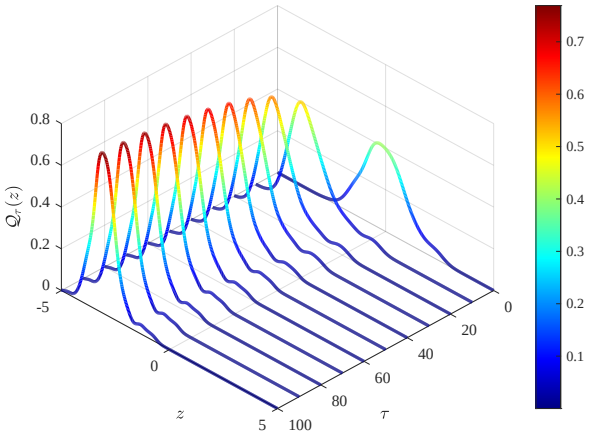
在本节，将对所提出的 InfO 算法和 InfO-EM 算法的有效性进行实验验证。本节的实验将从如下四个研究问题进行展开：

- **问题 1（有效性）：**InfO 算法能否有效地逼近后验分布 $\mathcal{P}(z|x)$？
- **问题 2（精确性）：**InfO 算法能否精确逼近后验分布 $\mathcal{P}(z|x)$？与其他逼近策略相比，其准确性有何优势？
- **问题 3（表现性）：**基于 InfO-EM 算法所训练的 InfO-PLVM 在下游软测量建模任务中的表现如何？
- **问题 4（敏感性）：**在超参数产生变动的情况下，InfO-PLVM 的性能会产生什么样的变化？
- **问题 5（收敛性）：**InfO-EM 算法在训练概率隐变量模型的过程中能否收敛？

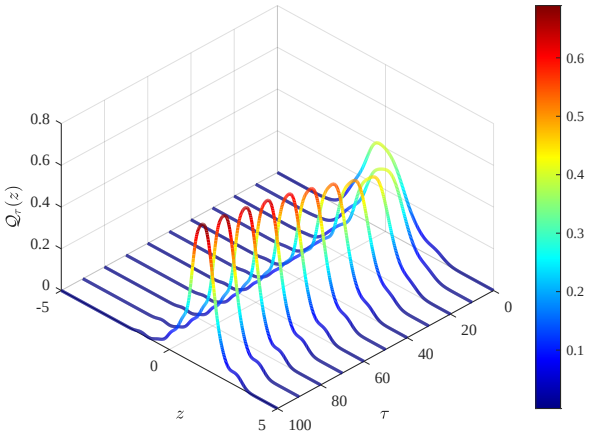
4.5.1 概率密度函数演化轨迹对比

本小节关注**问题 1**：“本章提出的 InfO 算法能否有效逼近后验分布 $\mathcal{P}(z|x)$？”。为回答这一问题，本小节设计了一维概率密度函数逼近的定性实验，通过可视化密度函数 $\mathcal{Q}_\tau(z)$ 随 τ 的演化轨迹，验证 InfO 算法的有效性。为此，实验选取了三种一维分布：高斯分布（Gaussian distribution）$\mathcal{N}(\mu,\sigma^2)$（其中 μ 和 σ 分别为位置参数和尺度参数）、学

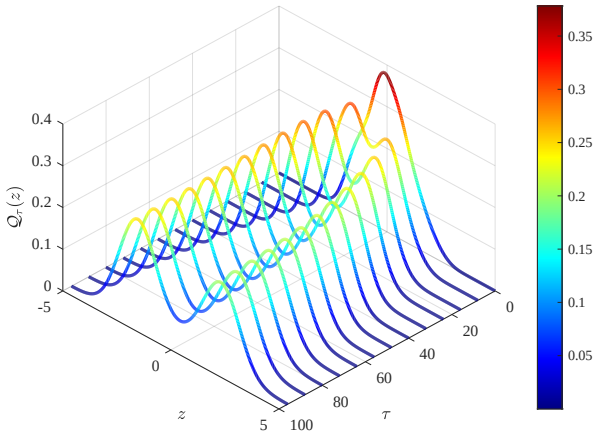
生-t 分布（Student's-t distribution）$St(\nu, \mu, \sigma)$（其中 ν、μ 和 σ 分别为自由度参数、位置参数和尺度参数），以及高斯混合分布。为确保实验公平性，初始分布 $\mathcal{Q}_0(z)$ 设定为标准高斯分布 $\mathcal{N}(0,1)$。图4.4(a)到图4.4(c)展示了 $\mathcal{Q}_\tau(z)$ 随时间 τ 的演化轨迹（其中演化轨迹通过核密度估计获得，核密度估计采用 RBF 核函数，带宽由 Scott 方法确定）。



(a) $\mathcal{P}(z|x) \propto \mathcal{N}(-3, 0.5^2)$



(b) $\mathcal{P}(z|x) \propto St(9, 1.5, 0.5)$



(c) $\mathcal{P}(z|x) \propto \frac{1}{2}\mathcal{N}(-2, 0.5^2) + \frac{1}{2}\mathcal{N}(2, 0.5^2)$

图 4.4 概率密度函数 $\mathcal{Q}_\tau(z)$ 随 τ 的演化轨迹

图4.4展示了以下实验现象：

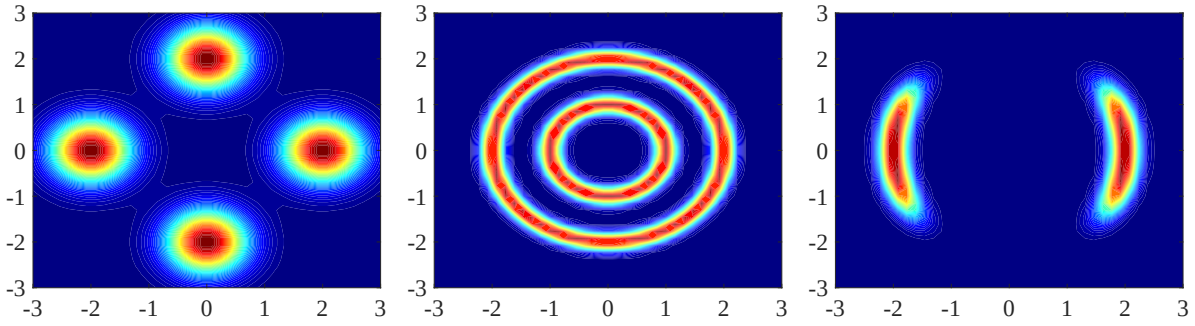
（1） 在图4.4(a)中，从标准高斯分布 $\mathcal{Q}_0(z)$ 出发，随时间推移，扰动过程改变了 $\mathcal{Q}_t(z)$ 的均值和方差，展现了 InfO 算法在调整分布位置和尺度方面的灵活性。

（2） 在图4.4(b)中，InfO 算法成功将 $\mathcal{Q}_\tau(z)$ 从标准高斯分布这一轻尾分布转变为学生-t 分布这一重尾分布，体现了其适应分布尾部行为的能力。

（3） 在图4.4(c)中，尽管初始分布 $\mathcal{N}(0,1)$ 与目标分布 $\frac{1}{2}\mathcal{N}(-2,0.5^2)+\frac{1}{2}\mathcal{N}(2,0.5^2)$ 的外形相似度较小，但 InfO 算法仍能推动 $\mathcal{Q}_\tau(z)$ 从集中于原点的单峰分布逐渐演变为具有两个中心的双峰分布，展示了其改变分布基本形态的能力。

通过上述实验，InfO 算法在分布逼近中的灵活性和有效性得到了充分验证，并为**问题 1** 提供了明确答案。

4.5.2 后验分布推断的精确度的对比



(a) 混合高斯　　(b) 混合环形　　(c) 双月

图 4.5 概率密度函数 $\mathcal{P}(z|x)$ 的概率密度等高线图

本小节将进一步探究所提出的 InfO 算法对隐变量的后验分布 $\mathcal{P}(z|x)$ 逼近的精确性，并探讨**问题 2**：“本章提出的 InfO 算法能否精确逼近后验分布 $\mathcal{P}(z|x)$？与其他逼近策略相比，其准确性有何优势？”为此，本小节考虑三种特定的后验分布类型：混合高斯（Mixture of Gaussian）、混合环形（Mixture of Ring）和双月（Two Moon）分布，对应的概率密度函数的概率密度等高线图在图4.5(a) 至图4.5(c) 中给出。

为了定量评估 InfO 算法的性能，本小节考虑将 InfO 算法得到的 $\mathcal{Q}(z)$ 与两种常见的 $\mathcal{Q}(z)$ 推断策略进行比较：其一，将 $\mathcal{Q}(z)$ 限制在单峰高斯分布分布族内进行推断; 其二，将 $\mathcal{Q}(z)$ 限制在高斯混合分布族内进行推断（本研究中混合成份数设置为 8）。在此基础上，本节拟采用核斯坦因差异[176]（kernelized Stein discrepancy，KSD）$\mathbb{S}$ 作为度量指标，其定义见式(4-70)和(4-71)，KSD 值越小表示 $\mathcal{Q}(z)$ 对 $\mathcal{P}(z|x)$ 逼近的准确性越高[176]。

$$\mathbb{S}\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right) \coloneqq \mathbb{E}_{z,z'\sim\mathcal{Q}(z)}\left[\mathcal{V}_{\mathcal{P}(z|x)}(z,z')\right] \approx \frac{1}{\mathrm{M}^2}\sum_{i=1}^{\mathrm{M}}\sum_{j=1}^{\mathrm{M}}\mathcal{V}_{\mathcal{P}(z|x)}\left(z_i,z_j\right), \tag{4-70}$$

$$\mathcal{V}_{\mathcal{P}(z|x)}(z,z') \coloneqq \left[\nabla_z \log \mathcal{P}(z|x)\right]^\top K(z,z') \left[\nabla_{z'} \log \mathcal{P}(z'|x)\right] + \left[\nabla_z \log \mathcal{P}(z|x)\right]^\top \nabla_{z'} K(z,z') + \left[\nabla_z K(z,z')\right]^\top \left[\nabla_{z'} \log \mathcal{P}(z'|x)\right] + \mathrm{Trace}\left(\nabla_{z,z'} K(z,z')\right). \tag{4-71}$$

此外，为了确保 $\mathbb{S}(\mathcal{Q}(z),\mathcal{P}(z|x))$ 计算结果的非负性，如式(4-70)的最后一行所示，本文采用冯·米塞斯统计量[176]（von Mises statistic）对 $\mathbb{S}(\mathcal{Q}(z),\mathcal{P}(z|x))$ 进行近似计算。

除了基于 KSD 对推断精确性进行定量评估外，本节还将 KSD 用作检验统计量开展拟合优度检验（goodness-of-fit），其具体计算流程详见附录A.1。在拟合优度检验中，设定的原假设如下：

- 原假设 H_0：样本 $z_i|_{i=1}^{\mathrm{M}}$ 来自于分布 $\mathcal{P}(z|x)$。
- 备择假设 H_1：样本 $z_i|_{i=1}^{\mathrm{M}}$ 不来自于分布 $\mathcal{P}(z|x)$。

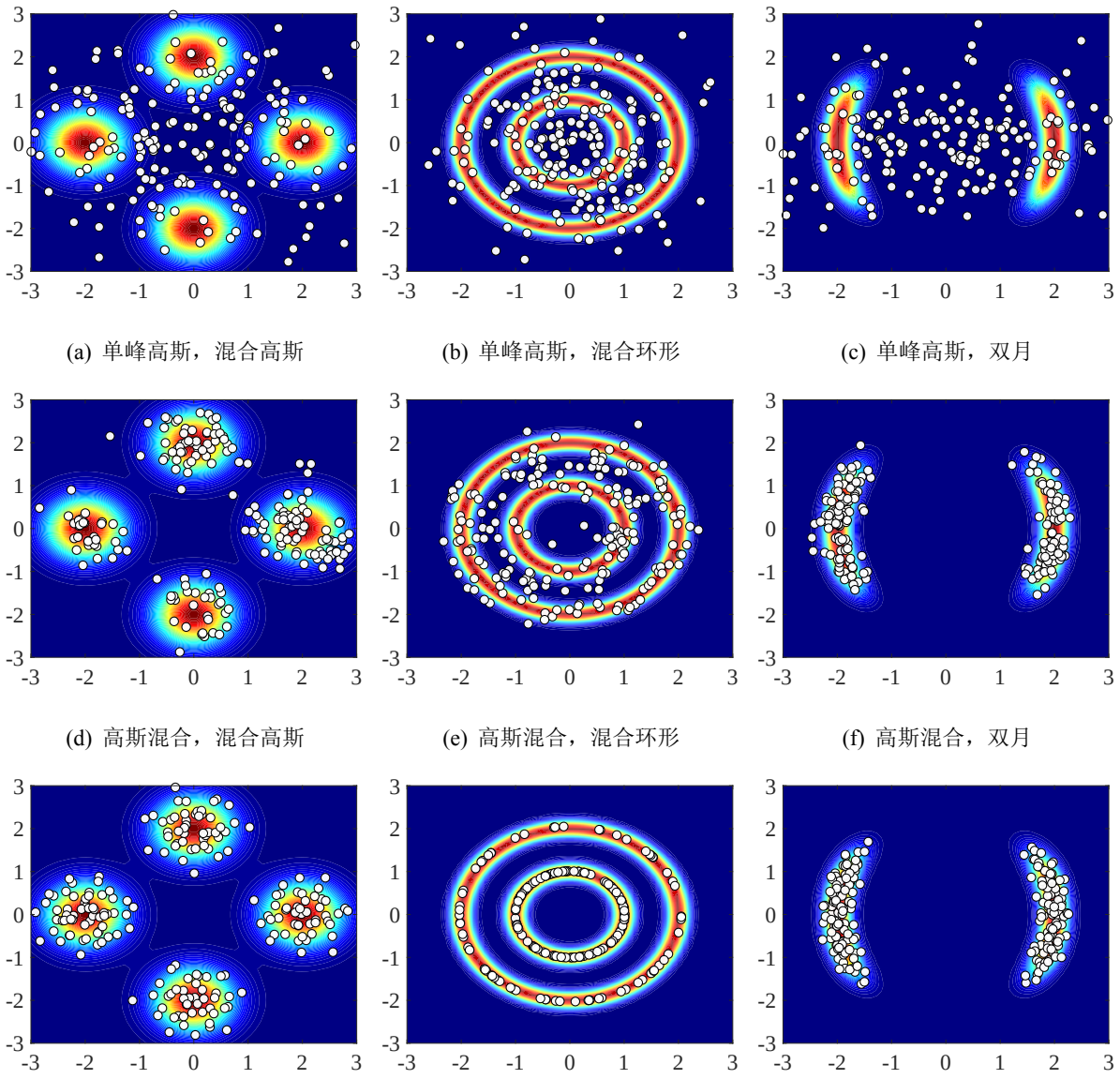
表 4.1 不同分布的逼近精度对比结果

方法	混合高斯		混合环形		双月	
	$\mathbb{S}(\mathcal{Q}(z),\mathcal{P}(z|x))$	H_0/H_1	$\mathbb{S}(\mathcal{Q}(z),\mathcal{P}(z|x))$	H_0/H_1	$\mathbb{S}(\mathcal{Q}(z),\mathcal{P}(z|x))$	H_0/H_1
单峰高斯	3.92E-1†	H_1	1.50E1†	H_1	5.76E1†	H_1
高斯混合	7.11E-2†	H_1	2.29E0†	H_1	4.19E-1†	H_1
InfO 算法	**5.53E-3**	H_0	**1.99E-3**	H_0	**9.66E-3**	H_0

注：匕首符号 † 表示在成对样本 t-检验中，InfO 算法相比其他基线方法在统计学上具有显著性差异（$p < 0.05$）。**粗体**标注的结果为各指标的最优表现。

本小节的实验结果在图4.6(a) 到 (i) 中进行了可视化（图其中白点为从变分分布 $\mathcal{Q}(z)$ 采出的样本），并在表4.1中总结。从图4.6(a) 到 (c) 中可以观察到，当 $\mathcal{Q}(z)$ 指定为单峰高斯时，大多数样本未能落入高概率密度区域，揭示了将 $\mathcal{Q}(z)$ 限制在单峰高斯函数族时，$\mathcal{Q}(z)$ 在逼近灵活性上的局限性。随后，当 $\mathcal{Q}(z)$ 指定为高斯混合时，尽管一些样本确实落入了高概率密度区域，但仍有相当数量的样本位于低密度区域。即使在增加高斯混合的成份数较大的情况下，这一现象仍然存在，如图4.6(d) 到 (f) 所示，进一步突显了对变分分布 $\mathcal{Q}(z)$ 的假设空间施加限制会很大程度地影响其对后验分布 $\mathcal{P}(z|x)$ 的逼近进度。最后，当应用所提出的 InfO 算法逼近 $\mathcal{P}(z|x)$ 时，如图4.6(g) 到 (i) 所示，几乎所有样本都集中在高密度区域，说明所提出的 InfO 算法的优越性。此外，这些可视化结果得到了表4.1中量化结果的进一步支持。具体而言，从表4.1中可以观察到，InfO 算法的 KSD 指标比单峰高斯和高斯混合这两个方法低一到两个数量级，并且在拟合优度检

验中取得了优越的结果。这些观察结果证明了所提出的对变分分布 $\mathcal{Q}(z)$ 的假设空间松弛策略的有效性，为**问题 2** 提供了明确答案，并进一步验证了 InfO 算法的优势。



(a) 单峰高斯，混合高斯　(b) 单峰高斯，混合环形　(c) 单峰高斯，双月

(d) 高斯混合，混合高斯　(e) 高斯混合，混合环形　(f) 高斯混合，双月

(g) InfO 算法，混合高斯，$v = 0.5$　(h) InfO 算法，混合环形，$v = 1.0$　(i) InfO 算法，双月，$v = 0.5$

图 4.6 InfO 算法的逼近结果

4.5.3　软测量建模精度对比实验

本小节将进一步研究**问题 3**：“基于 InfO-EM 算法所训练的 InfO-PLVM 在下游软测量建模任务中的表现如何？”为此，本小节考虑在脱丁烷精馏塔数据集上进行软测量建模，并对比所提出的 InfO-PLVM 和其他的概率隐变量模型之间的软测量预测精确度，

从而说明 InfO-PLVM 的优越性。

为此，本小节选择以下概率隐变量模型作为基线模型：深度贝叶斯概率慢特征分析[97]（deep Bayesian probabilistic slow feature analysis，DBPSFA）、概率性判别时间序列模型[96]（probabilistic discriminative time-series model，PDTM）、改进的无监督 VAE-监督深度 VAE[88]（modified unsupervised deep variational autoencoder-supervised deep variational autoencoder，MUDVAE-SDVAE）、非线性概率隐变量回归[13]（nonlinear probabilistic latent variable regression，NPLVR）和高斯混合变分自编码器[90]（Gaussian mixture variational autoencoder，GMVAE）。值得注意的是，DBPSFA、PDTM、MUDVAE-SDVAE 和 NPLVR 的变分分布设置为标准高斯分布，而 GMVAE 的变分分布则设置为具有八个模态的混合高斯分布。选择这些基线模型的原因和相关的实验细节在附录C中给出。为了进行更为全面的评估，根据2.3.1节，本小节采用 R^2、RMSE、MAPE 和 MAE 作为评估稳态软测量模型优越性的性能指标，在此基础上，本小节的实验结果在表4.2进行呈现。

表 4.2 脱丁烷精馏塔数据集上的软测量精度对比结果

模型	R^2	RMSE	MAPE	MAE
DBPSFA	2.639E-1†	1.738E-1†	2.804E+2†	1.421E-1†
PDTM	6.942E-1†	1.012E-1†	<u>9.951E+1</u>	8.487E-2†
MUDVAE-SDVAE	2.651E-2†	1.990E-1†	2.660E+2†	1.591E-1†
NPLVR	-8.196E-2†	2.091E-1†	2.390E+2†	1.659E-1†
GMM-VAE	<u>8.279E-1</u>†	<u>8.368E-2</u>†	1.186E+2†	<u>6.793E-2</u>†
InfO-PLVM	**9.649E-1**	**3.262E-2**	**5.090E+1**	**2.554E-2**
超越基线模型数	5	5	5	5

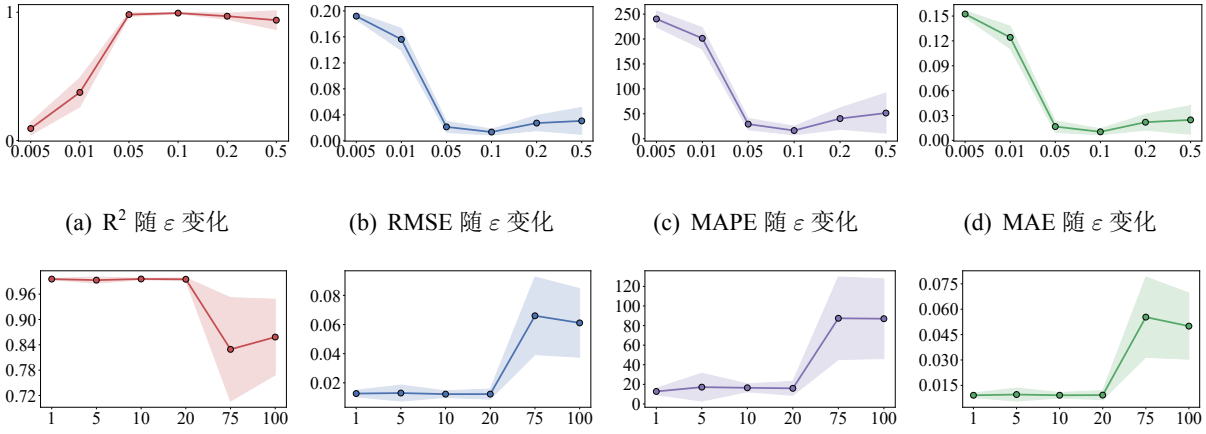
注：匕首符号 † 表示在成对样本 t-检验中，InfO-PLVM 相比其他基线模型在统计学上具有显著性差异（$p < 0.05$）。**粗体**标注的结果为各指标的最优表现，而<u>下划线</u>标注的结果则为各指标的次优表现。

从表4.2中可以观察到，变分分布的选择显著影响模型在下游任务上的表现，这与4.3.2节中讨论所提出的观点相吻合。具体而言，从表4.2中的实验结果可以发现，采用单峰高斯分布作为变分分布的模型，如 DBPSFA、PDTM、MUDVAE-SDVAE 和 NPLVR，模型通常会表现出次优的性能。相反，当放松变分分布 $\mathcal{Q}(z)$ 为混合高斯分布时候，模型的性能得到了较大的提升，正如表4.2所展示的 GMVAE 所呈现的结果那样。在这个基础上，如果进一步放松 $\mathcal{Q}(z)$ 至沃瑟斯坦空间 $\mathscr{P}_2(\mathrm{D_{LV}})$ 时，模型的性能可以得到进一步的提升，正如表4.2中 InfO-PLVM 所展示的结果那样。这一现象说明了通过放松变分分布的所属的分布族对改进隐变量在下游工业软测量建模任务中的重要性，进一步从侧

面展示了所提出的 InfO-EM 算法的优越性，进一步从实践上回答了**问题 3**。

4.5.4　敏感性分析

本小节将进一步研究**问题 4**：“在超参数产生变动的情况下，InfO-PLVM 的性能会产生什么样的变化?”为此，本小节考虑调整 InfO-PLVM 在训练过程中的离散步长 ε 和粒子数目 M 从而从实践上回答这个问题。为此，图4.7(a)至图4.7(h)展示了对 InfO-PLVM 进行灵敏度分析的实验结果，其中的阴影部分代表了 ±0.5 倍标准差。



(a) R^2 随 ε 变化　(b) RMSE 随 ε 变化　(c) MAPE 随 ε 变化　(d) MAE 随 ε 变化

(e) R^2 随 M 变化　(f) RMSE 随 M 变化　(g) MAPE 随 M 变化　(h) MAE 随 M 变化

图 4.7 InfO-PLVM 的敏感性分析实验结果

从图4.7(a)至图4.7(d)可以看出，随着离散步长 ε 的增大，InfO-PLVM 的预测精度，如 R^2，呈现出先增加后下降的趋势。而 RMSE、MAPE、MAE 等指标则表现出先下降后上升的趋势。这些现象表明，随着离散步长 ε 的增加，InfO-PLVM 的建模精度呈现出先提升后降低的趋势。这与 ODE 的离散化特性相一致——在较小的离散步长下，InfO 算法无法在有限时间 T 内有效逼近真实的后验分布 $\mathcal{P}(z|x)$，从而难以达到最优解，导致 InfO-PLVM 未能充分训练，进而降低了模型的建模精度。相应地，随着离散步长的增加，前向欧拉法的近似误差也随之增大，使得 InfO-PLVM 逼近一个误差较大的终点，最终影响了模型的整体性能。

从图4.7(e)至图4.7(h)可以看出，随着粒子数目 M 的增加，InfO-PLVM 的性能，例如 R^2，呈现出明显下降的趋势。同时，RMSE、MAPE、MAE 等指标则显示出上升的趋势。这些现象表明，粒子数的增加与引入 RKHS 模型所产生的梯度之间存在复杂的相互作用。理论上，粒子的增多应有助于提高模型的表达能力[177]，因为更高的粒子数量可

以提供更丰富的特征表示。然而，实际表现却可能与预期相悖，真正的原因可能在于对 $\psi(x)$ 限制在 RKHS 这一假设的带来了额外复杂性和潜在噪音。随着粒子数的增加，模型可能面临过拟合的风险，尤其是在训练数据有限或噪声较大的情况下，这将直接影响到模型的预测准确性和稳定性。因此，在选取粒子数量时，需要在模型的复杂性和预测精度之间找到一个合理的平衡点，方能确保其在实际应用中的有效性。综上所述，在将 InfO-PLVM 应用在软测量建模的过程中，需要综合考虑离散步长和粒子数目的设置，以在模型精度的性能和计算效率取得平衡，进一步从实践上回答了**问题 4**。

4.5.5 收敛性分析

最后，本小节将讨论**问题 5**:“InfO-EM 算法在训练概率隐变量模型的过程中能否收敛?”为此，本小节考虑将 InfO-EM 算法在脱丁烷精馏塔数据集上的训练过程进行展示从而从实践上回答这一问题。需要指出的一点是由于 $\mathcal{Q}_\tau(z)$ 在计算过程中不能显示地进行估计，因此 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(z)||\mathcal{P}(z|x)]$ 难以显式进行计算。尽管如此，仍旧可以通过观察对数似然函数 $\mathbb{E}_{\mathcal{Q}(z)}\left[\log p_\theta(x|z)\right]$ 在迭代过程中的演化从实践上验证 InfO-EM 算法的收敛性。为此，图4.8展示了 InfO-PLVM 在脱丁烷精馏塔数据集上对数似然函数 $\mathbb{E}_{\mathcal{Q}(z)}\left[\log p_\theta(x|z)\right]$ 随着迭代轮数的变化过程，其中的阴影部分代表了 ±0.5 倍标准差。

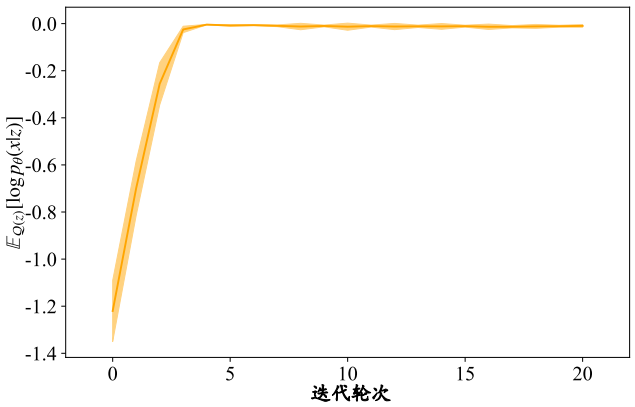


图 4.8 对数似然函数 $\mathbb{E}_{\mathcal{Q}(z)}\left[\log p_\theta(x|z)\right]$ 的演化过程示意图

如图4.8所示，概率隐变量模型的对数似然函数 $\mathbb{E}_{\mathcal{Q}(z)}\left[\log p_\theta(x|z)\right]$ 随着训练轮数的增加而稳步上升。值得注意的是，对数似然函数的平均曲线（橙色点划线）在前 10 个迭代轮数内接近零，且标准差在迭代后期逐渐变小（阴影部分）。上述现象表明 InfO-EM 算法在对 InfO-PLVM 的训练过程中呈现较快且较为稳定的收敛特性，进一步从实践上回答了**问题 5**。

4.6 本章小结

本章提出了一种基于无穷时域最优控制的方法，对稳态概率隐变量模型的训练过程进行系统性重构，以提升其在软测量任务中的性能。具体而言，本章通过“量子化”技术将近似分布表示为有限数量的粒子，进而将隐变量分布的推断问题重新表述为无穷时域最优控制问题。这一转化突破了传统方法在预定义归一化分布族 $\mathbb{F}$ 中进行推断的局限，将归一化函数族 $\mathbb{F}$ 假设空间中的分布优化问题提升为无穷时域路径空间 $\mathscr{C}([0,\infty),\mathbb{R}^{\mathrm{D_{LV}}})$ 中的最优控制策略求解问题，从而有效规避了对隐变量分布的规范化要求，提升了模型的灵活性。基于这一理论框架，本章利用庞特里亚金极大值原理推导了最优控制策略，并借助 RKHS，提出了一个易于实现该最优控制策略的计算方法。在此基础上，本章进一步提出了基于无穷时域最优控制的隐变量推断算法（InfO 算法）和对应的概率隐变量模型训练算法（InfO-EM 算法），并严格证明了所提出的 InfO-EM 算法的收敛性。本章在最后部分通过后验分布推断的有效性实验、后验分布推断的精确性实验、软测量建模的精确性实验、软测量建模的敏感性分析实验和收敛性实验这五个方面的实验证实了所提出隐变量推断算法和隐变量模型训练算法的有效性和优越性。

5　基于镜像下降策略的限制域隐变量模型构建方法

摘要： 针对支撑域为约束域的稳态概率隐变量模型（如图神经网络和变压器网络）训练问题，提出了一种基于镜像下降方法的概率隐变量模型训练框架。首先，本章深入分析了第4章所提出的概率隐变量模型训练框架在隐变量处于约束域场景下失效的根本原因，即其迭代过程对欧几里得空间的隐式依赖性。为克服这一局限，引入布雷格曼散度对迭代过程进行重构，并基于镜像下降方法设计了一种适用于约束域的隐变量推断算法。进一步地，以图神经网络为例，将对该方法进行具体化说明，并在 RKHS 框架下，提出了一种全新的图神经网络的图结构推断方法，并严格证明了该方法的收敛性。最后，实验部分从隐变量推断的准确性、软测量建模精度等多个层面进行了系统性验证，充分证实了所提出方法的有效性和优越性。

关键词： 约束优化　镜像下降　再生核希尔伯特空间　软测量建模

5.1　引言

尽管第4章通过引入无穷时域最优控制策略对概率隐变量模型的建模框架进行了系统性重构，但其对隐变量 z 的分布支撑域做出了一个显式假设——即最优分布 $\mathcal{P}(z|x)$ 和逼近分布 $\mathcal{Q}(z)$ 均支撑在 $\mathbb{R}^{\mathrm{D_{LV}}}$ 上。这一假设虽然在大部分概率隐变量模型中都能成立，但并非对所有概率隐变量模型都成立。例如，在变压器网络或图神经网络（graph neural network，GNN）中，若将归一化图矩阵或自注意力矩阵作为隐变量进行推断，所得到的隐变量必须满足非负性和行和为 1 的基本约束。在这种情况下，隐变量 z 的分布支撑域不再位于 $\mathbb{R}^{\mathrm{D_{LV}}}$ 上，而是被限制在一个特定的约束域（constrained domain）内，这使得第4章提出的隐变量推断策略无法直接适用。因此，如何在现有框架的基础上进行扩展，提出适用于支撑在约束域的隐变量推断方法，成为本章研究的核心问题。

为此，本章引入镜像下降方法，对第4章提出的概率隐变量推断策略进行系统性重构，提出了一种适用于限制域支撑的概率隐变量推断框架。为具体说明该方法的实际应用，本章以 GNN 的归一化邻接矩阵（normalized adjacency matrix）推断为研究案例，详细阐述了所提出方法的实现过程，并严格推导了其在 GNN 归一化邻接矩阵推断中的理

论收敛性。最后，为全面验证方法的有效性，本章从隐变量推断的准确性、推理效率、下游软测量任务建模精度以及模型收敛性四个维度开展了系统性实验，实验结果充分证实了所提出方法的优越性。

5.2 相关工作回顾与科学问题分析

在构建概率隐变量模型的过程中，隐变量分布的支撑并不总是 $\mathbb{R}^{D_{LV}}$，这一特性在实际应用中具有重要意义。具体而言，在统计学习的早期阶段，以高斯混合模型为代表的有限混合模型中[42]，隐变量通常服从分类分布（categorical distribution），其支撑为有限个整数集合。这种离散支撑的特性使得模型在处理多峰数据时表现出色，并成功应用于许多工业过程的软测量建模任务中，为处于多工况操作条件的质量变量预测提供了可靠的理论基础[178-180]。随着研究的深入，进入深度学习时代，GNN 和变压器网络的兴起进一步推动了这一领域的发展。这些网络中的归一化邻接矩阵和自注意力矩阵需要同时满足非负性和行归一化这两个约束[37,181]，使得其分布表现出迪利克雷分布的特性。这种限制域上的分布特性不仅为模型的结构化建模提供了新的视角，也为高维数据和大规模系统的软测量任务带来了新的挑战和机遇。因此，构建支撑在“限制域”上的概率隐变量模型的推断流程[182-183]，并建立起对应的软测量模型和训练框架，成为隐变量建模理论中的一个重要组成部分。

为此，在过去的几十年中，以 KL 散度作为差异度度量，构建模型的损失函数并训练隐变量模型成为研究的重点目标。例如，Beal[57]针对高斯混合模型的分类分布，提出了变分推断流程，为这一领域奠定了理论基础。在此基础上，Yao 等人[184]从随机变分推断的角度出发，建立了适用于大数据集的高斯混合模型，并开发了对应的并行化框架，显著提升了模型的计算效率。后续研究进一步扩展了这一框架，针对工业过程中的离群点问题、流式数据建模问题以及维数灾问题，分别提出了学生-t 混合模型[185-186]、流式贝叶斯框架[187-188]和混合因子模型[84]，极大地丰富了模型的应用场景和鲁棒性。随着深度学习的兴起，崔琳琳等人[89,179]将高斯混合模型与变分自编码器相结合，通过引入重参数化技术构建了软测量模型的训练目标，为复杂工业系统的建模提供了新的思路。后续研究以摊销变分推断为基础，通过将限制域支撑的隐变量与常规的隐变量模型如非线性状态空间模型[94,99]、慢特征分析模型[98,189]等进行结合，进一步推动了这一领域的发

展。尽管这些工作在统计学习时代和深度学习时代成功解决了诸如多峰数据建模、高维非线性数据特征提取等方面的问题取得了显著进展，这些模型的设计通常需要考虑限制域支撑的概率分布之间的 KL 散度的计算问题，并且在重参数化过程中通常需要涉及到对这类限制域支撑的概率分布的采样过程[190]。这些问题会增加在深度学习后端下模型的实现难度，减少模型的计算效率，甚至出现梯度难以回传的问题[190-192]。因此，根据上述分析，本章旨在解决下列两个科学问题：

（1） **限制域支撑的概率隐变量模型的高效推断策略：**是否可以设计一种推断策略，避免显式计算 KL 散度，同时有效推断限制域内的隐变量分布？

（2） **高效推断策略在概率隐变量模型训练中的实现：**是否能够开发一种具有收敛性保证的隐变量模型训练算法，以实现上述推断策略？

5.3　限制域定义下的图神经网络信息传递机制

根据 Ma 和 Tang 的著作[181]，信息传递神经网络（message passing neural network，MPNN）是一个通用的 GNN 框架，图卷积网络（graph convolution network）[193]，图 SAGE[194]和图注意力网络（graph attention network，GAT）[195]可以视为 MPNN 的特殊变体。具体而言，对于 MPNN，节点 v_i 的特征 F_i 依次由式(5-1)和(5-2)更新。

$$\mathscr{M}_i = \sum_{v_j \in \mathscr{N}(v_i)} \mathrm{Mess}(F_i, F_j, \mathscr{E}_{(v_i,v_j)}) \tag{5-1}$$

$$F_i^{'} = \mathrm{Update}(F_i, \mathscr{M}_i) \tag{5-2}$$

其中，F_i 代表了节点 v_i 的特征，“Mess”是信息传递函数，$\mathscr{N}(v_i)$ 是节点 v_i 的相邻节点的集合，其由描述工业过程观测数据观测变量之间关系的邻接矩阵 $\widehat{\mathcal{A}} \in \{0,1\}_{\mathrm{D_{PV}} \times \mathrm{D_{PV}}}$ 决定，$\mathscr{E}$ 为边特征，而“Update”代表更新函数，表示对节点 v_i 的特征进行的更新。通过结合原始特征和来自邻居的聚合消息，来自节点 v_i 邻居的消息由消息函数 Mess 生成，并传递给节点 v_i。通过改变求和算子（$\sum$）、消息函数（Mess）和更新函数（Update），MPNN 可以等效于其他 GNN 的结果实现对数据的空间特征的提取。

需要指出的一点是，在 GNN 的推理阶段，通常对需要对邻接矩阵 $\widehat{\mathcal{A}}$ 进行行归一化[181,193]。具体来说，可以根据图5.1给出的示意图描绘这个计算流程，其中对角矩阵 D

被称为度矩阵（degree matrix），对角元素是 $\widehat{\mathcal{A}}$ 的行和，$\mathcal{A}$ 是归一化邻接矩阵（normalized adjacency matrix，NAM），其计算流程为 $\mathcal{A} = D^{-0.5}\widehat{\mathcal{A}}D^{-0.5}$，图中的“RowSum”和“MakeDiag”分别代表求行和构造对角矩阵。归一化操作通过保持节点特征数量级的统一，进而给 GNN 带来更稳定和有效的学习流程；鉴于图神经网络在工业过程建模的广泛应用特性及其特殊的结构要求，本章以 GNN 中归一化邻接矩阵作为隐变量，并将归一化邻接矩阵的推断问题为研究问题，探讨限制域隐变量推断方法的推导。

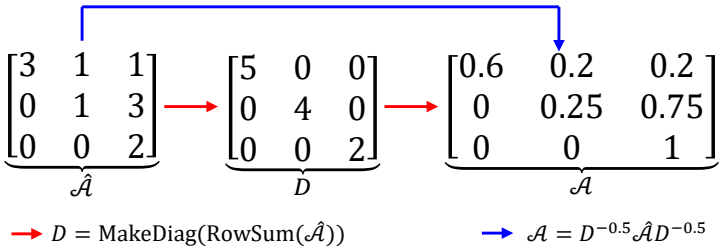


图 5.1 GNN 中归一化邻接矩阵 $\mathcal{A}$ 的计算流程示意图

5.4　限制域隐变量推断框架：基于归一化邻接矩阵的研究

本节将以图神经网络的归一化连接矩阵 $\mathcal{A}$ 视为隐变量，并以其推断作为研究问题以说明本章所提出的在限制域上进行隐变量推断的方法。

5.4.1　问题阐述

考虑以下问题：给定 D 维过程观测数据集 $\{x_i | x_i = [u_i, y_i] \in \mathbb{R}^{\mathrm{D}\times 1}, u_i \in \mathbb{R}^{\mathrm{D_{PV}}\times 1}, y_i \in \mathbb{R}^{\mathrm{D_{QV}}\times 1}, i = 1, \ldots, \mathrm{N}\}$。本章需要利用图神经网络 $p_{\hat{\theta}}(y|x)$ 对该数据进行软测量建模。其中描述过程变量 $u_i|_{i=1}^{\mathrm{D_{PV}}}$ 归一化邻接矩阵 $\mathcal{A} \in \mathbb{R}^{\mathrm{D_{PV}}\times\mathrm{D_{PV}}}$（模型的参数集 $\hat{\theta}$ 定义为 $\widehat{\theta} = \{\mathcal{A}\} \cup \theta$）。在进行建模的过程中，需要归一化邻接矩阵 $\mathcal{A}$ 的分布进行推断。其中归一化邻接矩阵 $\mathcal{A}$ 所属的最优分布 $\mathcal{P}(\mathcal{A}|x)$ 样本需要满足如下两点要求：

- **非负性：** $\mathcal{A}_{i,j} \geq 0,\ \text{for}\ i = 1, \ldots, \mathrm{D_{PV}}, j = 1, \ldots, \mathrm{D_{PV}}$；
- **行归一化性：** $\sum_{j=1}^{\mathrm{D_{PV}}} \mathcal{A}_{i,j} = 1,\ \text{for}\ i = 1, \ldots, \mathrm{D_{PV}}$。

此外，为了简化推导过程，本章假设归一化邻接矩阵 $\mathcal{A}$ 的先验分布 $\mathcal{P}(\mathcal{A})$ 为均匀分布，换言之，概率密度函数 $\log \mathcal{P}(\mathcal{A})$ 是一个常数。

5.4.2 研究动机分析

注意到，归一化邻接矩阵 $\mathcal{A}$ 的每一行之间相互独立，因此，$\mathcal{A}$ 可以被分解为 $\mathcal{A}=\left[\alpha_1^\top,\ldots,\alpha_{\mathrm{D_{PV}}}^\top\right]^\top$，其中 $\alpha_i|_{i=1}^{\mathrm{D_{PV}}}\in\mathbb{R}^{\mathrm{D_{PV}}\times 1}$。在此基础上，针对 $\mathcal{A}$ 的所属分布的推断可以转化为针对 $\alpha_i|_{i=1}^{\mathrm{D_{PV}}}$ 的所属分布的推断，因此5.4.1节对 $\mathcal{A}$ 的两点要求可以转化为如下：

- **非负性：** $\alpha_{i,j}\geq 0,\ \text{for}\ i=1,\ldots,\mathrm{D_{PV}},j=1,\ldots,\mathrm{D_{PV}}$；
- **行归一化性：** $\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}=1,\ \text{for}\ i=1,\ldots,\mathrm{D_{PV}}$。

可以发现 $\alpha_i|_{i=1}^{\mathrm{D_{PV}}}$ 属于迪利克雷分布（记为 Dir），其支撑为单纯形 $\Delta^{\mathrm{D_{PV}}-1}$，在此基础上 $\mathcal{P}(\mathcal{A}|x)$ 可以通过下式进行因子化（factorization）：

$$\mathcal{P}(\mathcal{A}|x)=\prod_{i=1}^{\mathrm{D_{PV}}}\mathcal{P}(\alpha_i|x). \tag{5-3}$$

因此，对归一化邻接矩阵 $\mathcal{P}(\mathcal{A}|x)$ 的推断问题可以表述为对 $\mathcal{A}$ 的每一行分布 $\mathcal{P}(\alpha_i|x),i=1,\ldots,\mathrm{D_{PV}}$ 的推断问题，与之相对应地，第5.4.1小节中的假设：先验分布 $\mathcal{P}(\mathcal{A})$ 为均匀分布，则可以等价转化为先验分布 $\mathcal{P}(\alpha_i),i=1,\ldots,\mathrm{D_{PV}}$ 为均匀分布。此外，由于 $\mathcal{A}$ 各行之间的独立性，因此本章后续内容将关注于 α_i 的推断过程。

根据上述分析，根据第2.1.2小节所介绍的方法，一个很直接的方式是引入神经网络 $q_\varphi(\alpha_i|u)$ 对分布 $\mathcal{P}(\alpha_i|x)$ 进行逼近，通过构建如下的损失函数，即可以对网络进行训练，从而构建软测量模型：

$$\begin{aligned}&\arg\min_{\theta,\varphi}\ \ \mathbb{E}_{q_\varphi(\alpha_i|u)}\left[\log p_\theta(y|u,\alpha_i)\right]+\sum_{i=1}^{\mathrm{D_{PV}}}\mathbb{D}_{\mathrm{KL}}\left[q_\varphi(\alpha_i|u)\|\mathcal{P}(\alpha_i)\right]\\ \overset{\text{(i)}}{\Rightarrow}&\arg\min_{\theta,\varphi}\ \ \mathbb{E}_{q_\varphi(\alpha_i|u)}\left[\log p_\theta(y|u,\alpha_i)\right]+\sum_{i=1}^{\mathrm{D_{PV}}}\mathbb{E}_{q_\varphi(\alpha_i|u)}\left[\log q_\varphi(\alpha_i|u)\right],\end{aligned} \tag{5-4}$$

其中步骤“(i)”基于假设：先验分布 $\mathcal{P}(\alpha_i)$ 为均匀分布。可以发现，直接针对上式构建隐变量模型的训练过程存在负熵项 $\mathbb{E}_{q_\varphi(\alpha_i|u)}\left[\log q_\varphi(\alpha_i|u)\right]$ 的计算问题。具体而言，正如上文所阐述的那样，$q_\varphi(\alpha_i|u)$ 应属于迪利克雷分布，而以 $\theta^{\mathrm{Dir},1}$ 为参数的迪利克雷分布的熵泛函 $\mathbb{H}[\mathrm{Dir}(\theta^{\mathrm{Dir},1})]$ 的计算表达式如下所示：

$$\begin{aligned}&\mathbb{H}[\mathrm{Dir}(\theta^{\mathrm{Dir},1})]\\ =&\log[\frac{\Gamma(\sum_{l=1}^{\mathrm{D_{PV}}}\theta_l^{\mathrm{Dir},1})}{\prod_{l=1}^{\mathrm{D_{PV}}}\Gamma(\theta_l^{\mathrm{Dir},1})}]+\sum_{l=1}^{\mathrm{D_{PV}}}(\theta_l^{\mathrm{Dir},1}-1)\mathrm{DiGa}(\theta_l^{\mathrm{Dir},1})-(\sum_{l=1}^{\mathrm{D_{PV}}}\alpha_i-\mathrm{D_{PV}})\mathrm{DiGa}(\sum_{l=1}^{\mathrm{D_{PV}}}\alpha_i),\end{aligned} \tag{5-5}$$

其中，$\Gamma(\theta)$ 是伽马函数（gamma function），其基本表达式如下所示：

$$\Gamma(\theta) \coloneqq \theta e^{\gamma\theta} \prod_{n=1}^{\infty} [(1+\frac{\theta}{n}) \exp(-\frac{\theta}{n})], \tag{5-6}$$

DiGa 是双伽马函数（digamma function），基本定义式如下：

$$\mathrm{DiGa}(\theta) = \frac{\mathrm{d}\Gamma(\theta)}{\mathrm{d}\theta}. \tag{5-7}$$

通过观察式(5-5)，(5-6)和(5-7)可以发现迪利克雷分布的熵泛函计算过程需要显示地计算伽马函数和双伽马函数等复杂函数，有较高的计算成本，并且其难以适应如 PyTorch[139]和 JAX[140]为代表的深度学习后端。因此，如何显式地规避迪利克雷分布的熵泛函计算的同时实现 $\alpha_i|_{i=1}^{\mathrm{D_{PV}}}$ 的所属分布的推断是本章需要进行解决的核心科学问题。

5.4.3 限制域隐变量推断方法

正如前文所述，从直觉上，可以通过对变分分布 $\mathcal{Q}(\alpha_i)$ 进行“量子化”离散，对第4章导出的式(4-51)进行模拟，即可以在规避显式地 KL 散度计算的同时，对 KL 散度进行下降，从而实现对 $\mathcal{P}(\alpha_i|x)$ 的推断。但是在图神经网络这一类建模框架下，直接执行这一模拟策略会存在以下两个问题：

- **隐式性：**区别于隐变量 z 可以直接解码产生特征的常规概率隐变量模型，根据5.3节，在图神经网络这一类深度学习结构中，归一化邻接矩阵的 $\mathcal{A}$ 并不直接解码质量变量 y，而是要与过程变量 u 共同作用，将 M 个样本得到的特征进行平均后，才能给出 y 的预测。换言之，如果要以第4章介绍的“量子化”的策略表示 $\mathcal{Q}(\alpha_i)$，则变分推断中的对数似然函数应该写为：

$$\arg\max_{\theta} \underbrace{\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}} \mathscr{G}(\alpha_{i,l},u),u], \alpha_{i,l} \overset{\text{i.i.d.}}{\sim} \mathcal{Q}(\alpha_i),}_{\approx \log p_\theta[\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u]} \tag{5-8}$$

其中 $\mathscr{G}$ 为基于信息传递机制设计的非线性算子，约等号成立的理论基础是蒙特卡洛近似。在训练的过程中需要极大化的似然项为 $\log p_\theta[y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u]$ 而不是 $\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log p_\theta[y|\mathscr{G}(\alpha_i,u),u]]$。这种由于神经网络结构的引入导致的不一致性在本文被称为“隐式性”。这种隐式性对式(5-4)这一类基于重参数化为基础构建的训练流程并不会有太大影响，因为导数算符作用在参数 φ 上。但是在第4章的 InfO 算

法的推断框架下，导数算符是作用在变量 α_i 上的。最终这种隐式性导致无法直接以 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 为目标（其中 $\mathcal{P}(\alpha_i|x) \propto p_\theta[\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u]\mathcal{P}(\alpha_i)$）构造对 $\mathcal{Q}(\alpha_i)$ 的推断流程。

- **良定性：**第4章导出的迭代流程可以保证每一步迭代都使得 z 处于实数域 $\mathbb{R}^{\mathrm{D}}$ 内，而不能保证 α_i 处在单纯形 $\Delta^{\mathrm{D_{PV}}-1}$ 上。

因此，如何克服隐式性（即给出减少 KL 散度 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 的迭代式）和良定性（保证迭代过程维持在单纯形 $\Delta^{\mathrm{D_{PV}}-1}$）两个核心问题，从而构建 $\mathcal{Q}(\alpha_i)$ 的推理过程，是本节接下来的内容需要着重解决的问题。

5.4.3.1 限制域上的推断策略导出及其解析表达式

可以发现尽管 KL 散度和式(5-8)所给出的似然函数之间的主要差别在于期望项 $\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\cdot]$ 的算符的位置差异，但是二者存在一定的相似性。因此，为了探寻二者的之间的内在联系以寻求合适的变分推断迭代式，本小节首先研究在式(5-8)的基础上通过添加负熵泛函 $-\mathbb{H}[\mathcal{Q}(\alpha_i)]=\mathbb{E}_{\mathcal{Q}(\alpha_i)}\left[\log\mathcal{Q}(\alpha_i)\right]$ 所形成的泛函的最小化策略：

$$\begin{aligned}
&\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)\\
\coloneqq&-\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathcal{G}(\alpha_{i,l},u),u]+\mathbb{E}_{\mathcal{Q}(\alpha_i)}\left[\log\mathcal{Q}(\alpha_i)\right]\\
\overset{\text{(i)}}{\approx}&-\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathcal{G}(\alpha_{i,l},u),u]+\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}[\log\mathcal{Q}(\alpha_{i,l})],
\end{aligned}\tag{5-9}$$

其中步骤“(i)”是基于蒙特卡洛估计，而 ER 是熵正则（entropy regularization）的英文首字母缩写。在此基础上，与4.3.3节类似，对 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 引入无穷小量 τ 和微扰方向 $\phi(\alpha_i):\mathbb{R}^{\mathrm{D_{PV}}}\to\mathbb{R}^{\mathrm{D_{PV}}}$ 对 α_i 进行扰动，$\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 在微扰前后的变化量可以通过如下定理进行给出：

定理 5.1： 在引入无穷小量 ε 和微扰方向 $\phi(\alpha_i):\mathbb{R}^{\mathrm{D_{PV}}}\to\mathbb{R}^{\mathrm{D_{PV}}}$ 对 α_i 进行微扰后，泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 在微扰前后的变化结果可以通过下式进行给出：

$$\begin{aligned}
&\mathcal{L}^{\mathrm{ER}}(y,\alpha_i+\varepsilon\phi(\alpha_i),u)-\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)\\
=&-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\varepsilon\phi^{\top}(\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u)+\log|\det(\mathcal{I}+\varepsilon\nabla_{\alpha_i}\phi(\alpha_i))|].
\end{aligned}\tag{5-10}$$

证明：在扰动前后，$\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 的变化量可以通过下式进行给出：

$$\begin{aligned}
&\mathcal{L}^{\mathrm{ER}}(y,\alpha_i+\varepsilon\phi(\alpha_i),u)-\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)\\
=&-\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l}+\varepsilon\phi(\alpha_{i,l}),u),u]+\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l},u),u]\\
&+\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}[\log\mathcal{Q}(\alpha_{i,l}+\varepsilon\phi(\alpha_{i,l}))]-\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}[\log\mathcal{Q}(\alpha_{i,l})].
\end{aligned}\tag{5-11}$$

其中 $\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l}+\varepsilon\phi(\alpha_{i,l}),u),u]$ 可以化为：

$$\begin{aligned}
&\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l}+\varepsilon\phi(\alpha_{i,l}),u),u]\\
=&\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l},u),u]+\frac{\varepsilon}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\phi^\top(\alpha_{i,l})\nabla_{\alpha_{i,l}}\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l},u),u]\\
&+\mathbf{H.O.T.}(\varepsilon^2)\\
=&\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l},u),u]+\varepsilon\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)].
\end{aligned}\tag{5-12}$$

因此，可以得到如下结论：

$$\begin{aligned}
&-\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l}+\varepsilon\phi(\alpha_{i,l}),u),u]+\log p_\theta[y|\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\mathscr{G}(\alpha_{i,l},u),u]\\
=&-\varepsilon\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)].
\end{aligned}\tag{5-13}$$

与之相对应，考虑将微扰的变化过程记为函数 $\boldsymbol{T}(\alpha_i)=\alpha_i+\varepsilon\phi(\alpha_i)$，微扰前后的概率密度函数分别记为 $\mathcal{Q}(\alpha_i)$ 和 $\mathcal{Q}_{\boldsymbol{T}}(\alpha_i)$，则熵泛函 $\mathbb{H}[\mathcal{Q}_{\boldsymbol{T}}(\alpha_i)]$ 与 $\mathbb{H}[\mathcal{Q}(\alpha_i)]$ 的差值可记为：

$$\begin{aligned}
&\mathbb{H}[\mathcal{Q}_{\boldsymbol{T}}(\alpha_i)]-\mathbb{H}[\mathcal{Q}(\alpha_i)]\\
=&-\int_{\mathcal{M}}\mathcal{Q}_{\boldsymbol{T}}(\alpha_i)\log\mathcal{Q}_{\boldsymbol{T}}(\alpha_i)\mathrm{d}\alpha_i+\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\log\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
=&-\int_{\mathcal{M}}\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{T}(\alpha_i))\log\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{T}(\alpha_i))\mathrm{d}\boldsymbol{T}(\alpha_i)+\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\log\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
=&-\int_{\mathcal{M}}\frac{\mathcal{Q}(\alpha_i)}{|\det(\nabla_{\alpha_i}\boldsymbol{T}(\alpha_i))|}\log\frac{\mathcal{Q}(\alpha_i)}{|\det(\nabla_{\alpha_i}\boldsymbol{T}(\alpha_i))|}|\det(\nabla_{\alpha_i}\boldsymbol{T}(\alpha_i))|\mathrm{d}\alpha_i\\
&+\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\log\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
=&-\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\log\frac{\mathcal{Q}(\alpha_i)}{|\det(\nabla_{\alpha_i}\boldsymbol{T}(\alpha_i))|}\mathrm{d}\alpha_i+\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\log\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
=&\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\log|\det(\nabla_{\alpha_i}\boldsymbol{T}(\alpha_i))|\mathrm{d}\alpha_i\\
=&\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log|\det(\mathcal{I}+\varepsilon\nabla_{\alpha_i}\phi(\alpha_i))|].
\end{aligned}\tag{5-14}$$

将式(5-14)和(5-13)代入式(5-11)，即可得式(5-10)。　　**证毕**

紧接着，考察 $\varepsilon \to 0$ 时，可以得到如下的泛函导数（functional derivative）：

$$
\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{\mathrm{d}[\mathcal{L}^{\mathrm{ER}}(y,\alpha_i+\varepsilon\phi(\alpha_i),u)-\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)]}{\mathrm{d}\tau}\\
=&\lim_{\varepsilon\to 0}\frac{\mathrm{d}-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\tau\phi^{\top}(\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)-\log|\det(\mathcal{I}+\tau\nabla_{\alpha_i}\phi(\alpha_i))|]}{\mathrm{d}\tau}\\
=&-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\phi^{\top}(\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]\\
&\quad+\lim_{\varepsilon\to 0}\frac{\mathrm{d}\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log|\det(\mathcal{I}+\tau\nabla_{\alpha_i}\phi(\alpha_i))|]}{\mathrm{d}\tau}\\
=&-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\phi^{\top}(\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]\\
&\quad+\lim_{\varepsilon\to 0}\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\frac{1}{|\det(\mathcal{I}+\tau\nabla_{\alpha_i}\phi(\alpha_i))|}\frac{\mathrm{d}(\mathcal{I}+\varepsilon\phi(\alpha_i))}{\mathrm{d}\tau}]\\
=&-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\phi^{\top}(\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}\phi(\alpha_i)].
\end{aligned}
\tag{5-15}
$$

在此基础上，通过模拟式(5-15)所给出的泛函导数即可对泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 进行最小化。但是，需要指出的是式(5-15)所给出的泛函导数并没有给出一个解析的 $\phi(\alpha_i)$ 的表达式，因此并没有办法直接优化泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$。为此，可以考虑寻求式(5-15)的一个下界，通过最大化该下界，从而寻求泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 的导数的解析表达式。接下来通过下列定理给出泛函导数的表达式：

定理 5.2： 当微扰方向 $\phi(\alpha_i)$ 处于 $\mathrm{D_{PV}}$ 维 RKHS（记为 $\mathscr{H}^{\mathrm{D_{PV}}}$）时，并且 $\mathscr{H}^{\mathrm{D_{PV}}}$ 的核函数 $K(\alpha_i',\alpha_i):\mathbb{R}^{\mathrm{D_{PV}}}\to\mathbb{R}^{\mathrm{D_{PV}}}$ 满足边界条件 $\lim_{\|\alpha_i\|\to\infty}K(\alpha_i',\alpha_i)=0$ 时，式(5-15)的一个具有解析表达式的下界为：

$$
\phi_{\mathrm{RKHS}}(\alpha_i)=-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[K^{\top}(\alpha_i',\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}K(\alpha_i',\alpha_i)]
\tag{5-16}
$$

证明： 当微扰方向 $\phi(\alpha_i)\in\mathscr{H}^{\mathrm{D_{PV}}}$ 时，式(5-15)可以重构为：

$$
-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\phi^{\top}(\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}\phi(\alpha_i)]-\frac{1}{2}\|\phi(\alpha_i)\|^2_{\mathscr{H}^{\mathrm{D_{PV}}}}
\tag{5-17}
$$

而对于核函数 $K(\alpha_i',\alpha_i)$，可以定义其形式如 $\xi(\alpha_i):\mathbb{R}^{\mathrm{D_{PV}}}\to\mathscr{H}$ 的特征映射，从而将核函数分解为 $K(\alpha_i',\alpha_i)=\langle\xi(\alpha_i'),\xi(\alpha_i)\rangle_{\mathscr{H}^{\mathrm{D_{PV}}}}$。在此基础上，进一步将核函数的谱分解 $K(\alpha_i',\alpha_i)=\sum_{i=1}^{\infty}\Lambda_i\Xi_i(\alpha_i')\Xi_i(\alpha_i)$（其中 Λ_i 和 $\Xi_i:\mathbb{R}^{\mathrm{D_{PV}}}\to\mathbb{R}$ 分别是特征值和归一化正交

基）应用在微扰方向 $\phi(\alpha_i) \in \mathscr{H}^{\mathrm{D_{PV}}}$ 上可以得到 $\phi(\alpha_i) = \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\Lambda_i} \Xi_i(\alpha_i)$。其中特征重要性权重满足 $\widehat{\psi}_i \in \mathbb{R}^{\mathrm{D_{PV}}}$，特征正交基满足 $\sum_{i=1}^{\infty} \|\widehat{\psi}_i\|_2^2 < \infty$。因此式(5-17)可以重构为下式：

$$\begin{aligned} \underset{\widehat{\psi}_i \in \mathscr{H}^{\mathrm{D_{PV}}}}{\arg\max} \; & -\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\sum_{i=1}^{\infty} \sqrt{\Lambda_i} \nabla_{\alpha_i} \log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i, u)], u) \widehat{\psi}_i \Xi_i(\alpha_i)] \\ & - \mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i} \cdot \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\Lambda_i} \Xi_i(\alpha_i)] - \frac{1}{2} \sum_{i=1}^{\infty} \|\widehat{\psi}_i\|_{\mathscr{H}^{\mathrm{D_{PV}}}}^2 . \end{aligned} \tag{5-18}$$

在此基础上，最优的特征重要性权重 $\widehat{\psi}_i^*$ 可以通过式(5-18)对 $\widehat{\psi}_i$ 求偏导并使之等于 0 进行给出，$\widehat{\psi}_i^*$ 的具体表达式如下：

$$\widehat{\psi}_i^* = -\sqrt{\Lambda_i} \mathbb{E}_{\mathcal{Q}(\alpha_i)} \left[\nabla_{\alpha_i} \Xi(\alpha_i) + \nabla_{\alpha_i} \log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i, u)], u) \Xi(\alpha_i) \right]. \tag{5-19}$$

将式(5-19)代入 $\phi(\alpha_i) = \sum_{i=1}^{\infty} \widehat{\psi}_i \sqrt{\Lambda_i} \Xi_i(\alpha_i)$，可以得到：

$$-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[K^\top(\alpha_i', \alpha_i) \nabla_{\alpha_i} \log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i, u)], u)] - \mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i} K(\alpha_i', \alpha_i)].$$

则式(5-16)得证。 **证毕**

通过观察上述定理，发现有如下注释：

注释 5：式(5-20)所给出的泛函导数的实现流程中中与概率密度函数 $\mathcal{Q}(\alpha_i)$ 有关的运算仅仅出现在期望项 $\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\cdot]$ 中。由于该期望项可通过蒙特卡洛模拟近似，即 $\mathbb{E}_{\mathcal{Q}(\alpha_i)}[f(\alpha_i)] \approx \frac{1}{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} f(\alpha_{i,j})$，其中 $\alpha_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{Q}(\alpha_i)$，从而降低了式(5-10)所定义的泛函导数的实现难度。

此外，由于在定理中引入了边值条件 $\lim_{\|\alpha_i\| \to \infty} K(\alpha_i', \alpha_i) = 0$，因此本章采用了与第3章类似的策略将 RBF 函数作为核函数：

$$K(\alpha_i, \alpha_i') \coloneqq \exp(-\frac{\|\alpha_i - \alpha_i'\|_2^2}{2v}).$$

由于 α_i 处于单纯形 $\Delta^{\mathrm{PV}-1}$ 上，因此除非特别说明，带宽 v 在本章中被设置为 1。

接下来将探究式(5-16)给出的表达式能否作为下降 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(\alpha_i) \| \mathcal{P}(\alpha_i|x)]$ 的导数方向。为了说明这一个问题，本节首先引入如下引理给出 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(\alpha_i) \| \mathcal{P}(\alpha_i|x)]$ 下降的充分条件：

引理 5.1： 微扰 $\phi(\alpha_i)$ 使 KL 散度 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 下降的一个充分条件是：

$$\mathbb{E}_{\mathcal{Q}(\alpha_i)}\left\{\phi^{\top}(\alpha_i)[\nabla_{\alpha_i}\log p_{\theta}(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)-\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)]\right\}\geq 0. \tag{5-20}$$

证明： 首先回顾式(2-27)，可以得到 KL 散度 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 随 τ 演化的动力系统：

$$\begin{aligned}
&\frac{\partial\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]}{\partial\tau}\\
=&\int_{\mathcal{M}}\frac{\partial\mathcal{Q}(\alpha_i)}{\partial\tau}\log\frac{\mathcal{Q}(\alpha_i)}{\mathcal{P}(\alpha_i|x)}+\mathcal{Q}(\alpha_i)\frac{\partial\log\mathcal{Q}(\alpha_i)}{\partial\tau}-\mathcal{Q}(\alpha_i)\frac{\partial\log\mathcal{P}(\alpha_i|x)}{\partial\tau}\mathrm{d}\alpha_i\\
=&\int_{\mathcal{M}}\frac{\partial\mathcal{Q}(\alpha_i)}{\partial\tau}[\log\frac{\mathcal{Q}(\alpha_i)}{\mathcal{P}(\alpha_i|x)}+1]\mathrm{d}\alpha_i\\
=&\int_{\mathcal{M}}[-\nabla\cdot(\mathcal{Q}(\alpha_i)\phi(\alpha_i))][\log\frac{\mathcal{Q}(\alpha_i)}{\mathcal{P}(\alpha_i|x)}+1]\mathrm{d}\alpha_i\\
=&\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\phi^{\top}(\alpha_i)[\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)-\nabla_{\alpha_i}\log\mathcal{P}(\alpha_i|x)]\mathrm{d}\alpha_i\\
=&\mathbb{E}_{\mathcal{Q}(\alpha_i)}\{\phi^{\top}(\alpha_i)[\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)-\nabla_{\alpha_i}\log\mathcal{P}(\alpha_i|x)]\}\\
=&\mathbb{E}_{\mathcal{Q}(\alpha_i)}\{\phi^{\top}(\alpha_i)[\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)-\nabla_{\alpha_i}\log p_{\theta}(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]\}.
\end{aligned}\tag{5-21}$$

在此基础上，一个让 KL 散度 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 随着 τ 的演化单调递减的一个充分条件如下：

$$\mathbb{E}_{\mathcal{Q}(\alpha_i)}\{\phi^{\top}(\alpha_i)[\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)-\nabla_{\alpha_i}\log p_{\theta}(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]\}\leq 0, \tag{5-22}$$

故式(5-20)得证。 **证毕**

在此基础上，引入如下定理说明式(5-16)给出的微扰方向可以使 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 随着 τ 的演化单调下降，进而实现对 $\mathcal{P}(\alpha_i|x)$ 的推断：

定理 5.3： 微扰方向 $\phi_{\mathrm{RKHS}}(\alpha_i)$ 可以使 $\mathbb{D}_{\mathrm{KL}}\left[\mathcal{Q}(\alpha_i)\|\mathcal{P}(\alpha_i|x)\right]$ 随着 τ 的演化单调下降。

证明：由于式(5-16)解决的是最小化问题，因此首先对 $\phi_{\text{RKHS}}(\alpha_i)$ 取相反数得到如下结果：

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q}(\alpha_i)}[K^\top(\alpha_i',\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]+\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}K(\alpha_i',\alpha_i)]\\
=&\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)K^\top(\alpha,\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)\mathrm{d}\alpha_i\\
&\quad+\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)\nabla_{\alpha_i}K(\alpha_i',\alpha_i)\mathrm{d}\alpha_i\\
\overset{\text{(i)}}{=}&\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)K^\top(\alpha,\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)\mathrm{d}\alpha_i\\
&\quad-\int_{\mathcal{M}}K^\top(\alpha_i',\alpha_i)\nabla_{\alpha_i}\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
=&\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)K^\top(\alpha,\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)\mathrm{d}\alpha_i\\
&\quad-\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)K^\top(\alpha_i',\alpha_i)\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
=&\int_{\mathcal{M}}\mathcal{Q}(\alpha_i)K^\top(\alpha_i',\alpha_i)[\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)-\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)],
\end{aligned}
\tag{5-23}
$$

其中步骤“(i)”应用了分部积分法。在此基础上，将式(5-23)代入式(5-20)可以得到如下结果：

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{Q}(\alpha_i)}\left\{\phi^\top(\alpha_i)[\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)-\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)]\right\}\\
=&\iint_{\mathcal{M}}[\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)-\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)]\\
&\quad\times K^\top(\alpha_i',\alpha_i)\\
&\quad\times[\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)-\nabla_{\alpha_i}\log\mathcal{Q}(\alpha_i)]\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\mathcal{Q}(\alpha_i)\mathrm{d}\alpha_i\\
\overset{\text{(ii)}}{\geq}&0.
\end{aligned}
\tag{5-24}
$$

其中步骤“(ii)”的导出是基于核函数的半正定性（positive semi-definite），$K(\alpha_i',\alpha_i)\succeq 0$，因此式(5-16)给出的微扰方向满足式(5-20)，则定理得证。　　**证毕**

需要指出的是，本小节在推导过程中引入的微扰方向 $\phi(\alpha_i)$ 的签名（signature）是 $\mathbb{R}^{\text{D}_{\text{PV}}}\to\mathbb{R}^{\text{D}_{\text{PV}}}$，并不能维持 α_i 在推断过程中的良定性。为了保证推断得到的样本始终维持在单纯形 $\Delta^{\text{D}_{\text{PV}}-1}$ 上，接下来将引入2.2.1节介绍的镜像下降技术对迭代过程进行修改。

5.4.3.2 基于镜像下降方法的限制域良定性保持策略

本小节将对如何保持 α_i 在迭代过程中的良定性进行分析。考虑到 $\sum_{j=1}^{\text{D}_{\text{PV}}}\alpha_{i,j}=1$，且 $\alpha_{i,j}\geq 0$，可以被视为一种特殊的在“概率分布”，因此根据 Shi 等人的论文[183]，本节

考虑采用式(5-25)给出的熵泛函作为布雷格曼势，进行镜像下降迭代流程的构建：

$$\Psi(\alpha_i)=\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}\log(\alpha_{i,j})-\alpha_{i,j}, \tag{5-25}$$

根据式(5-25)，式(5-25)对应的布雷格曼散度可以重构为：

$$\mathbb{B}_{\text{entropy}}[\alpha_i\|\alpha_{i,\tau}]=\sum_{j=1}^{\mathrm{N_{PV}}}\alpha_{i,j}\frac{\alpha_{i,j}}{\alpha_{i,j,\tau}}-\alpha_{i,j}+\alpha_{i,j,\tau}. \tag{5-26}$$

而式(5-26)可以重构为式所示的优化命题：

$$\begin{aligned}\underset{\alpha_i\in\Delta^{\mathrm{D_{PV}}-1}}{\arg\min}\quad & \mathcal{L}\coloneqq\frac{1}{\tau}\sum_{j=1}^{\mathrm{N_{PV}}}\alpha_{i,j}\frac{\alpha_{i,j}}{\alpha_{i,j,\tau}}-\alpha_{i,j}+\alpha_{i,j,\tau}+[-\phi_{\text{RKHS}}(\alpha_{i,j,\tau})]^{\top}\alpha_i\\ \text{s.t.}\quad & \begin{cases}\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}=1\\ \alpha_{i,j}\geq 0\end{cases}\end{aligned} \tag{5-27}$$

在此基础上，可以通过下列定理对式(5-27)所定义的优化问题进行求解以保证 α_i 在迭代过程的对限制域 $\Delta^{\mathrm{D_{PV}}-1}$ 的良定性：

定理 5.4： 保证 α_i 处于单纯形 $\Delta^{\mathrm{D_{PV}}-1}$ 上的迭代过程可以通过下列方程组给出：

$$\begin{cases}\log\alpha_{i,\tau+1}=\log\alpha_{i,\tau}+\varepsilon\phi_{\text{RKHS}}(\alpha_{i,\tau})+\lambda_i\\ \exp(-\lambda_i)=\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}[\exp(\varepsilon\phi_{\text{RKHS}}(\alpha_{i,\tau}))]_j.\end{cases} \tag{5-28}$$

证明： 考虑式(5-27)所给出的优化命题，引入拉格朗日乘子 $\lambda_i\in\mathbb{R}$ 和 $\beta_i\in\mathbb{R}^{\mathrm{D_{PV}}\times 1}$ 分别处理式(5-27)中的等式约束和不等式约束可以得到如下的无约束优化目标函数：

$$\mathcal{L}^{\text{uncons}}=\frac{1}{\tau}\sum_{j=1}^{\mathrm{N_{PV}}}\alpha_{i,j}\frac{\alpha_{i,j}}{\alpha_{i,j,\tau}}-\alpha_{i,j}+\alpha_{i,j,\tau}-\varepsilon\phi_{\text{RKHS}}^{\top}(\alpha_{i,\tau})\alpha_i+\lambda_i(\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}-1)+\alpha_i\beta_i, \tag{5-29}$$

其中上标 uncons 代表为无约束（unconstrained）的英文缩写。在此基础上，对 $\mathcal{L}^{\text{uncons}}$ 应用一阶条件 $\nabla_{\alpha_i}\mathcal{L}^{\text{uncons}}=0$，可以得到如下结果：

$$\log\alpha_i=\log\alpha_{i,\tau}+\varepsilon\phi_{\text{RKHS}}(\alpha_{i,\tau})+\lambda_i-\beta_i. \tag{5-30}$$

换言之，最优 α_i 的解满足下式：

$$\alpha_i=\exp(\lambda_i)\exp(\log\alpha_{i,\tau}+\varepsilon\phi_{\text{RKHS}}(\alpha_{i,\tau})-\beta_i). \tag{5-31}$$

注意到由于$\exp(\cdot)\geq 0$恒成立,根据优化命题在最优点的“卡鲁什-库恩-塔克条件”(Karush-Kuhn-Tucker conditions)：

$$\begin{cases}\beta_{i,j}\geq 0\\ \alpha_i\beta_i=0\\ \sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}=1\\ [\nabla_{\alpha_i}\mathcal{L}^u]_j=0\end{cases}. \tag{5-32}$$

可以导出下式：

$$\beta_{i,j}=0. \tag{5-33}$$

在此基础上，可以发现$\exp(\lambda_i)$发挥了归一化因子的作用。根据上述分析，基于$\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}=1$可以得到如下的$\lambda_i$的表达式：

$$\exp(-\lambda_i)=\sum_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j}[\exp\varepsilon[\phi_{\mathrm{RKHS}}(\alpha_{i,\tau})]_j. \tag{5-34}$$

根据式(5-31)和(5-34)，则定理得证。　　**证毕**

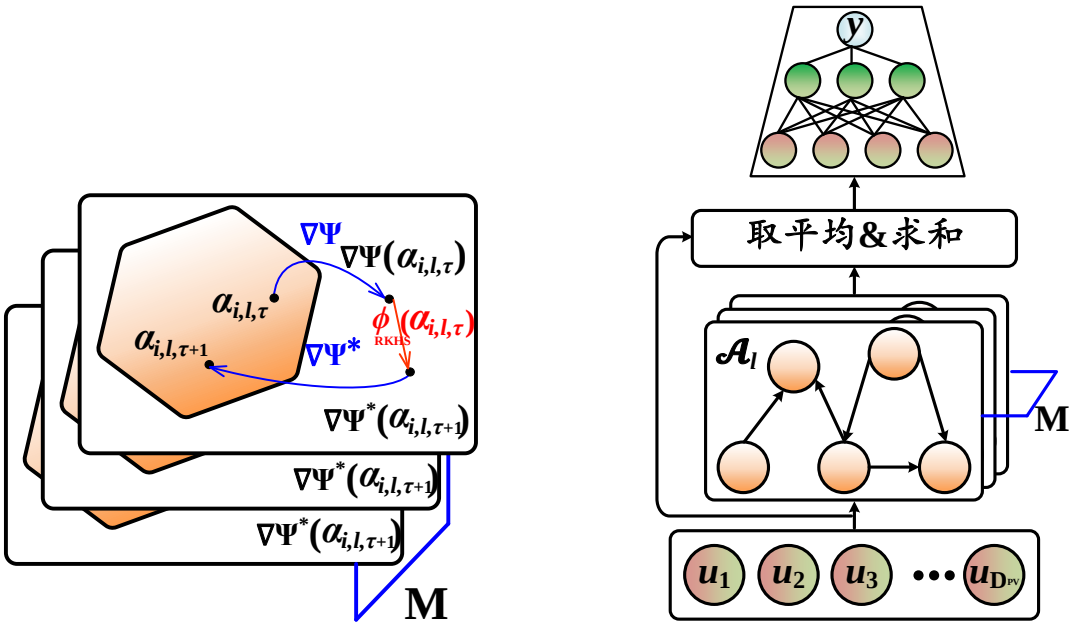
由于该算法基于RKHS和镜像下降技术导出，因此该算法被命名为“基于核镜像下降的单纯形迭代”(Kernelized Mirror Descent-based iteration over Simplex，KEMS）算法。

5.4.4　算法与模型结构总结

尽管5.4.3节提出了基于KEMS算法对图神经网络中归一化邻接矩阵进行变分推断的方法，但并未详细阐述以及KEMS算法的完整流程、KEMS算法收敛性分析以及对应的图神经网络的具体结构。为了弥补这一不足，本节通过图5.2(a)和图5.2(b)分别展示了利用KEMS算法推断$\mathcal{Q}(\alpha_i)$的示意图以及对应的图神经网络模型结构。由于所设计的图神经网络整合了多个图结构以实现特征梯度计算，并在推断过程中通过引入熵正则项对损失函数进行正则化，因此该模型被命名为熵正则集成自适应图（Entropy-regularized Ensemble Adaptive Graph，E^{2}AG）模型。

5.4.4.1 KEMS算法基本流程和收敛性分析

根据5.4.3节的内容，本小节首先在算法5.1中给出KEMS的具体算法流程，其对应的示意图在图5.2(a)中给出。可以看出KEMS的迭代流程图可以分为两步：首先模型



(a) $\mathcal{Q}(\alpha_i)$ 的迭代示意图　　(b) E^{2}AG 模型结构图

图 5.2 $\mathcal{Q}(\alpha_i)$ 的迭代流程和 E^{2}AG 模型结构示意图

算法 5.1 KEMS 算法伪代码

输入： 以 θ 为参数的模型：$p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u)$，在 $\tau=0$ 时刻用以构建起始变分分布 $\mathcal{Q}_0(\alpha_i)=\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}\delta(\alpha_i-\alpha_{i,l})$ 的样本：$\alpha_{i,l}|_{l=1}^{\mathrm{M}}$，终止时间：T，离散步长 ε。

输出： T 时刻的样本：$\alpha_{i,l,\mathrm{T}}|_{l=1}^{\mathrm{M}}$。

初始化样本 $\alpha_{i,l,0}|_{l=1}^{\mathrm{M}} \leftarrow \dfrac{1}{\mathrm{D_{PV}}}$。
for $\tau \leftarrow 0$ **to** $\mathrm{T}-1$ **do**
　$\exp(-\lambda_i) \leftarrow \sum_{j=1}^{\mathrm{D_{PV}}} \alpha_{i,j}[\exp(\varepsilon\phi_{\mathrm{RKHS}}(\alpha_{i,\tau}))]_j$　　> 计算拉格朗日乘子 λ_i
　$\log\alpha_{i,l,\tau+1} \leftarrow \log\alpha_{i,l,\tau}+\varepsilon\phi_{\mathrm{RKHS}}(\alpha_{i,l,\tau})+\lambda_i|_{l=1}^{\mathrm{M}}$　　> 撤回至单纯形 $\Delta^{\mathrm{D_{PV}}-1}$
end for
return $\alpha_{i,l,\mathrm{T}}|_{l=1}^{\mathrm{M}}$

在布雷格曼势诱导的函数 $\nabla_{\alpha_i}\Psi(\alpha_i)$ 的投影下从单纯形 $\Delta^{\mathrm{D_{PV}}-1}$ 投影至实数空间 $\mathbb{R}^{\mathrm{D_{PV}}}$（图5.2(a)中上半部蓝色箭头），并在此基础上受到微扰方向 $\phi_{\mathrm{RKHS}}(\alpha_i)$ 的扰动产生移动（图5.2(a)中红色箭头）。紧接着移动后的结果在布雷格曼势诱导的函数的反函数 $\nabla_{\alpha_i}\Psi^*(\alpha_i)$（图5.2(a)中下半部蓝色箭头）从实数空间 $\mathbb{R}^{\mathrm{D_{PV}}}$ 被拉回至单纯形 $\Delta^{\mathrm{D_{PV}}-1}$ 以保证迭代过程的良定性并完成本次迭代。

由于 α_i 属于模型参数 $\widehat{\theta}$ 的一部分，而关于参数 θ 的收敛性证明已经在第4章中进行了讨论，因此接下来关于收敛性的证明将着重关注在 KEMS 的收敛性证明上。为此，在定义1的基础上，本小节给出如下定理说明 KEMS 算法的收敛性：

定理 5.5： 设 $\{\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}_\tau(\alpha_{i,\tau})\|\mathcal{P}(\alpha_{i,\tau}|x)]\}_{\tau=1}^{\mathrm{T}}$ 为 KEMS 算法生成的迭代序列。若离散步长 ε 充分小，则存在 $\mathcal{C}\geq 0$，使得对于任意给定的 $\gamma>0$，存在正整数 N 使得当 $\tau>N$ 时有 $\|\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(\alpha_{i,\tau})\|\mathcal{P}(\alpha_{i,\tau}|x)]-\mathcal{C}\|<\gamma$ 成立。

证明： 对 KEMS 算法的收敛性证明可以分为以下三个部分:

（1） **单纯形 $\Delta^{\mathrm{D_{PV}}-1}$ 与 $\mathbb{R}^{\mathrm{D_{PV}}}$ 之间的关系：**式(2-13)所给出的镜像下降流程可以重构为下式：

$$\begin{aligned} z_{\tau+1}&=\arg\min_{z\in\mathcal{M}}\ \frac{1}{\varepsilon}\mathbb{B}\left[z\|z_\tau\right]+\left[\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}\right]z \\ \Rightarrow 0&=\nabla_z\boldsymbol{F}(z)|_{z=z_\tau}+\nabla_z\frac{\mathbb{B}\left[z\|z_\tau\right]}{\varepsilon}|_{z=z_\tau} \\ \Rightarrow\frac{\mathrm{d}\nabla\Psi(z)}{\mathrm{d}\tau}&=\lim_{\varepsilon\to 0}\frac{\nabla\Psi(z_{\tau+1})-\nabla\Psi(z_\tau)}{\varepsilon}=-\nabla_z\boldsymbol{F}(z). \end{aligned} \tag{5-35}$$

从上述推导流程可以看出，镜像下降的迭代流程可以看作是用步长为 ε 的前向欧拉法[175]对式(5-35)最后一行的 ODE 的离散化。值得注意的是，对于单纯形约束，$\nabla\Psi(\alpha_i)=\log\alpha_i$，下列相等成立：

$$\begin{aligned} &\frac{\mathrm{d}\alpha_i}{\mathrm{d}\tau}\times\mathrm{MakeDiag}(\alpha_i)^{-1} \\ =&\frac{\mathrm{d}\log\alpha_i}{\mathrm{d}\tau} \\ =&-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[K^\top(\alpha_i',\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)]-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\nabla_{\alpha_i}K(\alpha_i',\alpha_i)], \end{aligned} \tag{5-36}$$

其中“MakeDiag”代表了将一个 $\mathrm{D_{PV}}$ 维的行向量转化为一个 $\mathrm{D_{PV}}$ 维的对角矩阵（diagonal matrix）的函数。

（2） **单调递减条件验证：**由于KEMS在优化泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$，因此考察泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 随时间 τ 的演化过程可以得到如下关系：

$$\begin{aligned} &\frac{\mathrm{d}\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)}{\mathrm{d}\tau} \\ =&[\frac{\mathrm{d}\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)}{\mathrm{d}\alpha_i}][\frac{\mathrm{d}\alpha_i}{\mathrm{d}\tau}]^\top \\ =&[\frac{\mathrm{d}\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)}{\mathrm{d}\alpha_i}]\mathrm{MakeDiag}(\alpha_i)[\frac{\mathrm{d}\log\alpha_i}{\mathrm{d}\tau}]^\top \\ =&-\|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[K^\top(\alpha_i',\alpha_i)\nabla_{\alpha_i}\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathscr{G}(\alpha_i,u)],u)+\nabla_{\alpha_i}K(\alpha_i',\alpha_i)]\|_2^2\times\prod_{j=1}^{\mathrm{D_{PV}}}\alpha_{i,j} \\ \leq&0. \end{aligned} \tag{5-37}$$

观式(5-37)可以发现 α_i 的演化可以降低泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$。

（3） **有界条件：**这一部分验证 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 的有界性。注意到 $-\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u)$ 构建的是软测量任务，因此其损失函数可以写为如下的平均最小二乘函数：

$$\|\widehat{y}-y\|\geq 0. \tag{5-38}$$

此外，注意到 $\mathcal{Q}(\alpha_i)$ 和分布在单纯形 $\Delta^{\mathrm{D_{PV}}-1}$ 上的均匀分布 $\mathcal{U}(\alpha_i)$ 之间的 KL 散度可以转化为下式：

$$\begin{aligned}&\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(\alpha_i)\|\mathcal{U}(\alpha_i)]=-\mathbb{H}[\mathcal{Q}(\alpha_i)]-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log\mathcal{U}(\alpha_i)]\geq 0\\ \Rightarrow&-\mathbb{H}[\mathcal{Q}(\alpha_i)]\geq\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log\mathcal{U}(\alpha_i)]=\log\mathcal{U}(\alpha_i)=\log\prod_{j=1}^{\mathrm{D_{PV}}-1}j^{-1}.\end{aligned}$$

因此，$\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 满足下列不等式：

$$\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)\geq 0-\mathrm{D_{PV}}\sum_{j=1}^{\mathrm{D_{PV}}-1}\log(j)=-\mathrm{D_{PV}}\sum_{j=1}^{\mathrm{D_{PV}}-1}\log(j). \tag{5-39}$$

根据上述不等式，$\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 有界条件得证。

从第（2）和（3）步，可以发现式(5-28)给出的 KEMS 算法的迭代流程对应的 ODE 能够让有下界的泛函 $\mathcal{L}^{\mathrm{ER}}(y,\alpha_i,u)$ 单调递减，此外离散步长 ε 越小，越接近式(5-37)所定义的动力系统，进而 KEMS 算法的收敛性得证。 **证毕**

5.4.4.2 E^{2}AG 模型结构论述

在给出 KEMS 算法的基础上，本节将进一步对图5.2(b)中给出的 E^{2}AG 模型整体结构进行详细说明。在 E^{2}AG 模型中，过程变量 $u\in\mathbb{R}^{1\times\mathrm{N_{PV}}}$ 首先经过嵌入层（embedding layer）得到嵌入层提取的特征 $g_{i,l}^{\mathrm{embed}}$：

$$g_{i,l}^{\mathrm{embed}}=x_i\times W_{i,l}^{\mathrm{PV}}+b_{i,l}^{\mathrm{embed}}|_{i=1}^{\mathrm{N_{PV}}}|_{l=1}^{\mathrm{M}}. \tag{5-40}$$

紧接着，经过消息传递模块，进行消息交换后的特征可以通过下式得到：

$$g^{\mathrm{mess}}=\frac{1}{\mathrm{M}}\sum_{l=1}^{\mathrm{M}}g_l^{\mathrm{embed}}\mathcal{A}_l. \tag{5-41}$$

紧接着更新后的特征 g^{update} 可以通过下式得到：

$$g^{\text{update}} = xW^{\text{id}} + b^{\text{id}} + g^{\text{mess}}. \tag{5-42}$$

最后质量变量通过 L 层的多层感知机以 g^{update} 作为输入进行预测：

$$\widehat{y} = \left\{ \left[...\text{ReLU}(g^{\text{update}}W^1 + b^1) \right] W^{\text{L}} + b^{\text{L}} \right\}, \tag{5-43}$$

其中 ReLU(x) 是整流线性单位函数（rectified linear unit）的缩写，其基本定义式如下：

$$\text{ReLU}(x) \coloneqq \max\{0, x\}. \tag{5-44}$$

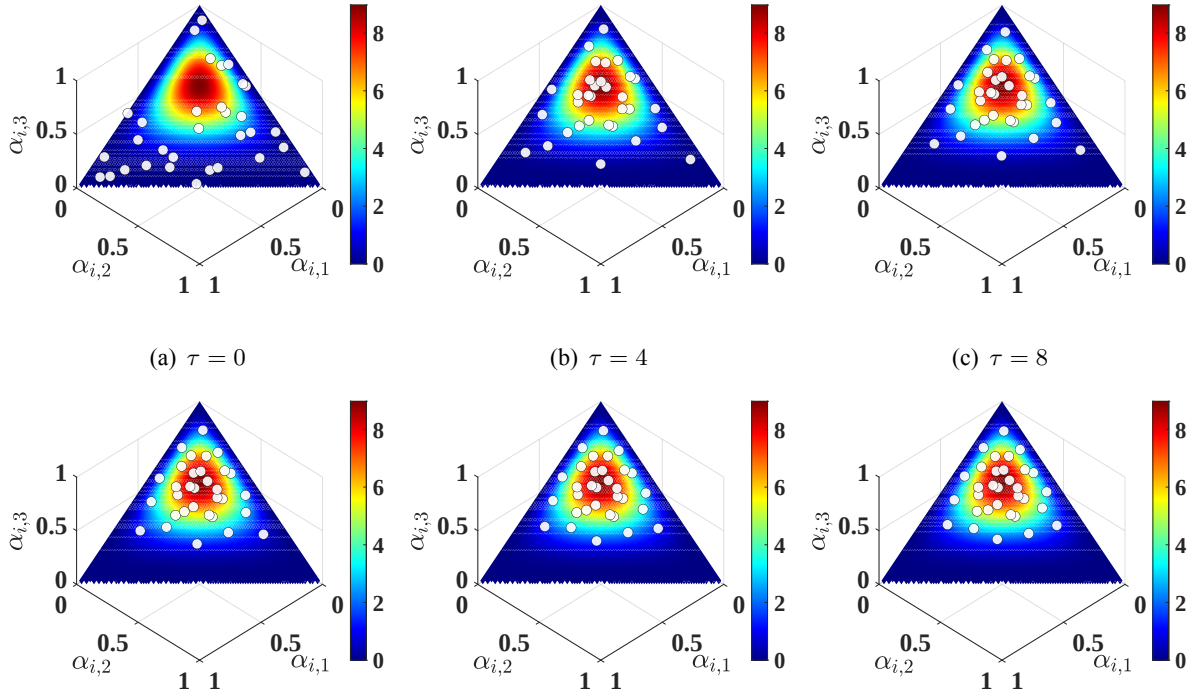
5.5 实验验证

在本节，将对本章所提出的 KEMS 算法和 E^{2}AG 模型的有效性进行实验验证。本节的实验将从如下五个研究问题进行展开（由于 $\mathcal{P}(\alpha_i|x)$ 的支撑为单纯形 $\Delta^{\text{D}_{\text{PV}}-1}$，而 KSD 的假设空间为实数空间 $\mathbb{R}^{\text{D}_{\text{PV}}}$，因此本章不展开基于 KSD 的拟合优度检验实验）：

- **问题 1（有效性）：** KEMS 算法能否有效逼近支撑在单纯形 $\Delta^{\text{D}_{\text{PV}}-1}$ 上的后验分布 $\mathcal{P}(\alpha_i|x)$？
- **问题 2（表现性）：** E^{2}AG 模型在软测量建模任务中的表现性如何？
- **问题 3（增益机理）：** 是什么让 E^{2}AG 模型取得了如此好的表现效果？
- **问题 4（敏感性）：** 随着超参数的调整，E^{2}AG 模型的性能变化趋势如何？
- **问题 5（收敛性）：** KEMS 算法能否在 E^{2}AG 模型的训练过程中收敛？

5.5.1 后验分布逼近过程的可视化分析

本小节聚焦于**问题 1**：“KEMS 算法能否有效地逼近后验分布 $\mathcal{P}(\alpha_i|x)$?”。为回答这一问题，本小节设计了一项针对支撑为三维单纯形 Δ^2 的分布 $\mathcal{P}(\alpha_i|x)$ 的逼近实验，通过可视化密度函数 $\mathcal{Q}(\alpha_i)$ 随 τ 的演化结果，并将其与 $\mathcal{P}(\alpha_i|x)$ 的对数概率密度等高线图进行对比，以定性地验证 KEMS 算法的有效性。为此，设 $\mathcal{P}(\alpha_i|x) \propto \text{Dir}([2.5, 2.5, 5.0])$，图5.3中展示 KEMS 算法对 $\mathcal{P}(\alpha_i|x)$ 的逼近过程。



(a) $\tau = 0$　(b) $\tau = 4$　(c) $\tau = 8$

(d) $\tau = 16$　(e) $\tau = 24$　(f) $\tau = 32$

图 5.3 $\mathcal{Q}(\alpha_i)$（白色点）随着 τ 的演化对分布 $\mathcal{P}(\alpha_i|x)$ 的逼近过程示意图

通过观察图5.3(a)至图5.3(c)，可以发现在演化的初期阶段，$\mathcal{Q}(\alpha_i)$ 在 $\tau = 0$ 时并未与 $\mathcal{P}(\alpha_i|x)$ 的高概率密度区域重合。然而，随着 τ 的持续演化，$\mathcal{Q}(\alpha_i)$ 显示出逐渐向 $\mathcal{P}(\alpha_i|x)$ 的高密度区域聚集的趋势。进一步，从图5.3(d)至图5.3(f)中可以看到，随着 τ 的演化，$\mathcal{Q}(\alpha_i)$ 逐渐与 $\mathcal{P}(\alpha_i|x)$ 的高概率密度区域重合，并沿等高线排列。通过这一实验过程，可以得出结论：本章提出的 KEMS 算法能够快速且精准地逼近支撑在单纯形上的后验分布，并为**问题 1** 提供了明确的答案。

5.5.2　软测量建模精度对比实验

本小节聚焦于**问题 2**：“E^{2}AG 模型在软测量建模任务中的表现如何？”为了回答这一问题，本小节将在水煤气变换单元和二氧化碳吸收单元这两个经典化工过程的反应和分离单元操作中进行软测量建模实验，以展示 E^{2}AG 模型的优越性。

为了更加充分地展示 E^{2}AG 模型的性能，本节选择了以下基于图神经网络设计的软测量模型作为基线模型进行相关实验。这些基线模型根据其结构设计原理可分为以下几个大类：

- **变压器网络类：**快速变压器网络[196]（Flashformer）、转置变压器网络[197]（inverted-Transformer，iTransformer）、随机合成变压器网络[198]（Synthesizer (Random)）和稠密合成变压器网络[198]（Synthesizer (Dense)）；
- **自适应图神经网络类：**动态图深度学习模型[199]（dynamic graph-based deep learning，DGDL）和堆叠图神经网络[200]（stacked graph convolution network，S-GCN）；
- **非自适应图神经网络类：**融入图卷积网络的双流潜在特征分析[201]（two-stream graph convolutional network-incorporated latent feature analysis，TGLFA）、保持对称性的双流图神经网络[202]（symmetry-preserving dual-stream graph neural networks，SDGNN）、自适应图对比学习（adaptive graph contrastive learning，AdaGCL）和随机权重的图卷积网络[203]（graph convolution network with random weight，GCN-RW）。

相关实验超参数在附录D中详细说明，而软测量精度的对比结果则列在表5.1中。表5.1展示了以下实验现象：

（1） 与基于非自适应图神经网络的模型相比，基于变压器网络和自适应图神经网络设计的模型通常在不同的软测量建模场景下表现出更为稳定的性能和更加精确的预测结果。

（2） 在水煤气变换单元数据集上，E^2AG 模型在多个指标上均优于基线模型。具体而言，E^2AG 模型的软测量建模在 R^2 指标上相比大多数基线模型提高了 0.33% 至 40529%，RMSE 降低了 2.48% 至 64.18%，MAPE 减少了 3.42% 至 96.82%，而 MAE 则降低了 3.48% 至 64.2%。

（3） 在 R^2 指标上，几乎所有模型在水煤气变换单元数据集上的表现均优于在二氧化碳吸收单元数据集上的表现。

（4） 在二氧化碳吸收单元数据集上，E^2AG 模型在多个指标上超越了大多数基线模型，具体表现为将 R^2 提高了 0.57% 至 35091%，RMSE 降低了 0.84% 至 51.38%，MAPE 减少了 0.75% 至 46.03%，而 MAE 则降低了 0.75% 至 46.57%。

现象（1）表明，将图神经网络的归一化邻接矩阵作为可学习的变量，能够显著提升模型在工业软测量建模任务中的预报精度和稳定性。这一发现进一步验证了本章将

表 5.1 水煤气变换单元和二氧化碳吸收单元数据集上的软测量建模精度对比结果

模型 / 数据集	水煤气变换单元			
评估指标	R^2	RMSE	MAPE	MAE
Flashformer	6.644E-1	2.699E-1	1.103E-1	2.167E-1
iTransformer	9.395E-1	1.452E-1	5.851E-2	1.151E-1
Synthesizer (Random)	9.306E-1	1.549E-1	6.194E-2	1.218E-1
Synthesizer (Dense)	9.339E-1	1.516E-1	6.110E-2†	1.201E-1†
S-GCN	9.332E-1†	1.527E-1†	6.184E-2†	1.216E-1†
TGLFA	9.377E-1†	1.475E-1†	5.893E-2†	1.159E-1†
SDGNN	8.993E-1†	1.871E-1†	7.317E-2†	1.438E-1†
AdaGCL	5.490E-1†	3.953E-1†	1.579E-1†	3.103E-1†
GNNRW	2.320E-3†	**7.278E-3**†	1.776E+0†	**5.219E-3**†
UniFilter	8.318E-1†	2.413E-1†	9.625E-2†	1.891E-1†
DGDL	9.341E-1	1.516E-1	6.099E-2	1.199E-1
E^2AG	**9.426E-1**	1.416E-1	**5.651E-2**	1.111E-1
超越基线模型数	11	10	11	10
模型 / 数据集	二氧化碳吸收单元			
评估指标	R^2	RMSE	MAPE	MAE
Flashformer	7.597E-1	3.569E-3	9.651E-1	2.808E-3
iTransformer	7.494E-1†	3.646E-3†	9.797E-1	2.854E-3
Synthesizer (Random)	7.427E-1†	3.696E-3†	9.976E-1†	2.905E-3†
Synthesizer (Dense)	7.507E-1	3.636E-3	9.855E-1	2.869E-3
S-GCN	6.553E-1†	4.278E-3†	1.163E+0†	3.387E-3†
TGLFA	6.756E-1†	4.148E-3†	1.110E+0†	3.231E-3†
SDGNN	7.441E-1†	3.684E-3†	**9.423E-1**	**2.750E-3**
AdaGCL	2.081E-1†	6.479E-3†	1.646E+0†	4.827E-3†
GNNRW	2.171E-3†	7.279E-3†	1.775E+0†	5.216E-3†
UniFilter	5.354E-1†	4.966E-3†	1.321E+0†	3.853E-3†
DGDL	7.288E-1†	3.794E-3†	1.023E+0†	2.980E-3†
E^2AG	**7.640E-1**	**3.539E-3**	9.579E-1	2.787E-3
超越基线模型数	11	11	10	10

注：匕首符号 † 表示在成对样本 t-检验中，E^2AG 模型相比其他基线模型在统计学上具有显著性差异（$p < 0.05$）。**粗体字**标注的结果为各指标的最优表现，下划线标注结果为各指标的次优表现。由于部分实验结果处于 10^{-3} 的数量级，为了统一结果格式，表格中的数值均采用科学计数法进行展示。

图神经网络的图结构作为可学习变量的合理性，同时也为本文引入变分推断对归一化邻接矩阵进行推断提供了经验支持。紧接着，现象（2）揭示了软测量建模的性能对数据集的依赖性，这一现象可以从化工单元操作的角度进行解释。具体而言，在氨合成过程中，二氧化碳吸收单元的输入气体可能受到天然气原料中杂质的污染（如硫化氢），从而导致一氧化碳出口浓度受到显著影响，并在数据集中引入大量噪声。相比之下，水煤气变换单元位于靠近产品侧的反应环节，其输入组分更为纯净。因此，在水煤气变换单

元数据集上进行软测量建模所得到的 R^2 分数显著优于二氧化碳吸收单元数据集上的结果。最后，现象（3）和（4）表明，尽管 E^{2}AG 模型属于自适应图神经网络模型，但其性能优于大多数基线模型。这一现象不仅验证了将图神经网络视为一种特殊的概率隐变量模型，在其中引入变分推断机制进行归一化邻接矩阵推断的合理性，同时也进一步凸显了本章所提出的模型和方法的优越性。总而言之，本小节通过对水煤气变换单元和二氧化碳吸收单元的实验分析从实验上证明了 E^{2}AG 模型在软测量建模任务中的优越性，并从实践层面上回答了**问题 2**，为在工业场景下部署 E^{2}AG 提供了实验基础。

5.5.3 消融实验

在5.5.2节的基础上，本小节进一步探讨 E^{2}AG 模型在软测量建模精度方面取得优异表现的原因，并尝试回答**问题 3**：“是什么促成了 E^{2}AG 模型如此出色的效果？”为此，本小节通过消融实验来研究模型的三个关键组成部分，以阐明 E^{2}AG 模型的有效性。具体而言，本小节对以下三个关键模块进行消融实验：

- **镜像下降机制：**在此模块的消融实验中，α_i 的更新从式(5-28)替换为直接采用 $\phi_{\text{RKHS}}(\alpha_i)$ 进行更新；
- **多图逼近：**在此模块的消融实验中，粒子数 M 设定为 1；
- **熵泛函正则机制：**在此模块的消融实验中，负熵泛函正则项 $\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log \alpha_i]$ 被从损失函数 $\mathcal{L}^{\text{ER}}(y, \alpha_i, u)$ 中移除。

实验结果在表5.2中进行了呈现。

首先考虑“镜像下降”模块未被消融的情况。从表5.2中第 1 行至第 3 行的数据可以清晰地看出，消融“多图逼近”和“熵泛函正则化”模块会显著降低模型性能。这一实验结果突显了这两个组成部分在提升模型有效性方面的重要作用。值得注意的是，从第 3 行的结果可以发现，仅消融“多图逼近”模块，其对性能产生的负面影响似乎比同时删除两个模块更为明显。这可以从梯度下降的角度进行解释：当粒子数量减少到 1 个时，将梯度限制在 RKHS 内会导致偏置梯度的产生，最终造成性能下降。相反，当仅消融“熵泛函正则化”模块时，从表5.2中的第 3 行可以看到，模型的性能虽然有所下降，但与第 1 行和第 2 行的结果相比，退化并不明显。这一现象进一步强调了粒子数量在增

表 5.2 二氧化碳吸收单元和水煤气变换单元数据集上的消融实验结果

消融模块			水煤气变换单元			
镜像下降	多图逼近	熵泛函正则	R^2(↓)	RMSE(↑)	MAPE(↑)	MAE(↑)
✓	✗	✗	0.4002%	3.1382%	3.7956%	3.8193%
✓	✗	✓	0.6317%	4.8031%	5.2993%	5.3356%
✓	✓	✗	0.0219%	0.1766%	0.7987%	0.8039%
✗	✗	✗	0.6752%	4.8899%	5.5647%	5.5928%
✗	✗	✓	0.4382%	3.3788%	4.0457%	4.0706%
✗	✓	✗	0.2485%	1.9654%	2.8924%	2.9172%
✗	✓	✓	0.5748%	4.4117%	5.0163%	5.0418%
消融模块			二氧化碳吸收单元			
镜像下降	多图逼近	熵泛函正则	R^2(↓)	RMSE(↑)	MAPE(↑)	MAE(↑)
✓	✗	✗	1.7736%	2.7085%	2.0258%	2.1109%
✓	✗	✓	2.0442%	3.0919%	2.2813%	2.2750%
✓	✓	✗	1.3706%	2.0759%	1.4022%	1.4339%
✗	✗	✗	1.3229%	2.0332%	1.0118%	1.0046%
✗	✗	✓	3.0305%	4.3912%	3.7591%	3.8110%
✗	✓	✗	1.5287%	2.3597%	1.6969%	1.6747%
✗	✓	✓	1.8664%	2.8266%	1.7389%	1.6964%

强 E^2AG 模型性能中的关键作用，同时展示了熵泛函正则化项在归一化邻接矩阵学习过程中的积极影响，进一步从实践上说明了本章将图的归一化邻接矩阵的学习重构为变分推断问题的必要性。

此外，考虑到“镜像下降”模块被消融的情况，从表5.2中第 4 行至第 7 行与第 1 行至第 3 行的结果对比可以明显看出，消融“镜像下降”模块始终会导致性能显著退化，且程度大于保留“镜像下降”模块时的情况。这一现象突显了行归一化邻接矩阵约束在图神经网络模型训练阶段的重要性，并进一步验证了在归一化邻接矩阵的推断中引入镜像下降模块以保持迭代过程稳定性的必要性。综上所述，本小节的消融实验表明，整合多图逼近、熵泛函正则化项和镜像下降模块到 E^2AG 训练过程中的必要性，从实践层面上回答了**问题 3**，同时进一步证实了它们在确保模型最佳性能中的共同重要性。

5.5.4　敏感性分析

在前一小节消融实验的基础上，本小节将进一步探讨**问题 4**：“随着超参数的调整，E^2AG 模型的性能变化趋势如何?”，以深入分析 E^2AG 模型的软测量建模精度与超参数变化之间的关系，从而评估模型对超参数的敏感性为实际应用提供指导。为此，本小节

重点考察了离散步长（学习率）ε、批次大小 $\mathcal{B}$ 以及粒子数量 M 的变化对 E^{2}AG 模型的软测量建模精度的影响。实验结果如图5.4(a)至图5.5(x)所示，其中阴影部分表示 ±0.5 倍标准差。

图5.5展示了以下实验现象：

（1） 随着离散步长 ε 的增加，E^{2}AG 模型的软测量建模精度在两个数据集上都呈现了下降趋势。

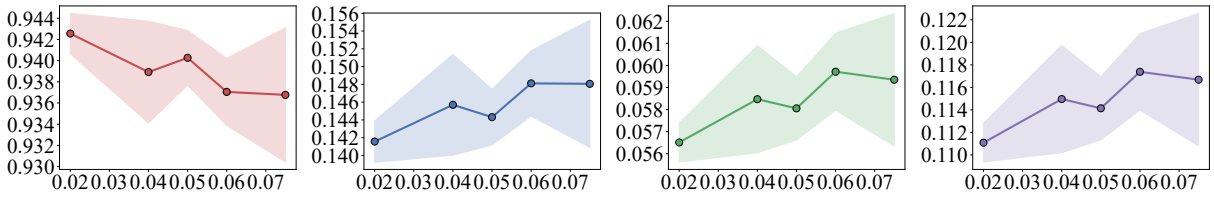
（2） 随着批次大小 $\mathcal{B}$ 的增加，E^{2}AG 模型的软测量建模精度在水煤气变换单元数据集上呈现下降趋势，而在二氧化碳吸收单元数据集上呈现上升趋势。

（3） 随着粒子数量 M 的增加，E^{2}AG 模型的软测量建模精度在两个数据集上呈现先上升后下降趋势。

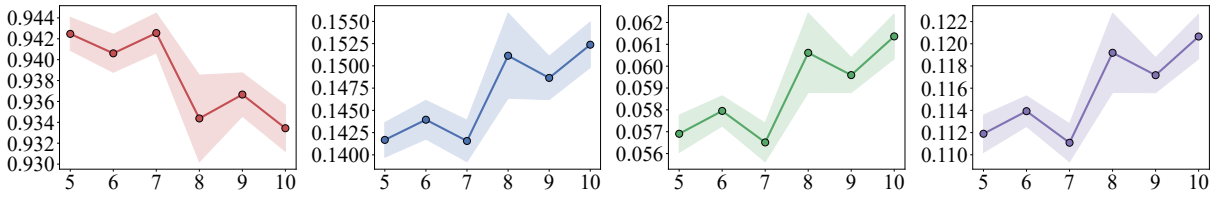
现象（1）表明，KEMS 算法对离散步长很敏感，较小的学习率可能会导致更好的模型性能。相反，现象（2）表明，最佳批量大小高度依赖于数据集属性。如5.5.2节所述，由于原料中的杂质，二氧化碳吸收单元可能会受到原料中杂质成分波动带来的噪声影响，因此更大的批量大小可能有助于模型减轻数据集噪声的影响。然而，对于水煤气变换单元数据集，这一数据集可能不会受到噪声问题的影响，更大的批大小可能会导致过度泛化和随后的性能下降。最后，现象（3）意味着粒子数量应该大于 1 以提高模型性能，然而，过多的粒子数可能会导致过拟合，降低模型性能。上述现象表明，在 E^{2}AG 模型的训练期间采用较小的学习率、合适的批大小和粒子数量对于确保 E^{2}AG 模型在实际部署的过程中保持最佳性能至关重要。这一结论从实践层面有效解答了**问题 4**，为后续模型在实际部署过程中的超参数的优化提供了指导。

5.5.5 收敛性分析

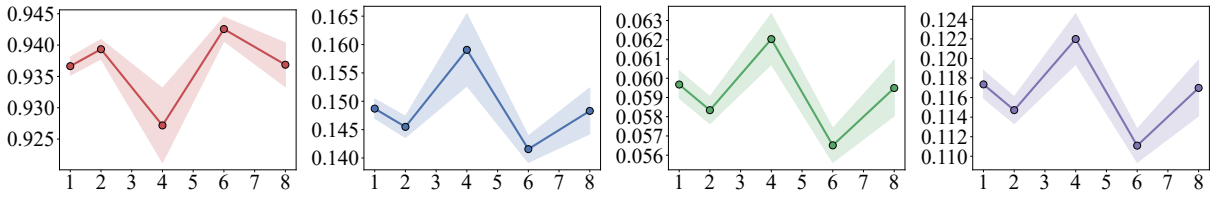
本小节将讨论**问题 5**：“KEMS 算法能否在 E^{2}AG 模型的训练过程中收敛？”为此，本小节考虑将 E^{2}AG 模型在水煤气变换单元和二氧化碳吸收单元数据集上的训练过程进行展示从而从实践上回答这一问题。需要指出的一点是由于 $\mathcal{Q}(\alpha_i)$ 在计算过程中不能显示地进行估计，因此 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(\alpha_i)||\mathcal{P}(\alpha_i|x)]$ 难以显式地直接计算。尽管如此，仍旧可以通过观察负对数似然函数 $-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u)]$ 随着迭代轮次的增加而变



(a) R^2 随 ε 变化　(b) RMSE 随 ε 变化　(c) MAPE 随 ε 变化　(d) MAE 随 ε 变化

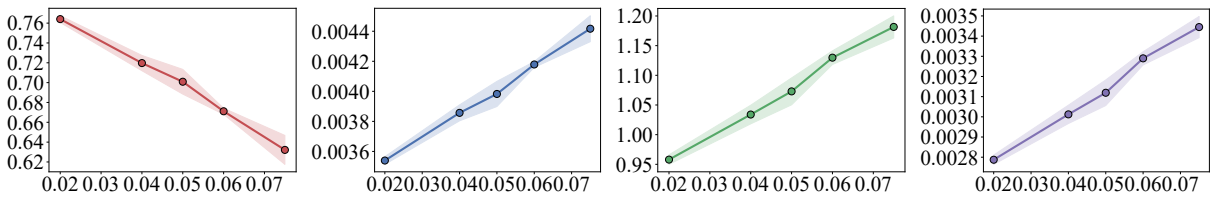


(e) R^2 随 $\log_2 \mathcal{B}$ 变化　(f) RMSE 随 $\log_2 \mathcal{B}$ 变化　(g) MAPE 随 $\log_2 \mathcal{B}$ 变化　(h) MAE 随 $\log_2 \mathcal{B}$ 变化

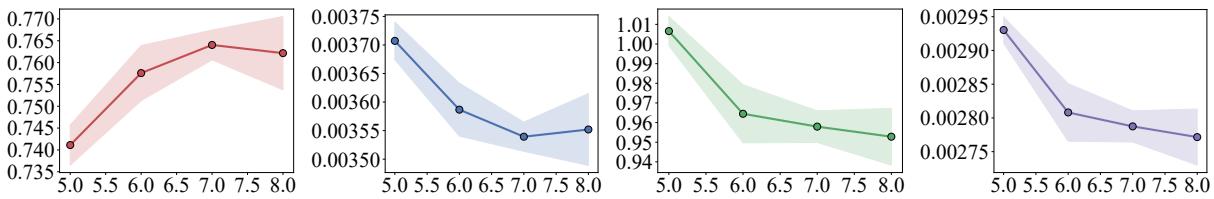


(i) R^2 随 M 变化　(j) RMSE 随 M 变化　(k) MAPE 随 M 变化　(l) MAE 随 M 变化

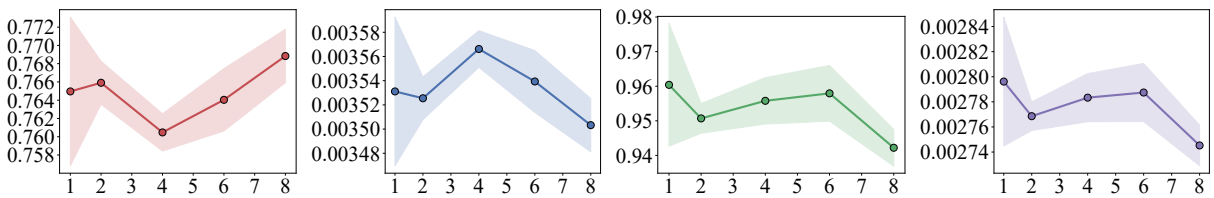
图 5.4 E²AG 在水煤气变换数据集中的敏感性分析结果



(m) R^2 随 ε 变化　(n) RMSE 随 ε 变化　(o) MAPE 随 ε 变化　(p) MAE 随 ε 变化



(q) R^2 随 $\log_2 \mathcal{B}$ 变化　(r) RMSE 随 $\log_2 \mathcal{B}$ 变化　(s) MAPE 随 $\log_2 \mathcal{B}$ 变化　(t) MAE 随 $\log_2 \mathcal{B}$ 变化



(u) R^2 随 M 变化　(v) RMSE 随 M 变化　(w) MAPE 随 M 变化　(x) MAE 随 M 变化

图 5.5 E²AG 在二氧化碳吸收数据集中的敏感性分析结果

化的结果进而从实践上验证 KEMS 算法的收敛性。为此，图5.6展示了模型在水煤气变换单元和二氧化碳吸收单元数据集上对数似然函数 $-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u)]$ 随着迭代轮次的增加而变化的结果，其中的阴影部分代表了 ±0.5 倍标准差。如图5.6所示，E^{2}AG 模型的负对数似然函数 $-\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\log p_\theta(y|\mathbb{E}_{\mathcal{Q}(\alpha_i)}[\mathcal{G}(\alpha_i,u)],u)]$ 随着训练轮数的增加而稳步下降。值得注意的是，负对数似然函数的平均曲线（红色和绿色点划线）在前 40 个迭代轮次内接近平坦，且标准差在迭代后期逐渐变小（红色和绿色阴影部分）。上述现象表明 KEMS 算法在对 E^{2}AG 模型模型的训练过程中呈现较快且较为稳定的收敛特性，从实践上回答了**问题 5**，并进一步从实践上验证了定理5.5。

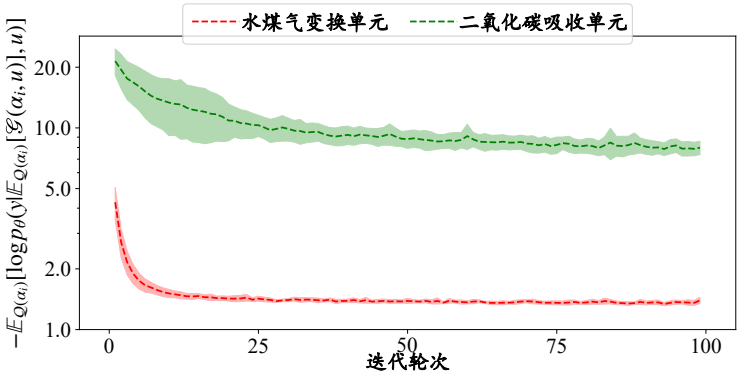


图 5.6 负对数似然函数 $-\mathbb{E}_{\mathcal{Q}(z)}[\log p_\theta(x|z)]$ 的演化过程图

5.6　本章小结

在第4章的基础上，本章基于泛函导数的方法，以图神经网络的归一化邻接矩阵的推断过程为研究案例，对限制域支撑的概率隐变量模型的推断过程进行了系统性重构。具体而言，本章分析了熵泛函正则化后的模型损失函数的泛函导数，导出了梯度优化方向，利用 RKHS 推导出了该导数的解析表达式，并从理论上证明了该解析表达式所诱导的迭代过程能有效降低模型的 KL 散度，进而实现对隐变量分布的推断。在此基础上，为了维持迭代过程样本始终保持在限制域上确保变分分布逼近过程对限制域的良定性，本章引入了镜像下降策略对迭代过程进行修改，并证明了修改后的迭代过程的收敛性。在总结上述理论推导的基础上，本章提出了 KEMS 算法，并描述了基于图神经网络设计的软测量模型 E^{2}AG 模型的相应结构。本章在最后部分通过后验分布逼近实验、软测量建模实验、消融实验、灵敏度分析实验和收敛性分析实验这五个方面的实验验证了所提出的 KEMS 算法和 E^{2}AG 模型的有效性与优越性。

6 基于有限时域最优控制的动态隐变量模型构建方法

摘要： 针对非线性动态概率隐变量模型的隐空间推断及其在深度学习后端下的实现问题，引入最优控制方法和随机微分方程框架对非线性动态概率隐变量模型的参数优化过程进行系统性分析。首先，基于随机微分方程的理论视角，将非线性动态概率隐变量模型的损失函数重新构造为最优控制问题，并通过分析该最优控制问题的解的结构，提出了非线性动态概率隐变量模型在摊销变分推断框架下的隐空间推断原理。在此基础上，为满足软测量建模对精确性的高要求，进一步引入随机微分方程的矩展开策略，重新构建了模型在深度学习后端下的实现框架。整合上述研究成果，针对动态场景下的软测量建模任务，提出了“最优控制-非线性动态概率隐变量模型”，并严格证明了其在基于梯度下降的深度学习后端训练过程中的收敛性。最后，通过软测量建模精度、模型敏感性分析等多个层面的实验验证，充分证实了所提出方法的有效性与优越性。

关键词： 随机微分方程 最优控制 矩展开 软测量建模

6.1 引言

第4章和第5章构建了实数集支撑和限制域支撑的非线性概率隐变量模型的建模方法。然而，在实际工业场景中，在非线性特性之外，工业过程通常表现出显著的动态特性，这使得传统的稳态建模方法难以满足实际需求。因此，深入研究非线性动态概率隐变量模型（nonlinear dynamical probabilistic latent variable model，NDPLVM）的构建问题，为复杂工业环境提供更为精确的建模解决方案具有非常重要的研究意义。与此同时，在深度学习时代，如文献[65]所总结的那样，NDPLVM 的设计通常基于线性动态概率隐变量模型（linear dynamical probabilistic latent variable model）的结构，通过在线性动态概率隐变量模型中引入非线性神经网络结构，如门控循环单元，以规避线性模型推导过程中的马尔可夫假设（Markov assumption），从而增强模型对动态特性数据的提取能力，实现 NDPLVM 的构建。在此基础上，这种 NDPLVM 可以视为一种特殊的动态变分自编码器（dynamical VAE）模型，并以摊销变分推断为基础，构建模型的训练流程，以实现工业过程的软测量建模。

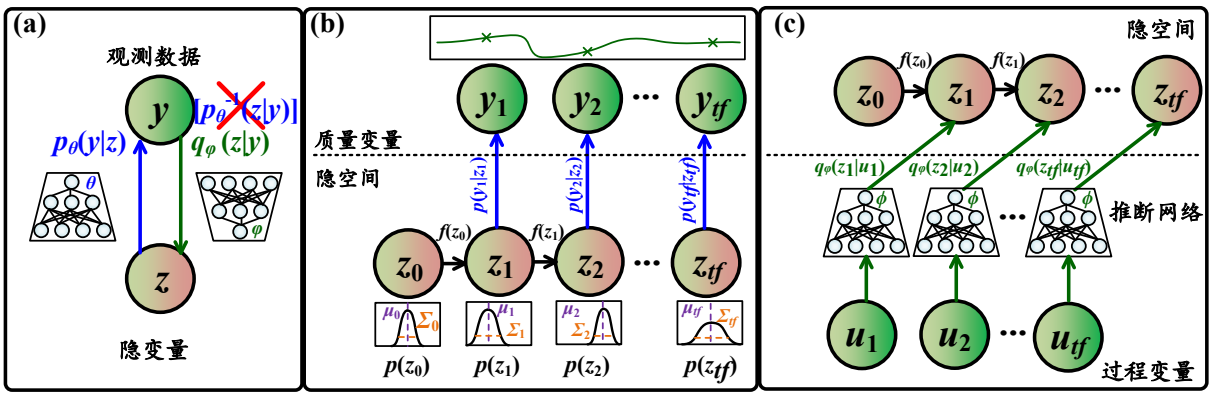


图 6.1 (a) 摊销变分推断示意图；(b) NDPLVM 的解码器过程示意图；(c) NDPLVM 的编码器过程示意图

尽管 NDPLVM 在软测量建模领域的非线性和动态特征提取方面已取得显著进展，但仍然存在两个关键问题未得到充分解决。为更清楚地阐述这一问题，首先在图6.1给出相关示意图，并定义 NDPLVM 中的几个关键组成部分：隐变量（$z \in \mathbb{R}^{D_{LV}}$）、过程变量（$u \in \mathbb{R}^{D_{PV}}$）、质量变量（$y \in \mathbb{R}^{D_{QV}}$）、以 φ 为参数的推断网络 q_φ、以 θ 为参数的生成网络 p_θ 以及转移函数 f。正如前文所提到的，NDPLVM 训练算法采用了摊销变分推断技术，该方法使用两个神经网络进行模型的训练：推断网络 q_φ 从过程变量 u 中推断出隐变量 z，而生成网络 p_θ 则从 z 解码生成质量变量 y。该方法旨在最小化两个主要成分：KL 散度的正则化项（表示推断和先验隐空间之间的差异）和似然项（量化了原始数据与生成数据之间的差异）。在此基础上，NDPLVM 在模型推理结构的设计和训练过程中面临着以下两个问题：

（1） **隐变量分布的不准确推断：**根据2.1.2节的内容和图6.1(a)，按照贝叶斯定理，推断网络 $q_\varphi(z|u)$ 的输入应设计为 $p(y|z)/p(u,y|z)$ 的“反函数”，而在软测量建模任务的背景下，其中生成网络需要生成质量变量数据 y，如图6.1(b) 所示，因此推断网络 $q_\varphi(z|u)$ 的设计应该考虑为 $q_\varphi(z|y)$，但是，如图6.1(c) 所示，当前大部分 NDPLVM 工作在设计推断网络时都不会考虑将 y 作为推断网络的输入，这导致了在推断隐空间 z 时候有很大可能会产生不准确的隐空间推断。

（2） **深度学习后端的实现：**需要指出的是，在 NDPLVM 的损失函数推导过程构建的是从概率密度函数到变量的泛函映射，但是以 PyTorch[139]和 JAX[140]为代表的神经网络的后端进行处理的并不是概率密度函数而是从概率密度函数中采出的样本，这一差异性导致如何在基于深度学习后端实现模型产生较大困难。为了更为形象

地表述这一问题，如图6.1(b) 所示，考虑转移函数 f 将 t 时刻的隐变量 z_t 映射为 $t+1$ 时刻的隐变量，在此基础上，$t+1$ 时刻的概率密度函数 $p_\theta(z_{t+1})$ 应该满足如下关系：$p_\theta(z_{t+1})|\det \nabla_{z_t} f(z_t)| = p_\theta(z_t)$。在这个过程中需要显式地计算转移函数的 f 的雅可比矩阵，最终会给模型在深度学习后端的实现带来较大困难。

因此，重新设计 NDPLVM 中推断网络的输入变量，并重构其在深度学习后端的实现方式，以改善 NDPLVM 在软测量建模任务中的性能，具有重要的研究意义。为此，本章引入最优控制和随机微分方程理论，从泛函优化的角度尝试解决这两个关键问题。具体而言，本章首先利用随机微分方程理论将 NDPLVM 的学习问题重构为优化命题，并引入 ADMM 对该优化命题进行分析和求解。在此基础上，推断网络可视为交替方向乘子法求解过程中的一个最优控制子问题的最优控制律模拟器。通过对该最优控制子问题解的深入分析，可以导出推断网络的输入变量，从而有效解决第一个问题。随后，进一步对 NDPLVM 中的矩表达式（moment expression）进行了严格分析，并提出了“最优控制-非线性动态概率隐变量模型”结构及其对应的训练和推理算法。同时，对所提出训练算法的收敛性进行了验证。最后，本章通过对软测量建模精度、模型敏感性分析等多个层面的实验，充分证实了所提方法的有效性与优越性。

6.2 相关工作回顾与科学问题分析

近年来，NDPLVM 在工业过程软测量建模中的研究和应用备受关注[14]。然而，传统动态概率隐变量模型受限于马尔可夫假设和线性假设，在提取复杂工业过程数据的时间和空间层次特征方面存在明显不足。因此，通过引入深度学习结构，打破传统概率隐变量模型的线性性和马尔可夫性，构建非线性概率隐变量模型已成为学术界的研究热点。为突破传统模型的局限性，学术界开始探索将深度学习结构引入概率隐变量模型，以打破其线性性和马尔可夫性。Shen 和 Ge[91]率先在转移函数 f 及输入输出特征提取模块中引入非线性函数，有效解决了模型的线性假设问题，并在后续研究[92]中通过互信息（mutual information）加权不同时刻的质量变量信息，在脱丁烷精馏塔数据集中验证了模型的有效性。然而，这些研究主要聚焦于非线性特征表达，尚未解决传统概率隐变量模型中的马尔可夫假设问题。针对这一问题，Lu 等人[96]引入循环神经网络结构和判别性损失函数，提出了概率判别时间序列模型，并在加氢裂化过程中验证了其有效性。

随后，Jiang 等人[97]将 GRU 作为记忆模块引入动态概率隐变量模型，结合化工过程变化缓慢的特性，提出了深度贝叶斯概率慢特征分析模型，并在后续研究中[98]引入有限混合模型架构以应对多工况操作条件带来的数据分布多峰特性。

尽管上述研究证实了引入不同深度学习结构可显著提升动态概率隐变量模型的性能，但在推断网络设计的合理性方面仍然存在一定的研究空间。具体而言，推断网络 q_φ 在摊销变分推断中承担着逆转生成网络 p_θ 的作用。若生成网络 p_θ 预测了质量变量 y，则推断网络 q_φ 应将质量变量 y 作为输入。然而，现有研究中普遍直接将过程变量 u 用作推断网络的输入，这种设计可能与摊销变分推断的设计理念存在一定偏离，从而在一定程度上影响了隐变量推断的准确性，并进一步对模型的预测精度产生了潜在影响。

在另一方面，从模型实现层面来看，现有研究通常采用“重参数化”技巧（reparameterization trick）处理概率密度函数在深度学习后端的实现问题。具体而言，对于符合高斯分布的概率密度函数的后验分布，推断网络 $q_\varphi(z)$，通常通过参数化高斯分布的均值 μ 和方差 σ^2，以 $z=\mu+\sigma\times\epsilon,\epsilon\sim\mathcal{N}(0,\mathcal{I})$ 的形式进行隐变量的推断。然而，重参数化后的样本 z 在计算概率密度转移函数 f 时面临显著的计算挑战。当重参数化后的 t 时刻隐变量 z_t 经神经网络参数化的转移函数 f 映射为 $t+1$ 时刻隐变量 z_{t+1} 时，根据贝叶斯滤波（Bayesian filtering）的原理[204-205]，z_{t+1} 时刻的样本概率密度函数可能不再保持高斯分布，导致模型在 z_{t+1} 时刻的变分分布和先验分布之间的 KL 散度难以计算，需要借助 PyRO[206]等概率编程语言实现。这一瓶颈严重阻碍了 NDPLVM 在常规深度学习后端的直接实现，不利于实现模型的大规模部署。

基于上述分析，本章研究旨在解决 NDPLVM 用于软测量建模中的以下关键科学问题：

（1） **隐空间推断的原理及推断网络的输入变量选择：**是否能够从原理上阐明 NDPLVM 的隐空间推断原理，并以此为基础设计推断网络的输入，以实现对隐空间的精确推断？

（2） **深度学习后端的实现：**如何在深度学习后端实现 NDPLVM？

6.3 随机微分方程与 ADMM

基于第2章的理论基础，本节将补充随机微分方程的基本理论框架以及 ADMM，为后续方法构建奠定数学基础。

6.3.1 随机微分方程

令三元组 $(\Omega, \mathscr{F}, \mathbb{P})$[204]为概率空间，其中 Ω 为样本空间，σ-代数 $\mathscr{F}$ 为事件集合（event set），$\mathbb{P}$ 为概率测度（probability measure）$\mathscr{F} \mapsto [0,1]$。令 $\mathscr{W}_t$ 为该空间上的 $\mathscr{F}_t$-适应维纳过程（Wiener process），$b(x,t)$ 与 L 为 $\mathscr{F}_t$-适应随机过程。可以定义如下的伊藤过程（Itô process）：

$$x(t) = x(0) + \int_0^t b(x,\tau)\mathrm{d}\tau + \int_0^t L\mathrm{d}\mathscr{W}_t, \tag{6-1}$$

作为如下的随机微分方程（stochastic differential equation）的解：

$$\mathrm{d}x(t) = b(x,t)\mathrm{d}t + L\mathrm{d}\mathscr{W}_t, \tag{6-2}$$

其中 $b(x,t)$ 为飘移项（drift term），L 为波动项（volatility term），在本文中固定为单位矩阵 $\mathcal{I}$，而维纳过程 $\mathrm{d}\mathscr{W}_t$ 的谱密度矩阵（spectral density matrix）记为 $\mathcal{Q}$。

根据上述概念，两组伊藤过程的密度比可以通过吉尔萨诺夫定理（Girsanov theorem）和拉东-尼科迪姆定理（Radon-Nikodym theorem）进行给出[204-205]：

定理 6.1 (伊藤过程的密度比，参见 Särkka 和 Solin 的著作[204]定理 7.4)： 引入形如式(6-2)的两组伊藤过程：

$$\begin{aligned} \mathrm{d}x &= f(x,t)\mathrm{d}t + \mathrm{d}\mathscr{W}_t, x(0) = x_0, \\ \mathrm{d}y &= g(y,t)\mathrm{d}t + \mathrm{d}\mathscr{W}_t, y(0) = x_0, \end{aligned} \tag{6-3}$$

其中 $f(x,t)$ 与 $g(y,t)$ 分别为飘移项，二者对应的路径空间 $\mathbb{P}$ 和 $\mathbb{Q}$ 的拉东-尼科迪姆导数（Radon-Nikodym derivative）为：

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x) = \exp(-\frac{1}{2}\int_0^t \|g(x,\tau) - f(x,\tau)\|_2^2 \mathrm{d}\tau + \int_0^t (g(x,\tau) - f(x,\tau))^\top \mathrm{d}\mathscr{W}_\tau). \tag{6-4}$$

6.3.2 ADMM

考虑由 $w \in \mathbb{R}^{\mathrm{D}}$ 和 $v \in \mathbb{R}^{\mathrm{D}}$ 这两组可分离变量组成的优化目标函数 $\boldsymbol{F}(w,v)$，该目标函数可以分解为：

$$\begin{aligned} \min \quad & \boldsymbol{F}(w,v) = \boldsymbol{f}(w) + \boldsymbol{g}(v) \\ \text{s.t.} \quad & \boldsymbol{A}w + \boldsymbol{B}v = c \end{aligned} . \tag{6-5}$$

根据增广拉格朗日乘子法，式(6-5) 可以重构为：

$$\mathcal{L}^u = \boldsymbol{f}(w) + \boldsymbol{g}(v) + \lambda^\top(\boldsymbol{A}w + \boldsymbol{B}v - c) + \frac{\rho}{2}\|\boldsymbol{A}w + \boldsymbol{B}v - c\|_2^2, \tag{6-6}$$

其中 $\lambda \in \mathbb{R}^{\mathrm{D}}$ 为拉格朗日乘子，ρ 为二次惩罚系数（quadratic penalty coefficient）。在此基础上，变量 w 和 v 以及乘子 λ 可以通过以下方式在每一轮迭代中分别优化：

$$\begin{cases} w_{\tau+1} & = \arg\min\limits_{w} \mathcal{L}(w, v_\tau) \\ v_{\tau+1} & = \arg\min\limits_{v} \mathcal{L}(w_{\tau+1}, v) \\ \lambda_{\tau+1} & = \lambda_\tau + \rho(\boldsymbol{A}w_{\tau+1} + \boldsymbol{B}v_{\tau+1} - c) \end{cases} . \tag{6-7}$$

而式(6-7)所示的方法即为 ADMM。

6.4 基于有限时域最优控制的动态隐变量推断框架

6.4.1 问题阐述

与前三章研究的稳态隐变量模型不同，本章专注于动态概率隐变量模型的设计，因此有必要在阐述所提出方法之前重新明确本章拟解决的问题。基于文献[207]中的定义，设软测量模型的预测范围为 $\mathcal{H}$，历史序列长度为 $\mathcal{T}$。在此框架下，本章拟解决的问题可表述为：给定质量变量 $y_{1:\mathcal{T}} \in \mathbb{R}^{\mathrm{D_{QV}} \times \mathcal{T}}$ 的历史序列和过程变量 $u_{1:\mathcal{T}+\mathcal{H}} \in \mathbb{R}^{\mathrm{D_{PV}} \times (\mathcal{T}+\mathcal{H})}$ 的序列，设计一种软测量模型，以预测未来 $\mathcal{H}$ 个时间窗的质量变量，即模型需能够生成对 $y_{\mathcal{T}+1:\mathcal{T}+\mathcal{H}} \in \mathbb{R}^{\mathrm{D_{QV}} \times \mathcal{H}}$ 的预测。此外，由于本章所处理的时间序列指标 t 是真实存在的，因此与前三章稳态模型中使用 τ 表示隐变量分布的概率密度函数演化过程不同，本章将时间序列指标表示为 t，用向量 $\vec{y}$ 和 $\vec{u}$ 表示质量变量序列和过程变量序列，并以空心体字符 $\mathbb{Q}$ 表示对应的路径空间。

6.4.2 研究动机分析

正如6.1节和6.2节所讨论的，如何有效引入泛函优化方法，以重新理解和设计 NDPLVM 的建模框架，是“选择变分分布函数的输入变量”和“模型的深度学习后端实现”这两个问题的核心步骤。然而，正如6.4.1节所指出的，本章所需研究的问题与前面三章存在显著差异。本章需要针对时间长度为 $\mathcal{T}+\mathcal{H}$ 的序列进行隐变量分布的推断。换言之，按照前面三章的策略，从“无穷时域”最优控制的角度对一个隐变量的后验分布进行逼近并不合适。因此，如何从“有限时域”的视角构建最优控制问题，并进而重构 NDPLVM 的建模框架，成为本章研究的重点。

为此，首先考虑在非线性概率隐变量建模场景下的变分推断目标[57,110]——如何利用变分分布 $\mathcal{Q}(z_{t+1})$ 逼近模型的后验分布 $\mathcal{P}(z_{t+1}|z_t,\vec{y},\vec{u})$：

$$\arg\min_{\mathcal{Q}(z_{t+1})} \ \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{y},\vec{u})] \ \text{ for } \ t=1,2,\ldots,\mathcal{T}+\mathcal{H}-1. \tag{6-8}$$

如果根据第4章的策略，通过引入无穷时域最优控制策略在降低 KL 散度的同时规避显式的 KL 散度项 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{y},\vec{u})]$ 的计算会带来推断效率的问题——因为需要推断的 t 有 $\mathcal{T}+\mathcal{H}-1$ 个，因此直接在 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{y},\vec{u})]$ 这一项上引入最优控制策略对非线性概率隐变量模型的学习进行重构，将 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{y},\vec{u})]$ 的正则转移到 $\mathcal{Q}(z_{t+1})$ 的演化过程中，并不是一个明智的选择。为此，可以考虑通过引入与 KL 散度等价的证据下界泛函，考察其引入最优控制策略进行分析的可能性。在此基础上，式(6-8)可以转化为下式：

$$\arg\min_{\mathcal{Q}(z_{t+1})} \ \mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]-\mathbb{E}_{\mathcal{Q}(z_{t+1})}[\log\mathcal{P}(\vec{y}|z_{t+1})] \ \text{ for } \ t=1,2,\ldots,\mathcal{T}+\mathcal{H}-1, \tag{6-9}$$

其中，对数似然函数项 $\mathbb{E}_{\mathcal{Q}(z_{t+1})}[\log\mathcal{P}(\vec{y}|z_{t+1})]$ 可以直接通过蒙特卡洛估计进行计算，而 KL 散度项 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 在计算上则存在较大的挑战，因为分布 $\mathcal{P}(z_{t+1}|z_t,\vec{u})$ 取决于 z_t 时刻的分布 $\mathcal{P}(z_t)$ 和转移函数 f，直接进行 $\mathcal{P}(z_{t+1})$ 的计算需要计算 $\mathcal{P}(z_t)$ 和转移函数 f 的雅可比矩阵。因此，本章研究的关键在于如何在 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 的计算过程中引入最优控制策略，将 KL 散度的正则项转移至分布 $\mathcal{Q}(z_{t+1})$ 的演化过程中，从而重构非线性概率隐变量模型的学习框架。

6.4.3　最优控制问题的构造

如何引入最优控制，在避免直接计算 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 的同时实现对隐变量分布的推断是本小节需要解决的核心问题。值得注意的是，$\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 的表达式涉及 $t+1$ 时刻的概率密度函数 $\mathcal{Q}(z_{t+1})$ 和 $\mathcal{P}(z_{t+1}|z_t)$ 之间的密度比[70]（density ratio）$\dfrac{\mathcal{Q}(z_{t+1})}{\mathcal{P}(z_{t+1}|z_t,\vec{u})}$ 的计算。而式(6-4)中引入的两条概率路径的密度比计算结果与二次型最优控制的代价泛函具有显著的相似性。因此，一个可行的方案是将 $\dfrac{\mathcal{Q}(z_{t+1})}{\mathcal{P}(z_{t+1}|z_t,\vec{u})}$ 这一时刻的密度比，转化为 $[t,t+1]$ 时间段内的概率密度函数密度比，从而将 $t+1$ “时刻”（timestamp）两个分布之间的 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 转化为 $[t,t+1]$ “时段”（time interval）两个随机过程之间的 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}_t(z)\|\mathbb{P}_t(z)]$，并进一步实现最优控制问题的构建。这一过程的基本示意图如图6.2所示。

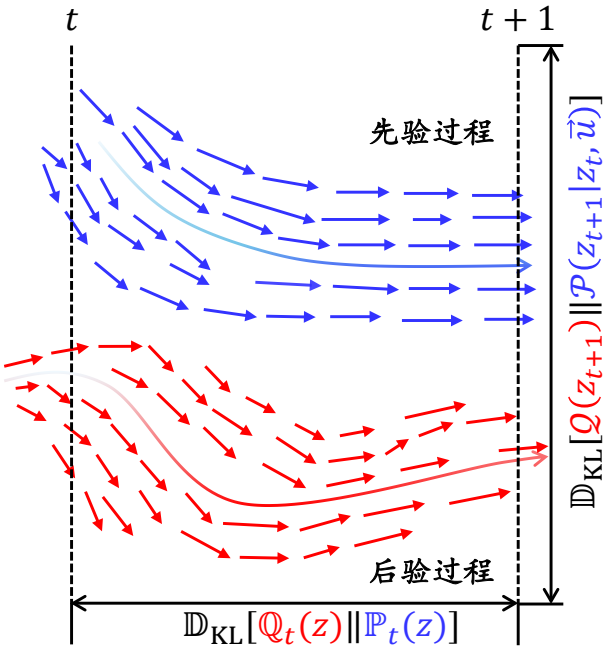


图 6.2 先验过程和后验过程的 KL 散度对比图

如图6.2所示，由于时间区间 $[t,t+1]$ 包含“无穷多”个时刻，因此直觉上可以认为，在 $t+1$ 时刻两个随机过程之间的 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 应当大于它们在整个时间区间 $[t,t+1]$ 上的 KL 散度 $\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}_t(z)\|\mathbb{P}_t(z)]$。因此，可以将 $\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_{t+1})\|\mathcal{P}(z_{t+1}|z_t,\vec{u})]$ 替换为 $\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}_t(z)\|\mathbb{P}_t(z)]$，从而得到一个上界作为合理的损失泛函。然而，上述分析主要基于直觉，尚缺乏严谨的数学证明。为了验证这一直觉的合理性，本小节首先提出如下定理，为上述分析过程提供理论支撑：

定理 6.2： 设 $\mathcal{Q}(z)$ 和 $\mathcal{P}(z)$ 是随机过程 $\mathbb{Q}(z)$ 和 $\mathbb{P}(z)$ 在某一预定义时刻 $\mathcal{T}$ 的边缘分布（即 $\mathcal{Q}_{\mathcal{T}}(z)=\mathcal{Q}(z)$ 和 $\mathcal{P}_{\mathcal{T}}(z)=\mathcal{P}(z)$）。则下列不等式成立：

$$\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z)\|\mathcal{P}(z)]\leq\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}(z)\|\mathbb{P}(z)]. \tag{6-10}$$

证明： 设 $\mathcal{G}$ 为 $\mathscr{F}$ 的子 σ-代数（即 $\mathcal{G}\subset\mathscr{F}$），$\mathbb{Q}_{\mathcal{G}}$ 和 $\mathbb{P}_{\mathcal{G}}$ 分别为 $\mathbb{Q}$ 和 $\mathbb{P}$ 在 $\mathcal{G}$ 上的限制。记 $\mathbb{E}_{\mathbb{P}}[\cdot|\mathcal{G}]$ 为概率测度 $\mathbb{P}$ 下给定 $\mathcal{G}$ 的条件期望。则对任意非负 $\mathcal{G}$-可测函数 $\boldsymbol{g}$，下列等式成立：

$$\int \mathrm{d}\mathbb{P}_{\mathcal{G}}(z)\boldsymbol{g}\mathbb{E}_{\mathbb{P}(z)}[\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\mathbb{P}(z)}|\mathcal{G}]=\int \mathrm{d}\mathbb{P}(z)\boldsymbol{g}\mathbb{E}_{\mathbb{P}(z)}[\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\mathbb{P}(z)}|\mathcal{G}]\overset{\text{(i)}}{=}\int \mathrm{d}\mathbb{P}(z)\boldsymbol{g}\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\mathbb{P}(z)}=\int \mathrm{d}\mathbb{Q}_{\mathcal{G}}(z)\boldsymbol{g}, \tag{6-11}$$

其中步骤“(i)”是基于随机过程的塔属性（tower property，参见 Billingsley 的著作[208]的定理 34.4）得到的结果。

因此在式(6-11)的基础上可以得到下列等式：

$$\frac{\mathrm{d}\mathbb{Q}_{\mathcal{G}}(z)}{\mathrm{d}\mathbb{P}_{\mathcal{G}}(z)}=\mathbb{E}_{\mathbb{P}(z)}(\frac{\mathrm{d}\mathbb{P}(z)}{\mathrm{d}\mathbb{Q}(z)}|\mathcal{G}). \tag{6-12}$$

而对于凸函数 $\Phi(z)\coloneqq z\log z$ 存在下列不等式：

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}_{\mathcal{G}}(z)\|\mathbb{P}_{\mathcal{G}}(z)]\\
=&\int \mathrm{d}\mathbb{P}_{\mathcal{G}}(z)\Phi(\frac{\mathrm{d}\mathbb{Q}_{\mathcal{G}}(z)}{\mathrm{d}\mathbb{P}_{\mathcal{G}}(z)})\\
=&\int \mathrm{d}\mathbb{P}(z)\Phi(\frac{\mathrm{d}\mathbb{Q}_{\mathcal{G}}(z)}{\mathrm{d}\mathbb{P}_{\mathcal{G}}(z)})\\
=&\int \mathrm{d}\mathbb{P}(z)\Phi[\mathbb{E}_{\mathbb{P}}(\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\mathbb{P}(z)}|\mathcal{G})]\\
\overset{\text{(ii)}}{\leq}&\int \mathrm{d}\mathbb{P}(z)\mathbb{E}_{\mathbb{P}(z)}[\Phi(\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\mathbb{P}(z)})|\mathcal{G}]\\
\overset{\text{(iii)}}{=}&\int \mathrm{d}\mathbb{P}(z)\Phi(\frac{\mathrm{d}\mathbb{Q}(z)}{\mathrm{d}\mathbb{P}(z)})\\
=&\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}(z)\|\mathbb{P}(z)],
\end{aligned} \tag{6-13}$$

其中步骤“(ii)”由琴生不等式导出，步骤“(iii)”由塔属性导出。因为 $\mathbb{Q}_{\mathcal{T}}(z)=\mathcal{Q}(z)$，并且 $\mathbb{P}_{\mathcal{T}}(z)=\mathcal{P}(z)$ 以及 $\mathcal{T}\in[0,\mathcal{T}]$，所以(6-12)所定义的不等式得证。 **证毕**

在式(6-4)和定理6.2的基础上，可以将式(6-9)所给出的证据下界重构为一个和二次型最优控制的代价泛函十分相似的目标泛函，以实现 NDPLVM 的训练框架的重构。根

据这一想法，在定理6-2的基础上，两组伊藤过程在时段 $[t,t+1]$ 的 KL 散度如下所示：

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_t(z)\|\mathbb{P}_t(z)) \\
=&\mathbb{E}_{\mathbb{Q}_t(z)}[\ln\frac{\mathcal{Q}(z_t)}{\mathcal{P}(z_t)}\frac{\mathrm{d}\mathbb{Q}_{(t,t+1]}(z)}{\mathrm{d}\mathbb{P}_{(t,t+1]}(z)}|z_t=z] \\
=&\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_t)\|\mathcal{P}(z_t)]+\mathbb{E}_{\mathbb{Q}_t(z)}[\frac{1}{2}\int_t^{t+1}\|\nu\|_2^2\mathrm{d}\tau+\int_t^{t+1}\nu^\top\mathrm{d}\mathscr{W}_\tau] \\
\overset{\text{(i)}}{=}&\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_t)\|\mathcal{P}(z_t)]+\mathbb{E}_{\mathbb{Q}_t(z)}[\frac{1}{2}\int_t^{t+1}\|\nu\|_2^2\mathrm{d}\tau] \\
\overset{\text{(ii)}}{=}&\mathbb{E}_{\mathbb{Q}_t(z)}[\frac{1}{2}\int_t^{t+1}\|\nu\|_2^2\mathrm{d}\tau],
\end{aligned} \tag{6-14}$$

其中 ν 定义为两组伊藤过程的飘移项之差：

$$\nu\coloneqq f(x,t)-g(x,t), \tag{6-15}$$

而步骤“(i)”利用了维纳过程的鞅性质，即 $\mathbb{E}_{\mathbb{Q}_t(z)}[\int_t^{t+1}\nu^\top\mathrm{d}\mathscr{W}_\tau]=0$；步骤“(ii)”则是基于两个随机过程具有相同的起始分布这一事实（$\mathcal{Q}(z_t)=\mathcal{P}(z_t)$）得到的结果。

在式(6-14)的基础上，根据极大似然原理，可以导出下述定理以确定非线性概率隐变量模型的代价泛函：

定理 6.3： 假设非线性隐变量模型的隐空间的先验过程为：

$$\mathrm{d}z=f_\theta(z,\vec{u})\mathrm{d}t+L\mathrm{d}\mathscr{W}_t, \tag{6-16}$$

后验过程为：

$$\mathrm{d}z=f_\varphi(z,\vec{u})\mathrm{d}t+L\mathrm{d}\mathscr{W}_t=f_\theta(z,\vec{u})\mathrm{d}t+L\nu\mathrm{d}t+L\mathrm{d}\mathscr{W}_t, \tag{6-17}$$

其中 θ 和 φ 分别为神经网络参数，ν 为 $f_\varphi(z,\vec{u})$ 与 $f_\theta(z,\vec{u})$ 之差。则在观测数据极大似然原理的基础上，非线性概率隐变量模型的损失泛函可以表示为：

$$\arg\min_{\theta,\nu}\ \sum_{t=1}^{\mathcal{T}+\mathcal{H}}\{\mathbb{E}_{\mathbb{Q}(z)}[-\log p_\theta(y_t|z_t,u_{1:t})+\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau]\} \tag{6-18}$$

$$\text{s.t.}\ \ \mathrm{d}z=f_\varphi(z,u)\mathrm{d}t+L\mathrm{d}\mathscr{W}_t=f_\theta(z,u)\mathrm{d}t+L\nu\mathrm{d}t+L\mathrm{d}\mathscr{W}_t \tag{6-19}$$

证明： 根据极大似然原理，非线性概率隐变量模型可以通过下式给出：

$$\arg\max_\theta\ \log p_\theta(\vec{y}|\vec{u})=\arg\max_\theta\ \log\int p_\theta(\vec{y},\vec{z}|\vec{u})\mathrm{d}\vec{z}. \tag{6-20}$$

则式(6-20)右端可以化为：

$$\log \int p_\theta(\vec{y},\vec{z}|\vec{u})\mathrm{d}\vec{z} = \log \int \mathcal{Q}(\vec{z})\frac{p_\theta(\vec{y},\vec{z}|\vec{u})}{\mathcal{Q}(\vec{z})}\mathrm{d}\vec{z} \overset{\text{(i)}}{\geq} \int \mathcal{Q}(\vec{z})\log\frac{p_\theta(\vec{y},\vec{z}|\vec{u})}{\mathcal{Q}(\vec{z})}\mathrm{d}\vec{z}, \tag{6-21}$$

其中步骤“(i)”是基于琴生不等式导出的结果。在此基础上，根据 Girin 等人的著作[65]，似然函数 $p_\theta(\vec{y},\vec{z}|\vec{u})$ 可以因子化为下式：

$$\begin{aligned}
&p_\theta(\vec{y},\vec{z}|\vec{u})\\
=&\prod_{t=1}^{\mathcal{T}+\mathcal{H}} p(y_t,z_t|y_{1:t-1},z_{1:t-1},u_{1:t})\\
=&\prod_{t=1}^{\mathcal{T}+\mathcal{H}} p_\theta(y_t|y_{1:t-1},z_{1:t},u_{1:t})p_\theta(z_t|y_{1:t-1},z_{1:t-1},u_{1:t})\\
=&\prod_{t=1}^{\mathcal{T}+\mathcal{H}} p_\theta(y_t|z_t)p_\theta(z_t|z_{t-1},u_{1:t}).
\end{aligned} \tag{6-22}$$

与之相对应地，变分分布 $\mathcal{Q}(\vec{z})$ 可以拆解为下式：

$$\mathcal{Q}(\vec{z}) = \prod_{t=1}^{\mathcal{T}+\mathcal{H}} \mathcal{Q}(z_t|z_{t-1}). \tag{6-23}$$

则式(6-21)可以重构为：

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathrm{D_{LV}}}} \mathcal{Q}(\vec{z})\log\frac{p_\theta(\vec{y},\vec{z}|\vec{u})}{\mathcal{Q}(\vec{z})}\mathrm{d}\vec{z}\\
=&\sum_{t=1}^{\mathcal{T}+\mathcal{H}} \{\underbrace{\mathbb{E}_{\mathcal{Q}(z_t|z_{t-1})}[\log p_\theta(y_t|z_t)]}_{=\mathbb{E}_{\mathbb{Q}_t(z)}[\log p_\theta(y_t|z_t)]} - \underbrace{\mathbb{D}_{\mathrm{KL}}[\mathcal{Q}(z_t|z_{t-1})\|p_\theta(z_t|z_{t-1},u_{1:t})]}_{\leq\mathbb{D}_{\mathrm{KL}}[\mathbb{Q}_t(z)\|\mathbb{P}_t(z)]}\}\\
\overset{\text{(ii)}}{\geq}&\sum_{t=1}^{\mathcal{T}+\mathcal{H}} \{\mathbb{E}_{\mathbb{Q}_t(z)}[\log p_\theta(y_t|z_t)] - \mathbb{D}_{\mathrm{KL}}[\mathbb{Q}_t(z)\|\mathbb{P}_t(z)]\},
\end{aligned} \tag{6-24}$$

其中步骤“(ii)”是基于定理6.2导出的结果。在此基础上，考虑先验和后验过程分别为下列方程组：

$$\begin{cases}
\mathrm{d}z &= f_\theta(z,\vec{u})\mathrm{d}t + L\mathrm{d}\mathscr{W}_t\\
\mathrm{d}z &= f_\varphi(z,\vec{u})\mathrm{d}t + L\mathrm{d}\mathscr{W}_t
\end{cases}, \tag{6-25}$$

并定义随机控制策略（stochastic control policy）$\nu \coloneqq f_\varphi(z,\vec{u}) - f_\theta(z,\vec{u})$，则根据式(6-14)和式(6-15)可以得到：

$$\arg\max_{\theta,\nu} \sum_{t=1}^{\mathcal{T}+\mathcal{H}} \{\mathbb{E}_{\mathbb{Q}_t(z)}[\log p_\theta(y_t|z_t,u_{1:t}) - \int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau]\}. \tag{6-26}$$

在此基础上，根据式(6-25)和式(6-26)，式(6-18)和式(6-19)得证。　　**证毕**

虽然定理 6.3在凸优化框架下重新表述了 NDPLVM 的参数学习优化问题，但由于不同时间区间的初始点不确定以及计算过程中需要进行“回溯”（backtracking）操作，式(6-18)所描述的学习目标在应用于软测量建模时仍面临较大挑战。具体而言，在软测量建模任务中，这种“回溯”操作是不切实际的，因为这些任务本质上具有因果性，即无法用未来的质量变量对过去的隐变量进行推断[209]。为了在软测量建模的背景下进一步说明这一问题，考虑如下的情况：y_t 是通过隐变量 z_t 进行预测的。然而，根据式(6-18)，z_t 依赖于未来的观测值 $y_{t+1:\mathcal{T}+\mathcal{H}}$，这在实际应用中是不可能实现的。这种时间上的不一致性对模型的训练带来了显著的挑战。尽管如此，Bertsekas 在其著作[210]中提出的一种“单步前瞻最小化”（one-step look-ahead minimization）方法可以用来解决这一问题，基于这一“单步前瞻最小化”策略，本小节进一步提出以下定理，以推导式(6-18)的一个上界从而简化 NDPLVM 的训练流程：

定理 6.4： 式(6-18)所定义的 NDPLVM 的目标泛函有如下的上界：

$$\begin{aligned}&\arg\min_{\theta,\nu}\ \sum_{t=1}^{\mathcal{T}+\mathcal{H}}\{-\mathbb{E}_{\mathbb{Q}_t(z)}[\log p_\theta(y_t|z_t,u_{1:t})+\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau]\}\\ \le&\sum_{t=1}^{\mathcal{T}+\mathcal{H}}\{\arg\min_{\theta,\nu}\ \mathbb{E}_{\mathbb{Q}_t(z)}[-\log p_\theta(y_t|z_t,u_{1:t})+\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau]\}\end{aligned}\tag{6-27}$$

证明：考虑一个更一般的情况：设 $\{\nu_1,\nu_2,\ldots,\nu_{\mathcal{T}+\mathcal{H}}\}$ 是一组控制策略，而 $\{\mathscr{L}_1,\mathscr{L}_2,\ldots,\mathscr{L}_{\mathcal{T}+\mathcal{H}}\}$ 是与之对应的函数。令 $\mathscr{Y}$ 表示如下寻找一组控制策略 $\nu_t|_{t=1}^{\mathcal{T}+\mathcal{H}}$ 的组合，使得对应的函数 $\mathscr{L}_t(\nu_t)$ 的总和达到最小的优化问题的最优解：

$$\mathscr{Y}\coloneqq\min\{\mathscr{L}_1(\nu_1)+\mathscr{L}_2(\nu_2)+\ldots+\mathscr{L}_{\mathcal{T}+\mathcal{H}}(\nu_{\mathcal{T}+\mathcal{H}})\}.\tag{6-28}$$

与此同时，再令 $\mathscr{S}$ 表示如下表达式的值：

$$\mathscr{S}\coloneqq\sum_{t=1}^{\mathcal{T}+\mathcal{H}}\min\{\mathscr{L}_t(\nu_t)\}.\tag{6-29}$$

上式的物理意义是对每个 $\mathscr{L}_t(\nu_t)$ 分别独立地求最小值，然后将这些最小值累加起来。

由于 $\mathscr{Y}$ 是在所有可能的 ν_t 组合下得到的最小总和，因此，对于任意给定的 t，$\mathscr{L}_t(\nu_t)$ 在 $\mathscr{Y}$ 中的贡献不可能超过其自身的最小值 $\min\{\mathscr{L}_t(\nu_t)\}$。这意味着，$\mathscr{Y}$ 不可能大于所有 $\mathscr{L}_t(\nu_t)$ 各自最小值的总和。因此，有如下结论：

$$\mathscr{Y}\le\sum_{t=1}^{\mathcal{T}+\mathcal{H}}\min\{\mathscr{L}_t(\nu_t)\}=\mathscr{S},$$

则式(6-27)得证。 **证毕**

至此，本小节完成了将 NDPLVM 的学习目标的重构为最优控制问题的过程。此外，与4.3.3节中的所讨论的内容类似，式(6-18)和式(6-19)所给出的目标函数有如下注释：

注释 6： 从演化过程的视角来看，式(6-18) 和式(6-19)中定义的最优控制问题，将一个从给定的归一化概率密度函数族 $\mathbb{F}$ 中推断 $\mathcal{Q}(z_t|z_{t-1})$ 的问题，转化为一个有限时域路径空间（finite-horizon path space）$\mathscr{C}\left([0,1],\mathbb{R}^{\mathrm{D_{LV}}}\right)$ 中求解最优控制策略 $\nu^*(z)$ 的问题。这一松弛通过扩展假设空间，显著提升了变分推断过程的灵活性。

6.4.4 基于最优控制的推断网络设计

通过对比式(6-5)与式(6-27)，可以发现两者在参数对应关系上具有以下特征：θ 对应于 ADMM 框架中的 w，ν 对应于 v。此外，对数似然项 $\log p_\theta(y_t|z_t,u_{1:t})$ 与控制策略项 $\int_{t-1}^{t}\|\nu\|_2^2\mathrm{d}\tau$ 在数学上具有可分离性。这一特性为问题求解提供了重要启发：基于式(6-7)所给出的 ADMM 的分解思想，可将 NDPLVM 的参数学习过程拆解为两个交替更新步骤，即最优控制策略 ν 的优化与生成网络参数 θ 的梯度更新，从而形成高效的联合学习框架。进一步分析表明，该优化过程与算法2.1中描述的 EM 算法在形式上具有内在一致性。具体而言，在本模型中，EM 算法的 E 步骤对应于确定最优控制策略 ν^* 的过程，而最优隐变量分布 $\mathcal{Q}^*(z)$ 的求解结构直接决定了摊销变分推断框架下推断网络输入的特征表达。因此，确定 ν^* 的解的结构是 NDPLVM 能够精确推断隐空间的推断网络的关键。基于上述理论分析，本节的研究目标是通过求解式(6-27)与式(6-19)所定义的有限时域最优控制问题，显式推导最优控制策略 ν^* 的数学表达式，并据此构选择断网络的输入变量，以实现精确的隐空间推断，从而完善 NDPLVM 在摊销变分推断框架下的学习流程。

注意到，根据定理6.4，优化式(6-18)中定义的“全局”代价泛函可以转化为优化每个时段 $[t-1,t],t\in\{1,\ldots,\mathcal{T}+\mathcal{H}\}$ 上的“局部”代价泛函问题。因此，接下来的内容将讨论关注于时段 $[t-1,t],t\in\{1,\ldots,\mathcal{T}+\mathcal{H}\}$ 上的最优控制问题的求解。在给出形式化的定理前，首先对式(6-27)右侧的局部代价泛函进行简化。根据 Jiang 等人的论文[97]将条件分布 $p(y|z_t,u_{1:t})$ 参数化为 $\mathcal{N}(\mu_t^y,\mathcal{I})$，并通过输出网络 $g_\theta(\cdot)$ 构造 μ_t^y 后，式(6-27)右

侧的局部代价泛函可以简化为下式：

$$\mathbb{E}_{\mathbb{Q}(z)}[-\log p_\theta(y_t|z_t)+\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau]=\mathbb{E}_{\mathbb{Q}(z)}[\frac{1}{2}(y_t-\mu_t^y)^\top(y_t-\mu_t^y)+\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau], \tag{6-30}$$

需要指出的是，这一过程可以理解为将 $p_\theta(y|z_t,u_{1:t})$ 理解为狄拉克测度（即 $\delta(y|z_t,u_{1:t})$）并通过正态分布 $\mathcal{N}(y-\mu_t^y,\mathcal{I})$ 对其进行近似。根据 Chen 的论文[211]，这种假设在带宽设置为 1 的核密度估计中较为常见。进一步分析表明，由于软测量建模通常关注预测精度问题即预测均值的估计问题，通过展开 $\mathbb{E}_{\mathbb{Q}(z)}[\frac{1}{2}(y_t-\mu_t^y)^\top(y_t-\mu_t^y)]$ 可以得到以下近似关系：

$$\mu_t^y=\mathbb{E}[g_\theta(z_t)]\approx g_\theta(\mu_t^z). \tag{6-31}$$

在忽略与方差相关的布朗运动 $\mathrm{d}\mathscr{W}_t$ 的情况下，可以构建如下确定性最优控制问题：

$$\begin{aligned}\arg\min_{\nu}\quad & (y_t-\mu_t^y)^\top(y_t-\mu_t^y)+\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau\\ \text{s.t.}\quad & \mathrm{d}z=f_\theta(z,u)\mathrm{d}t+L\nu\mathrm{d}t,\end{aligned} \tag{6-32}$$

其中，ν 表示控制策略的设计变量。在此基础上，最控制策略 ν^* 的解的结构可以通过以下定理进行表述：

定理 6.5： 式(6-32)的最优控制策略 ν^* 可参数化为下式：

$$\nu^*=q(y_t,z_t)=q(y_t,\mu_t,\Sigma_t), \tag{6-33}$$

其中 μ_t 和 Σ_t 为 z_t 的均值和协方差。

证明： 根据庞特里亚金极大值原理[107]，引入协态 $\lambda\in\mathbb{R}^{\mathrm{D_{LV}}}$ 构造下列哈密顿函数：

$$\boldsymbol{H}=\frac{1}{2}\|\nu\|_2^2+\lambda^\top[f_\theta(z,u)+L\nu]. \tag{6-34}$$

根据最优控制充分条件，可以得到如下结果：

$$\frac{\partial\boldsymbol{H}}{\partial\nu}=\nu+L\lambda=0\Rightarrow\nu=-L\lambda. \tag{6-35}$$

验证二阶条件，可以发现哈密顿函数对 ν 的二阶偏微分满足下列条件：

$$\frac{\partial\boldsymbol{H}}{\partial\nu}=\mathcal{I}\succ 0. \tag{6-36}$$

因此，根据勒让德条件[136]，可知式(6-35)为式(6-32)的最优控制策略。

在此基础上，分析最优控制的状态变量和协态变量满足下列微分方程组：

$$\begin{cases} \dfrac{\mathrm{d}\mu_t^z}{\mathrm{d}t} = \mathbb{E}_{\mathbb{Q}(z)}(\dfrac{\partial \mathcal{H}}{\partial \lambda}) = f_\theta(\mu_t^z, u) + L\nu \\ \dfrac{\mathrm{d}\lambda}{\mathrm{d}t} = -\dfrac{\partial \mathcal{H}}{\partial z} = -\dfrac{\partial f_\theta(z,u)}{\partial z} \end{cases}, \tag{6-37}$$

其对应的边值条件为：

$$\begin{cases} \mu_{t-1}^z = \mathbb{E}(z_{t-1}) \\ \lambda_t = 2(\mu_t^y - y_t)^\top \dfrac{\partial g_\theta(z_t)}{\partial z_t}|_{z_t=\mu_t^z} \end{cases}. \tag{6-38}$$

通过观察式(6-35)、式(6-37)和式(6-38)可以得到最优控制 ν^* 是一个输入为 y_t 和 z_t 的函数，而 z_t 又与均值 μ_t 和协方差 Σ_t 有关，因此式(6-33)得证。 **证毕**

在此基础上，采用参数为 φ 的神经网络 q_φ 对最优解 ν^* 进行参数化，可以确定推断网络的输入变量：

$$\nu^* = q_\varphi(y_t, \mu_t, \Sigma_t). \tag{6-39}$$

至此，本小节结合最优控制理论，将模型目标函数进行了系统性的重构，揭示了 NDPLVM 中推断网络的本质——即最优控制解的近似模拟器。该理论分析进一步明确了推断网络输入变量的设计原则，并指出其应严格依照最优控制策略的函数结构，以确保建模过程中达到较高的精度和理论一致性。

需要指出的是，经过最优控制策略 ν 修改后的隐状态从测度 $\mathbb{P}$ 转换为了测度 $\mathbb{Q}$，因此需要给出测度 $\mathbb{Q}$ 下的 z_t 分布以进行模型的递推计算。为此，根据吉尔萨诺夫定理，修改前后的 z_t 之间的关系可以通过下式给出：

$$\mathrm{d}z^{\mathbb{Q}} = \exp(\int_{t-1}^{t} -\frac{1}{2}\|\nu\|_2^2 \mathrm{d}\tau)\mathrm{d}z^{\mathbb{P}}. \tag{6-40}$$

由于在起始时间点 $t-1(t \in \{0, \ldots, \mathcal{T}+\mathcal{H}-1\})$，$z$ 在测度 $\mathbb{Q}$ 和 $\mathbb{P}$ 下的概率密度相同，因此可以通过测度 $\mathbb{P}$ 下的概率密度推导出测度 $\mathbb{Q}$ 下的概率密度。在此基础上，设 z 在时间 t 的概率密度为：

$$z_t \sim \mathcal{N}(\mu_t^{z,\mathbb{P}}, \Sigma_t^{z,\mathbb{P}}), \tag{6-41}$$

其中，上标 $z,\mathbb{P}$ 表示 z_t 定义在测度 $\mathbb{P}$ 下，μ 和 Σ 分别表示正态分布的均值和协方差矩阵。假设拉东-尼科迪姆导数 ν 的积分的解为：

$$\nu \sim \mathcal{N}(\mu_t^\nu, \Sigma_t^\nu), \tag{6-42}$$

则在时间 t，根据舒尔补[212]（Schur complement）操作，测度 $\mathbb{Q}$ 下 z_t 的概率密度可以通过下式导出：

$$z_t^{\mathbb{Q}} \sim \mathcal{N}((\Sigma_t^{\nu}+\Sigma_t^{z,\mathbb{P}})^{-1}(\Sigma_t^{\nu}\mu_t^{\nu}+\Sigma_t^{z,\mathbb{P}}\mu_t^{z,\mathbb{P}}),\Sigma_t^{\nu}+\Sigma_t^{z,\mathbb{P}})^{-1}(\Sigma_t^{\nu}\Sigma_t^{z,\mathbb{P}})). \tag{6-43}$$

综上所述，推断网络在 ADMM 框架下本质上等价于最优控制子问题的近似求解器，针对 NDPLVM 模型，推断网络的结构设计应遵循以下原则：

1. **优化命题的构造与分解：**首先，根据 NDPLVM 的证据下界构造整体优化目标，并借助 ADMM 方法将其分解为生成网络参数 θ 的优化子问题与隐变量推断相关的泛函优化子问题（即本章式(6-32)的最优控制问题所述）；

2. **求解有关推断网络的子优化命题：**固定生成网络参数的前提下，集中求解与推断相关的泛函优化子问题。通过应用泛函优化理论（如本节的庞特里亚金极大值原理）推导出最优函数的表达式，明确其依赖的输入变量（如定理 6.5 所示）；

3. **基于最优函数的网络结构设计：**引入参数为 φ 的神经网络对最优函数进行逼近，将神经网络的输入设置为最优函数的输入，即完成推断网络的设计。

6.4.5 基于矩展开策略的模型实现方法

在6.4.4节以最优控制方法给出的推断网络结构设计原理的基础上，本小节关注于解决模型在以 PyTorch[139]和 JAX[140]为代表的深度学习后端实现方法；为此，本小节首先给出下列定理以描述隐变量的一阶矩和二阶矩在不同时刻之间的演化：

定理 6.6：设由式(6-16)定义的先验伊藤过程驱动隐变量 z_t 与 z_{t+1} 在 t 至 $t+1$ 的演化，该演化过程的均值和协方差的变化可通过式(6-44)与式(6-45)定义的 ODE 描述：

$$\frac{\mathrm{d}\mu}{\mathrm{d}t}=f(\mu,t), \tag{6-44}$$

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}t}=\Sigma[\frac{\partial f(z,t)}{\partial z}|_{z=\mu}]^{\top}+\Sigma^{\top}[\frac{\partial f(z,t)}{\partial z}|_{z=\mu}]+L\mathcal{Q}L^{\top}. \tag{6-45}$$

证明：首先给出式(6-46)和式(6-47)给出的均值和协方差的定义式：

$$\mu_t=\mathbb{E}(z_t), \tag{6-46}$$

$$\Sigma_t = \mathbb{E}[(z_t - \mu_t)(z_t - \mu_t)^\top]. \tag{6-47}$$

在此基础上，均值和协方差对应的 ODE 可以导出为式(6-48)和式(6-49)：

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = \mathbb{E}[f(z,t)], \tag{6-48}$$

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}t} = \mathbb{E}[f(z,t)(z-\mu)^\top + (z-\mu)f(z,t)^\top + L(z,t)\mathcal{Q}L^\top(z,t)]. \tag{6-49}$$

基于式(6-16)所给出的先验伊藤过程，先验过程演化的均值和协方差可以进一步化为如下所示的式(6-50)和式(6-51)：

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = \int_{\mathbb{R}^{\mathrm{D_{LV}}}} f(z,t)\mathcal{N}(\mu,\Sigma)\mathrm{d}z, \tag{6-50}$$

$$\begin{aligned}&\frac{\mathrm{d}\Sigma}{\mathrm{d}t}\\ &= \int_{\mathbb{R}^{\mathrm{D_{LV}}}} [f(z,t)(z-\mu)^\top + (z-\mu)f(z,t)^\top]\mathcal{N}(\mu,\Sigma)\mathrm{d}z + \int_{\mathbb{R}^{\mathrm{D_{LV}}}} L(z,t)\mathcal{Q}L^\top(z,t)\mathcal{N}(\mu,\Sigma)\mathrm{d}z.\end{aligned} \tag{6-51}$$

为了规避积分项的计算，引入下式所示的斯坦因引理[204]（Stein's lemma）：

$$\int_{\mathbb{R}^{\mathrm{D_{LV}}}} f(z,t)(z-\mu)^\top\mathcal{N}(\mu,\Sigma)\mathrm{d}z = [\int_{\mathbb{R}^{\mathrm{D_{LV}}}} \frac{\partial f(z,t)}{\partial z}\mathcal{N}(\mu,\Sigma)\mathrm{d}z]\Sigma. \tag{6-52}$$

将式(6-52)代入式(6-50)和式(6-51)可以得到如下所示的式 (6-53)和式(6-54)：

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = \mathbb{E}[f(z,t)], \tag{6-53}$$

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}t} = \Sigma\mathbb{E}[\frac{\partial f(z,t)}{\partial z}]^\top + \mathbb{E}[\frac{\partial f(z,t)}{\partial z}]\Sigma^\top + \mathbb{E}[L(z,t)\mathcal{Q}L^\top(z,t)]. \tag{6-54}$$

此外，根据扩展卡尔曼滤波（extended kalman filtering）方法[205]，引入式(6-55)和式(6-56)对高斯分布进行展开：

$$f(z,t) \approx f(\mu,t) + [\frac{\partial f(z,t)}{\partial z}|_{z=\mu}](z-\mu), \tag{6-55}$$

$$L(z,t) = L(\mu,t) = \mathcal{I}. \tag{6-56}$$

则式(6-44)与式(6-45)得证。 **证毕**

由于在不同时刻之间的隐状态的方差表达式在实现时需要显式地计算雅可比矩阵，因此为了规避这一复杂的计算成本，本章引入下式所示的平面流（planar flow）模型以建模转移函数 $f_\theta(z)$：

$$f_\theta(z_t) = z_t + u_t\tanh(zW + b), \tag{6-57}$$

其中 tanh、W 和 b 分别为双曲正切激活函数、可学习的权重和偏置参数。在此基础上，对应的偏导数 $\frac{\partial f_\theta(z_t)}{\partial z_t}$ 可以通过下式显示地给出：

$$\frac{\partial f(z_t)}{\partial z_t} = \mathcal{I} + u_t[\frac{\mathrm{d}\tanh(z_t)}{\mathrm{d}z_t}|_{z=\mu_t W+b}]W = \mathcal{I} + u_t[1-\tanh^2(\mu_t W+b)]W. \tag{6-58}$$

在此基础上，本小节进一步给出下列定理说明在矩展开策略下生成网络损失函数的在深度学习后端的实现方法：

定理 6.7： 在软测量任务场景下，似然项 $\mathbb{E}_{\mathbb{Q}}[\|y_t - g(z_t, u_{1:t})\|_2^2]$ 可以近似为下式：

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[\|y_t - g(z_t, u_{1:t})\|_2^2\right] \\
\approx&\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathcal{I})}[\mathcal{L}^{\mathrm{MSE}}(\mu_t^z) + (\frac{\partial \mathcal{L}^{\mathrm{MSE}}(\mu_t^z)}{\partial z}|_{z=\mu_t^z})(\Sigma_t^z)^{\frac{1}{2}}(\frac{\partial \mathcal{L}^{\mathrm{MSE}}(\mu_t^z)}{\partial z}|_{z=\mu_t^z})^\top \times \epsilon] \\
:=&\mathcal{L}^{\mathrm{MLL}}(\mu_t^z),
\end{aligned} \tag{6-59}$$

其中上标“MLL”为“似然函数的矩”（moment of log-likelihood）。

证明： 在软测量建模任务的场景下，预测值应该在隐空间进行用观测数据后验更新后给出。换言之，从 z 解码出相关信息 y 的 z 必须是先验过程得到的 z 而不是用 y 进行滤波后的 z。因此，为了和软测量建模的场景保持一致，在 $\mathbb{Q}(z)$ 测度下的似然项应该如式(6-60)所示，应该替换为在 $\mathbb{P}(z)$ 测度下的似然项：

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}(z)}[\|y_t - g(z_t, u_{1:t})\|_2^2] \\
=&\mathbb{E}_{\mathbb{P}(z)}[\exp(\int_{t-1}^{t}\|\nu\|_2^2\mathrm{d}\tau)\|y_t - g(z_t, u_{1:t})\|_2^2] \\
\overset{\text{(i)}}{\approx}&\mathbb{E}_{\mathbb{P}(z)}[\|y_t - g(z_t, u_{1:t})\|_2^2].
\end{aligned} \tag{6-60}$$

其中步骤“(i)”的近似操作是考虑了实际应用场景的局限性。具体而言，根据定理6.5可以观察到 ν 含有 y_t 的信息；然而，在工业软测量建模任务中的测试阶段，由于任务的因果性，y_t 的预测是基于 $\mathbb{P}(z)$ 下的 z_t 进行预测的而不是经过 ν 修改后的 $\mathbb{Q}(z)$ 下的 z_t 进行预测的。因此，假设一个优化良好的模型将导致控制策略趋于零 ($\|\nu\|_2^2 \approx 0$)，则可以得到 $\exp(\int_{t-1}^{t}\|\nu\|_2^2\mathrm{d}\tau) \approx 1$（该假设的合理性将在本章后续的实验部分进行说明）。

在此基础上，可以定义如下式所示的与预测函数 $g_\theta(z_t)$ 有关的均方误差 $\mathcal{L}^{\mathrm{MSE}}(z_t)$：

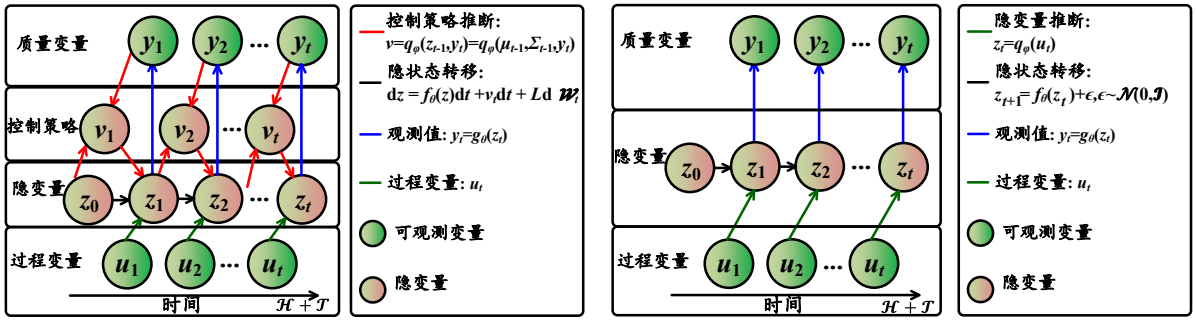
$$\mathcal{L}^{\mathrm{MSE}}(z_t) := \|y_t - g(z_t, u_{1:t})\|_2^2. \tag{6-61}$$

则对隐变量 z 进行如式(6-55)所示的矩展开，可以得到如下结果：

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}[\mathcal{L}(z_t)] \\
=&\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathcal{I})}[\mathcal{L}(\mu_t^z+\sigma_{z,t}\times\epsilon)] \\
\approx&\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathcal{I})}[\mathcal{L}^{\mathrm{MSE}}(\mu_t^z)+(\frac{\partial\mathcal{L}^{\mathrm{MSE}}(\mu_t^z)}{\partial z}|_{z=\mu_t^z})(\Sigma_t^z)^{\frac{1}{2}}(\frac{\partial\mathcal{L}^{\mathrm{MSE}}(\mu_t^z)}{\partial z}|_{z=\mu_t^z})^{\top}\times\epsilon] \\
:=&\mathcal{L}^{\mathrm{MLL}}(\mu_t^z).
\end{aligned} \tag{6-62}$$

则式(6-59)得证。　　**证毕**

6.4.6　模型结构论述及变分推断算法总结



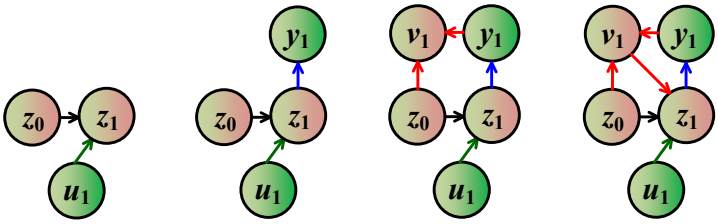
(a) OC-NDPLVM　　(b) 常规 NDPLVM

图 6.3 OC-NDPLVM 与常规 NDPLVM 结构对比图

基于第6.4.3小节和6.4.4小节的内容，本章提出的新的NDPLVM的模型结构如图6.3(a)所示。从图中可以观察到，该模型由两种不同颜色的节点构成：绿色节点表示可观测的过程测量数据，而橙色节点则代表不可观测的隐变量。结合摊销变分推断的思想，推断网络 $q_\varphi(z_{t-1},y_t)$ 用于生成最优控制策略 ν_t^*。由于所提模型依赖于最优控制这一泛函优化方法导出，因此本小节将该模型命名为“最优控制-非线性动态概率隐变量模型”（Optimal Control-Nonlinear Probabilistic Latent Variable Model，OC-NDPLVM）。为了更清晰地比较OC-NDPLVM与常规的NDPLVM的区别，本小节在图6.3(b)中展示了常规NDPLVM的结构。从图6.3(a)和图6.3(b)的对比中可以看出，OC-NDPLVM与传统NDPLVM的主要不同之处在于隐变量的推断方式。OC-NDPLVM需要基于观测数据，通过生成的最优控制策略 ν_t^* 进行校正后，才能开展隐变量的推断。而常规NDPLVM则直接利用过程变量 u_t 进行推断。这是因为，根据第6.4.3小节和6.4.4小节中提出的理论分析，OC-NDPLVM

的模型推导依赖最优控制解的结构特性。具体而言，根据定理6.5，最优控制策略 ν_t 的输入必须包含当前观测数据（y_t）和隐变量的历史信息（z_{t-1}），而根据摊销变分推断的思想，推断网络作为模拟最优策略的模拟器，其结构应与 ν_t 的解的形式保持一致。因此，与传统 NDPLVM 相比，OC-NDPLVM 在模型中额外引入了控制策略节点，以实现该推断步骤从而做到先验隐变量推断结果进行“后验更新”。

在此基础上，为了更直观地理解 OC-NDPLVM 的推断过程，图6.4进一步展示了 $t=1$ 时，$t-1$ 到 t 时刻 OC-NDPLVM 的推理流程示意图。



(a) 先验过程转移 (b) 质量变量预测 (c) 控制策略推断 (d) 后验过程校正

图 6.4 OC-NDPLVM 从 $t-1$ 到 t 时刻的推断流程示意图（图中 $t=1$）

结合图6.4的内容以及第6.4.5小节中讨论的矩方法实现策略，OC-NDPLVM 的推理流程可以分为以下几个步骤：

- **图6.4(a)，先验过程转移：**根据式(6-44)和式(6-45)中定义的 ODE，模拟从 $t-1$ 到 t 时刻隐变量的先验转移过程，得到测度 $\mathbb{P}$ 下的隐变量分布。
- **图6.4(b)，质量变量预测：**根据 $\hat{y}_t = g_\theta(\mu_t)$，利用以 θ 为参数的发射网络（emission network）$g_\theta(\cdot)$ 对质量变量 y_t 进行预测，获得 t 时刻的质量变量估计值。
- **图6.4(c)，控制策略推断：**根据 y_t（当 $t \in \{1,\ldots,\mathcal{H}\}$ 时）或 $\hat{y}_t$（当 $t \in \{\mathcal{H}+1,\ldots,\mathcal{H}+\mathcal{T}\}$）与 μ_{t-1} 和 Σ_{t-1} 作为推断网络 q_φ 的输入对最优控制策略 ν_t^* 进行推断。
- **图6.4(d)，后验过程校正：**利用式(6-43)对 t 时刻的隐变量的后验校正，得到测度 $\mathbb{Q}$ 下的隐变量分布。

通过重复上述步骤，模型能够递推至 $\mathcal{T}+\mathcal{H}$ 的时间步长，完成整个数据的建模与推断过程。在此基础上，算法6.1给出了 OC-NDPLVM 的推理流程。其中由于在 $\mathcal{T}+1$ 至

$\mathcal{T}+\mathcal{H}$ 时模型，质量变量 y_t 的信息是未知的，因此第 14 行将第 6 行的 y_{t+1} 替换为输出网络 g_θ 给出的 $\hat{\mu}^y_{t+1}$ 作为推断网络的输入进行最优控制策略的推断。

算法 6.1 OC-NDPLVM 的推理算法伪代码

输入： 起始质量变量序列：$\{y_1, y_2, ..., y_\mathcal{T}\}$ 和过程变量序列：$\{u_1, u_2, ..., u_{\mathcal{T}+\mathcal{H}}\}$。
参数： 推断网络参数 φ 和生成网络 θ。
输出： 预测质量变量序列 $[\hat{\mu}^y_{\mathcal{T}+1}, ..., \hat{\mu}^y_{\mathcal{T}+\mathcal{H}}]$。

1: 设置 $t=0$ 时刻的隐变量 z_0 的均值 $\mu^z_0=0$ 和协方差 $\Sigma^z_t=\mathcal{I}$；
2: **for** $t \leftarrow 0$ **to** $\mathcal{T}$ **do**
3: $\quad \mu^{z,\mathbb{P}}_{t+1} \leftarrow$ 式 (6-44) > 先验过程转移
4: $\quad \Sigma^{z,\mathbb{P}}_{t+1} \leftarrow$ 式 (6-45) > 先验过程转移
5: $\quad \hat{\mu}^y_{t+1} \leftarrow g_\theta(\mu^{z,\mathbb{P}}_{t+1})$ > 质量变量预测
6: $\quad [\mu^\nu_{t+1}, \Sigma^\nu_{t+1}] \leftarrow q_\phi(\mu^{z,\mathbb{P}}_t, \Sigma^{z,\mathbb{P}}_t, y_{t+1})$ > 控制策略推断
7: $\quad \mu^{z,\mathbb{Q}}_{t+1} \leftarrow (\Sigma^\nu_t + \Sigma^{z,\mathbb{P}}_t)^{-1}(\Sigma^\nu_t \mu^\nu_t + \Sigma^{z,\mathbb{P}}_t \mu^{z,\mathbb{P}}_t)$ > 后验过程校正
8: $\quad \Sigma^{z,\mathbb{Q}}_{t+1} \leftarrow (\Sigma^\nu_t + \Sigma^{z,\mathbb{P}}_t)^{-1}(\Sigma^\nu_t \Sigma^{z,\mathbb{P}}_t)$ > 后验过程校正
9: **end for**
10: **for** $t \leftarrow \mathcal{T}+1$ **to** $\mathcal{T}+\mathcal{H}-1$ **do**
11: $\quad \mu^{z,\mathbb{P}}_{t+1} \leftarrow$ 式 (6-44) > 先验过程转移
12: $\quad \Sigma^{z,\mathbb{P}}_{t+1} \leftarrow$ 式 (6-45) > 先验过程转移
13: $\quad \hat{\mu}^y_{t+1} \leftarrow g_\theta(\mu^{z,\mathbb{P}}_{t+1})$ > 质量变量预测
14: $\quad [\mu^\nu_{t+1}, \Sigma^\nu_{t+1}] \leftarrow q_\phi(\mu^{z,\mathbb{P}}_t, \Sigma^{z,\mathbb{P}}_t, \hat{\mu}^y_{t+1})$ > 控制策略推断
15: $\quad \mu^{z,\mathbb{Q}}_{t+1} \leftarrow (\Sigma^\nu_t + \Sigma^{z,\mathbb{P}}_t)^{-1}(\Sigma^\nu_t \mu^\nu_t + \Sigma^{z,\mathbb{P}}_t \mu^{z,\mathbb{P}}_t)$ > 后验过程校正
16: $\quad \Sigma^{z,\mathbb{Q}}_{t+1} \leftarrow (\Sigma^\nu_t + \Sigma^{z,\mathbb{P}}_t)^{-1}(\Sigma^\nu_t \Sigma^{z,\mathbb{P}}_t)$ > 后验过程校正
17: **end for**
18: **return** $[\hat{\mu}^y_{\mathcal{T}+1}, \ldots, \hat{\mu}^y_{\mathcal{T}+\mathcal{H}}]$

算法 6.2 OC-NDPLVM 的训练算法伪代码

输入： 训练数据集：$\mathcal{D}_{\text{train}} = \{[u_{1:\mathcal{T}+\mathcal{H},1}, y_{1:\mathcal{T}+\mathcal{H},1}], \ldots, [u_{1:\mathcal{T}+\mathcal{H},\text{N}_{\text{train}}}, y_{1:\mathcal{T}+\mathcal{H},\text{N}_{\text{train}}}]\}$、验证数据集：$\mathcal{D}_{\text{valid}} = \{[u_{1:\mathcal{T}+\mathcal{H},1}, y_{1:\mathcal{T}+\mathcal{H},1}], \ldots, [u_{1:\mathcal{T}+\mathcal{H},\text{N}_{\text{valid}}}, y_{1:\mathcal{T}+\mathcal{H},\text{N}_{\text{valid}}}]\}$、批次大小 $\mathcal{B}$、迭代轮次 $\mathcal{E}$、网络学习率 η。
参数： 推断网络参数 φ^* 和生成网络 θ^*。
输出： 预测质量变量序列 $[\hat{\mu}^y_{\mathcal{T}+1}, ..., \hat{\mu}^y_{\mathcal{T}+\mathcal{H}}]$。

1: 生成网络 θ^* 和推断网络参数 φ^*；
2: **for** $e \leftarrow 0$ **to** $\mathcal{E}-1$ **do**
3: $\quad$ 采样小批量数据 $\mathcal{D}_{\text{minibatch}} \subset \mathcal{D}_{\text{train}}$
4: $\quad [\hat{\mu}^y_{\text{T}+1}, ..., \hat{\mu}^y_{\text{T+H}}] \leftarrow$ Algorithm 6.1
5: $\quad$ 计算损失函数 $\mathcal{L}^{\text{MLL}} \leftarrow \frac{1}{2\mathcal{B}} \sum_{t=\mathcal{T}+1}^{\mathcal{T}+\mathcal{H}} \|y_t - \mu^y_t\|_2^2 + [\mu^\nu_t]^\top [\mu^\nu_t] + \text{Trace}(\Sigma_t)$
6: $\quad \theta_e \leftarrow \theta_{e-1} - \eta \nabla_\theta \mathcal{L}^{\text{MLL}}|_{\theta=\theta_{e-1}}$ > 更新生成网络参数
7: $\quad \varphi_e \leftarrow \varphi_{e-1} - \eta \nabla_\varphi \mathcal{L}^{\text{MLL}}|_{\varphi=\varphi_{e-1}}$ > 更新推断网络参数
8: **end for**
9: 在验证数据集 $\mathcal{D}_{\text{validate}}$ 上存储最好的生成网络参数 $\theta^* = \theta_{\text{best}}$ 和推断网络参数 $\varphi^* = \varphi_{\text{best}}$；
10: **return** 生成网络 θ^* 和推断网络参数 φ^*

基于算法6.1，本小节进一步总结了 OC-NDPLVM 的训练算法，详见算法6.2。同时，结合定义1，给出下列定理以说明 OC-NDPLVM 训练过程的收敛性（其中，条件（1）为“摊销变分推断”中常用的假设，详见2.1.2节）。

定理 6.8： 设 $\{\mathcal{L}_{\tau}^{\mathrm{MLL}}\}_{\tau=1}^{\mathcal{E}}$ 为 OC-NDPLVM 在训练过程中生成的损失函数序列。若（1）推断网络 q_{φ} 能模拟最优控制策略 ν^{*}；并且（2）学习率 η 充分小，则存在 $\mathcal{C}\geq 0$，使得对于任意给定的 $\gamma>0$，存在正整数 N 使得当 $\tau>N$ 时有 $\|\mathcal{L}_{\tau}^{\mathrm{MLL}}-\mathcal{C}\|<\gamma$ 成立。

证明： 假设第 e 轮迭代的损失函数值为 $\mathcal{L}_{e}^{\mathrm{MLL}}$，参数优化之前的损失函数值（即算法6.1第6-7 行）为 $\mathcal{L}_{e+\frac{1}{2}}^{\mathrm{MLL}}$。由于假设条件（1）成立，下述不等式成立：

$$\mathcal{L}_{e+\frac{1}{2}}^{\mathrm{MLL}}\leq\mathcal{L}_{e}^{\mathrm{MLL}} \tag{6-63}$$

这一不等式的成立的理由可以论述如下：由于推断网络能够生成最优控制策略，根据定理6.5，损失函数值 $\mathcal{L}_{e+\frac{1}{2}}^{\mathrm{MLL}}$ 必然小于或等于 $\mathcal{L}_{e}^{\mathrm{MLL}}$。

而当 $\eta\to 0$ 的时候，定义全体参数集合 $\Theta=\varphi\cup\theta$，可以得到如下条件：

$$\frac{\mathrm{d}\Theta}{\mathrm{d}\tau}=-\nabla_{\Theta}\mathcal{L}^{\mathrm{MLL}} \tag{6-64}$$

因此，随着 τ 的演化，下列不等式成立：

$$\frac{\mathrm{d}\mathcal{L}^{\mathrm{MLL}}}{\mathrm{d}\tau}=[\frac{\partial\mathcal{L}^{\mathrm{MLL}}}{\partial\Theta}]^{\top}[\frac{\mathrm{d}\Theta}{\mathrm{d}\tau}]=-[\nabla_{\Theta}\mathcal{L}^{\mathrm{MLL}}]^{\top}[\nabla_{\Theta}\mathcal{L}^{\mathrm{MLL}}]\leq 0. \tag{6-65}$$

在此基础上，在学习率足够小的条件下，则参数学习的过程越接近式(6-64)所给出的ODE。根据上述分析，可以得到如下结论：

$$\mathcal{L}_{e+1}^{\mathrm{MLL}}\leq\mathcal{L}_{e+\frac{1}{2}}^{\mathrm{MLL}} \tag{6-66}$$

将不等式(6-66)代入不等式(6-63)可以得到如下不等式：

$$\mathcal{L}_{e+1}^{\mathrm{MLL}}\leq\mathcal{L}_{e}^{\mathrm{MLL}} \tag{6-67}$$

这表明损失函数 $\mathcal{L}^{\mathrm{MLL}}$ 是单调递减的。此外，根据定理6.3，损失函数 $\mathcal{L}^{\mathrm{MLL}}$ 存在负对数似然函数下界限制：

$$\mathcal{L}^{\mathrm{MLL}}\geq-\log p(\vec{y}|\vec{u}) \tag{6-68}$$

结合不等式(6-67)和(6-68)，则 OC-NDPLVM 的训练过程过程收敛。 **证毕**

6.5 实验验证

在本节，将对本章所提出的 OC-NDPLVM 的有效性进行实验验证。本节的实验将从如下四个研究问题进行展开：

- **问题 1（表现性）：**OC-NDPLVM 在软测量建模任务中的表现性如何？
- **问题 2（增益机理）：**是什么让 OC-NDPLVM 取得了如此好的表现效果？
- **问题 3（敏感性）：**随着超参数的调整，OC-NDPLVM 的性能变化趋势如何？
- **问题 4（收敛性）：**OC-NDPLVM 的训练过程中是否收敛？

6.5.1 软测量建模精度对比实验

本小节聚焦于**问题 1**：“OC-NDPLVM 在软测量建模任务中的表现性如何?”为了回答这一问题，本小节将在脱丁烷精馏塔和水煤气变换单元这两个经典化工过程的反应和分离单元操作中进行软测量建模实验，以展示 OC-NDPLVM 的优越性。为了更加充分地展示 OC-NDPLVM 的性能，本节设置 $\mathcal{T}=5$ 并选择了以下软测量模型作为基线模型进行相关实验。这些基线模型根据其结构设计原理可分为以下几个大类：

- **循环神经网络类（自回归结构）：**自回归-时间卷积网络[39]（auto-regressive temporal convolution network，AR-TCN）、双重注意力长短期记忆网络[213]（dual- attention long-short-term-memory，DA-LSTM）；
- **概率隐变量类模型（稳态/自回归结构）：**非线性概率隐变量回归[13]（nonlinear probabilistic latent variable regression，NPLVR）、动态概率隐变量模型[58]（dynamic probabilistic latent variable model，DPLVM）；概率性判别时间序列模型[96]（probabilistic discriminative time-series model，PDTM）；深度贝叶斯概率慢特征分析模型[97]（deep bayesian probabilistic slow feature analysis，DBPSFA）
- **变压器网络类（非自回归结构）：**信息变压器网络[214]（Informer）、局部变压器网络[197]（LogTrans）；
- **有限混合模型类：**迪利克雷过程混合模型[215]（Dirichlet process mixture model，DPMM）和动态混合变分自编码器回归模型[99]（dynamical mixture variational autoencoder regression，DMVAER）。

相关实验超参数将在附录E中详细说明。

由于本章需要预测未来 $\mathcal{H}$ 个时间窗口内的质量变量，因此本小节借鉴文献[214]的思路，对式(2-37a)和式(2-37d)修改为式(6-69a)和式(6-69b)，采用基于时间窗口的评估指标来衡量软测量建模的精确性，其中“w”是窗口（window）的英文首字母：

$$\text{wRMSE} = \frac{1}{\text{N}_{\text{test}}}\frac{1}{\mathcal{H}}\sum_{n=1}^{\text{N}}\sum_{h=1}^{\text{N}_{\text{test}}}\sqrt{(\widehat{y}_{h,n}-y_{h,n})^2}, \tag{6-69a}$$

$$\text{wMAE} = \frac{1}{\text{N}_{\text{test}}}\frac{1}{\mathcal{H}}\sum_{n=1}^{\text{N}_{\text{test}}}\sum_{h=1}^{\text{H}}|\widehat{y}_{h,n}-y_{h,n}|, \tag{6-69b}$$

其中，$\widehat{y}$ 表示预测值，y 表示实际值，$\mathcal{H}$ 为预测窗口的长度，N_{test} 表示测试集的样本数量。对于 wRMSE 和 wMAE 而言，其值越低，说明软测量模型的建模精度越高。在此基础上，软测量建模的精度对比实验结果在表6.1列出。

表6.1展示了以下实验现象：

（1） 对于脱丁烷精馏塔数据集，当 $\mathcal{H}=2,3,4,5$ 时，wRMSE 相较于基线模型分别降低了 39.73%∼92.54%、33.02%∼89.46%、17.97% ∼86.29% 和 8.14%∼83.08%；而 wMAE 则分别降低了 42.68%∼93.00%、37.96%∼90.40%、23.14%∼87.71% 和 13.72％∼84.97%。

（2） 对于水煤气变换单元数据集，当 $\mathcal{H}=2,3,4,5$ 时，wRMSE 相较于基线模型分别降低了 0.45%∼80.49%、3.43%∼ 79.90%、0.79%∼78.86% 和 0.17%∼77.92%；而 wMAE 分别降低了 0.65%∼ 81.55%、4.06%∼ 81.32%、1.21%∼80.45% 和 0.28%∼79.21%。

（3） OC-NDPLVM 相较于大多数基线模型的性能提升是显著的，这一点可以通过成对样本 t-检验中 p 值小于 0.05 得到验证。

（4） 基于递归神经网络和自注意力机制的方法在性能上优于基于概率隐变量模型的方法以及基于混合模型的方法。

（5） 线性概率隐变量模型（如 DPLVM）和线性混合模型（如 DPMM）的性能低于其非线性版本（如 NPLVM 和 DMVAER）。

（6） 当预测窗口长度增大时，与 NDPLVM 和循环神经网络方法相比，自注意力模型的性能退化幅度更小。

表 6.1 脱丁烷精馏塔和水煤气变换单元数据集上的软测量建模精度对比结果

数据集	模型	$\mathcal{H}=2$		$\mathcal{H}=3$		$\mathcal{H}=4$		$\mathcal{H}=5$	
		wRMSE	wMAE	wRMSE	wMAE	wRMSE	wMAE	wRMSE	wMAE
脱丁烷精馏塔	AR-TCN	0.0234†	0.0230†	0.0298†	0.0292†	0.0367†	0.0356†	0.0442†	0.0427†
	DA-LSTM	0.0341†	0.0323†	0.0393†	0.0367†	0.0379†	0.0357†	0.0502†	0.0480†
	NPLVR	0.1393†	0.1390†	0.1399†	0.1392†	0.1401†	0.1389†	0.1405†	0.1389†
	DPLVM	0.1722†	0.1714†	0.1732†	0.1718†	0.1741†	0.1721†	0.175†	0.1723†
	PDTM	0.1512†	0.1499†	0.1433†	0.1419†	0.1427†	0.1411†	0.1413†	0.1393†
	DBPSFA	0.1424†	0.1412†	0.1413†	0.1394†	0.1420†	0.1400†	0.1445†	0.1415†
	LogTrans	0.0326†	0.0319†	0.0344†	0.0333†	0.0346†	0.0328†	0.0409†	0.0387†
	Informer	0.0243†	0.0238†	0.0314†	0.0305†	0.0318†	0.0302†	0.0351	0.0329†
	DPMM	0.1896†	0.1893†	0.1900†	0.1894†	0.1904†	0.1894†	0.1909†	0.1895†
	DMVAER	0.1394†	0.1389†	0.1395†	0.1387†	0.1401†	0.1388†	0.1407†	0.1388†
	OC-NDPLVM	**0.0141**	**0.0132**	**0.0200**	**0.0181**	**0.0260**	**0.0232**	**0.0322**	**0.0284**
超越基线模型数		10	10	10	10	10	10	10	10
数据集	模型	$\mathcal{H}=2$		$\mathcal{H}=3$		$\mathcal{H}=4$		$\mathcal{H}=5$	
		wRMSE	wMAE	wRMSE	wMAE	wRMSE	wMAE	wRMSE	wMAE
水煤气变换单元	AR-TCN	0.1036	0.0945	0.1119	0.0991†	0.1157	0.1007	0.1201	0.1037
	DA-LSTM	0.1793†	0.1631†	0.1965†	0.1713†	0.2153†	0.1836†	0.2268†	0.1910†
	NPLVR	0.5340†	0.5150†	0.5411†	0.5129†	0.5472†	0.5136†	0.5525†	0.5152†
	DPLVM	0.6046†	0.5843†	0.6162†	0.5856†	0.6230†	0.5861†	0.6277†	0.5864†
	PDTM	0.1309†	0.1204†	0.1460†	0.1298†	0.1595†	0.1401†	0.1563†	0.135†
	DBPSFA	0.1450†	0.1329†	0.1648†	0.1441†	0.1818†	0.1545†	0.1948†	0.1612†
	LogTrans	0.1065	0.0974	0.1136	0.1008	0.1186	0.1036	**0.1197**	**0.1033**
	Informer	0.1071†	0.0978†	0.1152	0.1023	0.1184	0.1033	0.1218	0.1054
	DPMM	0.6033†	0.5899†	0.6095†	0.5892†	0.6132†	0.5884†	0.6158†	0.5876†
	DMVAER	0.5296†	0.5103†	0.5381†	0.5098†	0.5440†	0.5102†	0.5472†	0.5098†
	OC-NDPLVM	**0.1031**	**0.0939**	**0.1081**	**0.0951**	**0.1148**	**0.0995**	0.1199	0.1034
超越基线模型数		10	10	10	10	10	10	9	9

注：匕首符号 † 表示在成对样本 t-检验中，OC-NDPLVM 相比其他基线模型在统计学上具有显著性差异（$p < 0.05$）。**粗体字**标注的结果为各指标的最优表现，下划线标注结果为各指标的次优表现。

从现象（1）和（2）可以看出，OC-NDPLVM 的表现优于其他基线模型。现象（3）表明，在大多数场景下，OC-NDPLVM 的性能确实优于其他基线模型。与此同时，现象（4）揭示了 NPLVM 在软测量建模任务上仍有很大的改进空间，这从实践角度进一步证明了本文在推断网络输入选择和矩展开策略上的合理性与必要性。现象（5）强调了引入深度学习架构对于提升 NDPLVM 性能的重要作用。现象（6）表明，自回归结构的模型在预测窗口增加时可能会面临误差累积的问题，最终导致模型性能出现下降，而自注意力模型由于其非自回归的设计能够有效缓解误差累积问题。总而言之，本小节通过对脱丁烷精馏塔和水煤气变换单元的软测量精度对比试验从实验上验证了 OC-NDPLVM 在软测量建模任务上的优越性，并从实践层面上回答了**问题 1**，为后续在工业场景下部署 OC-NDPLVM 提供了实验基础。

6.5.2 消融实验

在6.5.1节的基础上，本小节进一步探讨 OC-NDPLVM 在软测量建模精度方面取得优异表现的原因，并尝试回答**问题 2**：“是什么让 OC-NDPLVM 取得了如此好的表现效果?”为此，本小节通过消融实验来研究模型的两个关键组成部分，以阐明 OC-NDPLVM 的有效性。具体而言，本小节对以下两个关键模块进行消融实验：

- **基于最优控制设计的推断网络输入：**在此模块的消融实验中，OC-NDPLVM 在 t 时刻的推断网络输入从 y_t 和 z_{t-1} 替换为 u_t；
- **矩展开实现策略：**OC-NDPLVM 在 $t-1$ 到 t 时刻的转移函数通过模拟 $z_t = z_{t-1} + f_\theta(z_t) + \nu + \sqrt{2}\epsilon, \epsilon \sim \mathcal{N}(0, \mathcal{I})$ 实现，而 $\hat{y}_t$ 的预测有 z_t 输入发射网络 g_θ 后给出。

实验结果在表6.2中进行了展示。

表6.2展示了以下实验现象：

（1） 当推断网络输入的设计未充分考虑基于最优控制方法所得的最优控制策略的函数形式时，模型性能出现了明显下降（表中每个数据集及预测窗口的第 1 行与第 3 行）。这一现象与表6.1中的观察一致，进一步强调了利用最优控制方法推导最优控制策略的函数结构，并以此指导推断网络输入变量选择的关键作用。同时，这一实验结果从实践角度验证了6.4.4节所提出方法的必要性和有效性。

表 6.2 脱丁烷精馏塔和水煤气变换单元数据集上的消融实验结果

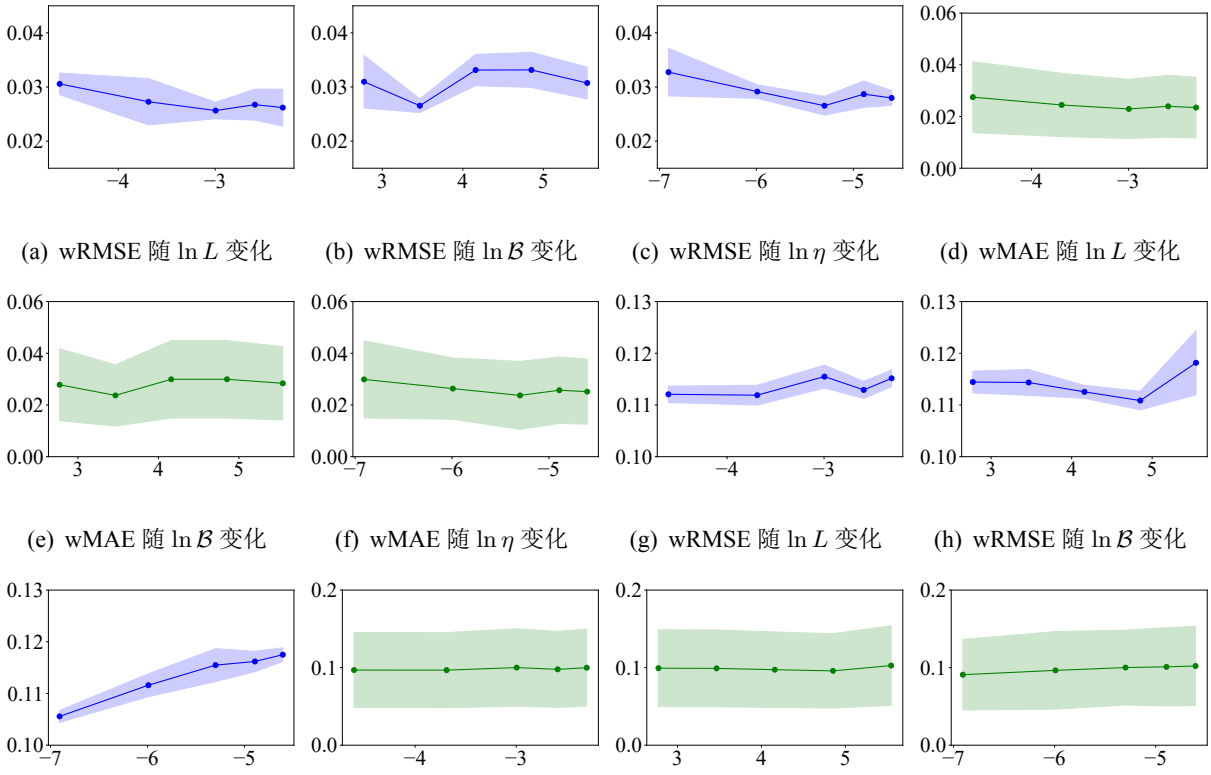
预测时间窗	最优控制	矩展开	脱丁烷精馏塔		水煤气变换单元	
			wRMSE(↑)	wMAE(↑)	wRMSE(↑)	wMAE(↑)
$\mathcal{H}=2$	✗	✓	1083%	1103%	239.0%	240.9%
	✓	✗	210.9%	218.3%	24.90%	26.30%
	✗	✗	903.9%	957.8%	316.4%	326.8%
$\mathcal{H}=3$	✗	✓	848.4%	846.7%	276.5%	281.8%
	✓	✗	149.0%	162.2%	34.80%	37.80%
	✗	✗	615.1%	673.5%	332.6%	347.6%
$\mathcal{H}=4$	✗	✓	641.7%	642.5%	296.2%	303.8%
	✓	✗	190.2%	211.9%	18.80%	20.60%
	✗	✗	453.8%	507.7%	355.3%	378.3%
$\mathcal{H}=5$	✗	✓	563.8%	545.6%	330.8%	338.2%
	✓	✗	133.7%	151.8%	18.50%	19.40%
	✗	✗	349.7%	397.5%	363.3%	382.7%

（2） 当移除基于矩展开的模型实现策略时（如表中每个数据集和预测窗口的第 2 行与第 3 行所示），模型性能同样出现下降。这一结果表明，矩展开策略在软测量建模任务中对于 NDPLVM 的实现与部署至关重要，并从实验上进一步证明了6.4.5节所提出方法的现实意义。

上述实验现象突出了两个关键点：第一，在推断网络中引入基于质量变量信息的最优控制策略的函数结构信息对于保证 NDPLVM 在实际运行过程中的优良性能至关重要；第二，遵循矩展开策略所提供的模型实现策略是确保 NDPLVM 在软测量建模任务中具备鲁棒性和可靠性的必要条件。综上所述，本小节的消融实验表明，根据最优控制的函数结构对推断网络进行设计并在矩展开的框架下实现 NDPLVM 是保证 OC-NDPLVM 在软测量建模任务中取得优良性能的关键，并进一步从实践层面上回答了**问题 2**。

6.5.3 敏感性分析

在前一小节消融实验的基础上，本小节将进一步探讨**问题 3**：“随着超参数的调整，OC-NDPLVM 的性能变化趋势如何？”，以深入分析 OC-NDPLVM 的性能与超参数变化之间的关系，从而评估模型对超参数的敏感性为实际应用提供指导。为此，本小节重点考察了波动项水平 L、批次大小 $\mathcal{B}$ 以及学习率 η 的变化对 OC-NDPLVM 性能的影响。实验结果如图6.5(a)至图6.5(l)所示，其中阴影部分表示 ±0.5 倍标准差。



(a) wRMSE 随 $\ln L$ 变化　(b) wRMSE 随 $\ln \mathcal{B}$ 变化　(c) wRMSE 随 $\ln \eta$ 变化　(d) wMAE 随 $\ln L$ 变化

(e) wMAE 随 $\ln \mathcal{B}$ 变化　(f) wMAE 随 $\ln \eta$ 变化　(g) wRMSE 随 $\ln L$ 变化　(h) wRMSE 随 $\ln \mathcal{B}$ 变化

(i) wRMSE 随 $\ln \eta$ 变化　(j) wMAE 随 $\ln L$ 变化　(k) wMAE 随 $\ln \mathcal{B}$ 变化　(l) wMAE 随 $\ln \eta$ 变化

图 6.5 OC-NDPLVM 在 (a)-(f) 脱丁烷精馏塔和 (g)-(l) 水煤气变换单元的敏感性分析结果

图6.5展示了以下实验现象：

（1） 随着波动项水平 L 的增加，OC-NDPLVM 的软测量建模精度在脱丁烷精馏塔数据集呈现了上升趋势而在水煤气变化单元上呈现了下降趋势。

（2） 随着批次大小 $\mathcal{B}$ 的增加，OC-NDPLVM 的软测量建模精度在脱丁烷精馏塔数据集和水煤气变化单元上呈现了先提升后下降的趋势。

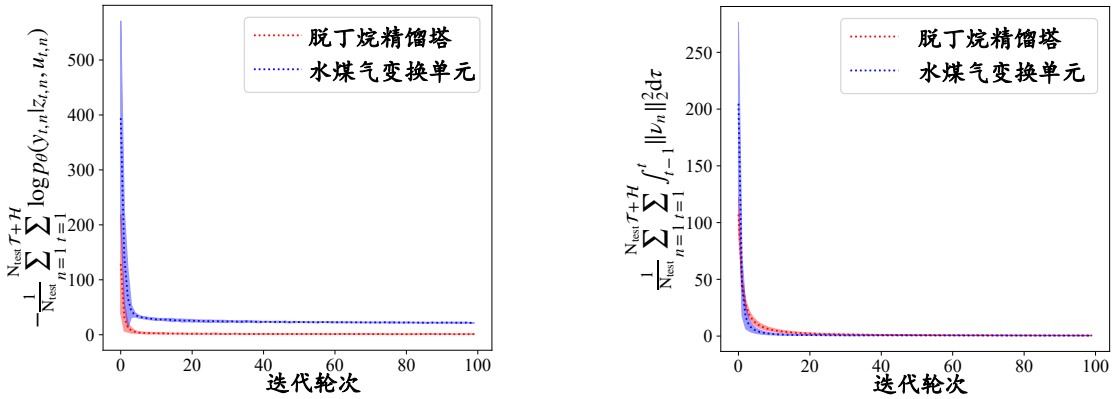
（3） 随着学习率 η 的增加，与波动项水平 L 类似地，OC-NDPLVM 的软测量建模精度在脱丁烷精馏塔数据集呈现了上升趋势而在水煤气变化单元上呈现了下降趋势。

从现象（1）和（3）可知，波动项水平 L 与学习率 η 对 OC-NDPLVM 性能的影响主要取决于数据集的噪声水平。在脱丁烷精馏塔这类含有多组分混合物的分离单元操作中，数据采集过程往往带有较高噪声，因此较高的波动项水平 L 和学习率 η 能为模型提供更强的先验正则化，从而有效减弱噪声带来的不利影响，提升模型精度。相比之下，在水煤气变换这样进口原料纯度较高的反应单元中，由于数据相对纯净，仅需较低的 L

和 η 即可获得更佳的建模精度。与此同时，现象（2）指出，在 OC-NDPLVM 的训练过程中需适度选择批次大小 $\mathcal{B}$：若 $\mathcal{B}$ 过小，模型参数会引入过多噪声；若批次大小 $\mathcal{B}$ 过大，则容易发生过拟合。综合上述现象可知，在训练阶段合理设置波动项水平 L、学习率 η 和批次大小 $\mathcal{B}$ 对于确保 OC-NDPLVM 在测试阶段保持最佳性能至关重要。该结论从实践角度有效解答了**问题 3**，并为 OC-NDPLVM 在工业场景的部署提供了指导。

6.5.4　收敛性分析

最后，本小节将讨论**问题 4**：“OC-NDPLVM 的训练过程中是否收敛?”为此，本小节考虑将 OC-NDPLVM 在脱丁烷精馏塔和水煤气变换单元数据集上的训练过程进行展示从而从实践上回答这一问题。由于 OC-NDPLVM 的代价泛函是基于证据下界进行构造的，因此本章分别在图6.6(a)和图6.6(b)展示了式(6-27)右端所定义的 OC-NDPLVM 的损失函数中的似然项目 $\sum_{t=1}^{\mathcal{T}+\mathcal{H}}\mathbb{E}_{\mathbb{Q}_t(z)}[-\log p_\theta(y_t|z_t,u_{1:t})]$ 和控制正则项 $\sum_{t=1}^{\mathcal{T}+\mathcal{H}}\int_{t-1}^{t}\frac{1}{2}\|\nu\|_2^2\mathrm{d}\tau]$ 在训练过程中的随着迭代轮次的变化结果。图中的阴影部分代表了 ±1.5 倍标准差。从图6.6(a)和图6.6(b)中可以发现，OC-NDPLVM 的似然项和控制正则项均随着迭代轮次的增加而单调递减（红色和蓝色点划线），并且在前 20 个迭代轮次内接近平坦，且标准差在迭代后期逐渐变小（红色和蓝色阴影部分）。上述现象表明 OC-NDPLVM 的训练过程中呈现较快且较为稳定的收敛特性，从实践上回答了**问题 4**，并进一步从实践上证明了定理6.7的证明过程中式(6-60)中步骤“(i)”的合理性和定理6.8的恰当性。



(a) 负对数似然项随迭代轮次演化结果　　(b) 控制策略正则化强度随迭代轮次演化结果

图 6.6 OC-NDPLVM 损失函数随迭代轮次变化结果图

6.6 本章小结

基于有限时域最优控制方法，对 NDPLVM 在软测量任务中的性能进行了系统性重构。具体而言，本章通过引入随机微分方程对隐变量的动态演化过程进行数学表征，引入有限时域最优控制问题将传统 NDPLVM 的学习目标进行重构，并通过严格的理论推导证明该重构目标作为传统 NDPLVM 学习目标上界的有效性。在上述重构的基础上，基于 ADMM 框架对目标泛函进行求解，并将隐变量推断的子问题提升至有限时域路径空间 $\mathscr{C}([0,1],\mathbb{R}^{D_{LV}})$ 中的最优控制策略求解，并通过引入庞特里亚金极大值原理对最优控制策略进行解析推导，提出了基于最优控制律的函数结构进行推断网络的设计准则。在上述分析的基础上，进一步结合矩展开方法对模型实现过程进行细化。综合上述框架和推导，提出了 OC-NDPLVM，系统总结了其训练与推理算法流程，并从理论上证明了 OC-NDPLVM 的收敛性。最后通过软测量建模精度对比、消融分析、超参数敏感性分析和收敛性分析这四个方面的实验证明了所提出的 OC-NDPLVM 的有效性。

7 总结与展望

摘要： 本章对全文的研究内容进行总结，并展望了利用泛函优化的概率隐变量建模理论在软测量建模的后续研究方向。

关键词： 总结 展望 泛函优化 隐变量模型 软测量建模

7.1 研究工作总结

随着工业传感技术与信息技术的深度融合，现代生产系统已迈入多源异构数据爆发式增长的时代。这种充足的数据供应为数据驱动的软测量建模提供了坚实的基础，使其在过去二十年间得到了广泛关注和深入研究。在众多建模方法中，基于隐变量模型的数据驱动过程建模方法因其在捕捉复杂过程潜在结构和动态特性方面的优势，被广泛应用于软测量建模中。本文利用泛函优化的相关理论对隐变量模型进行了系统性重构，并在软测量建模任务中验证了本文所提出方法的有效性。本文的研究工作和理论贡献可以总结为如下的四个小点：

（1） **构建隐变量驱动的数据补全方法框架：**在工业过程中，极端操作环境常导致传感器故障并引发数据丢失问题。针对此挑战，第3章深入分析了基于隐变量模型的缺失数据补全方法。第3章从泛函优化理论视角出发通过融合最小移动方案问题与RKHS对模型推理过程进行了系统性重构，有效缓解了传统隐变量模型因隐式引入熵泛函正则项而导致的推理发散问题。同时，第3章提出了一种与缺失机制无关的模型训练策略，从根本上降低了模型对特定缺失模式的依赖性，显著增强了其在多种缺失机制下的鲁棒性与一致性。基于上述理论创新，该章节进一步总结并形式化了面向工业过程的数据缺失值补全算法（KnewImp算法），并从理论上严格证明了该算法的收敛性。实验结果表明，所提方法在缺失数据补全的准确性、模型收敛性及下游软测量建模任务中的性能均表现出显著优势。

（2） **基于无穷时域最优控制角度的稳态隐变量模型：**为克服传统概率隐变量模型训练中因表达能力受限而导致的性能下降问题，第4章基于无穷时域最优控制理论，对

稳态概率隐变量模型的训练过程进行了系统性重构。第4章提出了一种变分分布量子化方法，将原本标量化的变分分布离散为一组定义在高维实数支撑上的有限粒子集合，从而将隐变量分布推断问题转化为无穷时域最优控制问题。在此基础上，通过结合 RKHS，第4章严格推到了该无穷时域最优控制问题的实现方法，并以此为基础提出了全新的隐变量分布推断算法（InfO 算法）与对应的 EM 算法（InfO-EM 算法），大幅提升了模型在处理复杂后验分布时的表达能力。此外，第4章从理论角度严格证明了所提出算法的收敛性，为隐变量模型训练算法的构建提供了坚实的理论保障和实践依据。实验结果表明，所提出的方法在隐变量推断准确性、模型收敛性及下游软测量建模精度等方面均体现了显著的优越性。

（3） **基于泛函优化的限制域概率隐变量模型建模方法：**针对隐变量分布支撑在约束域的稳态概率隐变量模型（如图神经网络和变压器网络）建模问题，第5章从泛函优化理论视角出发，对第4章的概率隐变量模型训练框架进行了理论拓展。第5章首先剖析了第4章的方法在约束域场景下失效的根本原因——即迭代过程对欧几里得空间的隐式依赖。为突破这一局限，第5章引入了布雷格曼散度与镜像下降策略，构建了适用于约束域的隐变量分布推断理论框架。在此基础上，以图神经网络为典型应用场景，本章在 RKHS 的约束下，提出了一种全新的图结构推断算法，并从理论上严格证明了该算法的收敛性和稳定性。最后，第5章通过隐变量推断准确性、推理效率、下游软测量建模精度及收敛性等多维度的系统性实验验证，全面证明了该方法在限制域建模中的有效性和优越性。

（4） **动态概率隐变量模型的构建方法：**在第6章，为应对非线性动态概率隐变量模型在隐空间推断和深度学习后端实现中的关键挑战，第6章基于泛函优化理论，提出了一种全新的非线性动态概率隐变量模型建模框架。第6章首先分析了现有非线性动态概率隐变量模型在软测量建模任务中存在的两个主要问题：隐变量分布推断的不准确性和深度学习后端实现的困难性。针对这些问题，本章引入随机微分方程理论，将非线性动态概率隐变量模型的学习问题重构为最优控制问题，并通过交替方向乘子法对该问题进行求解，提出了一种以最优控制策略为核心的隐空间推断方法。在此基础上，第6章进一步设计了基于矩表达式的模型训练与推理算法，构建了“最优控制-非线性动态概率隐变量模型”框架，并严格证明了其在深度学

习后端实现过程中的收敛性和稳定性。最后，通过软测量建模精度、模型敏感性分析等多个层面的实验验证，充分证明了所提出方法的有效性和优越性。

综上所述，本文围绕隐变量模型驱动的软测量建模系统，从数据预处理到模型构建的完整链路，基于泛函优化方法进行了系统性重构，为实际工业应用提供了切实可行的解决方案。研究内容涵盖了隐变量驱动的数据补全方法、稳态概率隐变量模型、限制域隐变量模型以及动态概率隐变量模型的构建，不仅从理论上拓展了隐变量模型的设计与推断框架，还通过实际工业场景的验证，证明了所提出方法的有效性与适用性。

7.2 研究工作展望

隐变量模型因其理论特性与工业过程软测量建模需求具有高度契合性，已成为从工业过程数据中提取有效信息的重要方法学工具。在智能制造与人工智能技术快速发展的背景下，本研究运用泛函优化理论，系统性地重构了概率隐变量模型的建模框架，并通过工业过程软测量应用验证了所提方法的有效性。基于当前数据驱动过程建模领域的发展趋势与实际应用需求，本研究在理论创新和工程应用两个维度仍存在广阔的拓展空间，具体体现在以下几个方面：

（1） **规避 RKHS 的限制：**在最优控制策略和微扰方向的推导中，为避免显式解决优化问题所带来的高昂计算开销，本文将待优化的函数类限制在 RKHS 中。这种正则化方法虽然降低了问题的复杂度，却也可能对最优控制策略和微扰方向施加不必要的约束，尤其在高维场景下或将限制隐变量推断的精度。此外，随着数据集规模的扩大，核函数的计算复杂度往往呈二次增长，进而增加了实际应用中的计算负担。因此，探索在 $\mathscr{P}_2(\mathrm{D})$ 空间中利用其他类型的正则化手段[216]以替代 RKHS 范数的正则是一条颇具潜力的研究方向。

（2） **高阶泛函优化方法：**本文的大部分推导与实现都基于一阶优化方法展开，所得到的迭代策略虽具有较好的可解释性和易实现性，但其性能尚有提升空间。未来的研究可尝试将哈密顿流[217-219]（Hamiltonian flow）等高阶优化思路引入到该方法当中，将问题扩展至动量空间进行迭代，从而提高模型的收敛效率、提供更好的局部泛函极值逃逸能力，进而提升隐变量推断结果的精确度。

（3） **沃瑟斯坦空间的假设：**本文在推导概率密度函数演化过程时，采用了沃瑟斯坦空间的几何框架，对应概率测度的“传输”[121]（transportation）过程。未来研究可考虑在费舍尔-饶（Fisher-Rao）流形上构建概率密度函数的演化路径[220-221]，实现概率测度的“传送”[222]（teleportation）。这一理论扩展有望增强模型处理特征空间不均衡样本的能力，从而更加适应工业过程数据中由于操作模式发生突变而产生的离群值，从而提升概率隐变量模型在实际应用场景中的鲁棒性和适应性。

（4） **泛函优化方法在“半监督学习”类隐变量模型中的应用：**由于工业过程中往往存在大量无标签数据或标签稀缺的情况[12,223]，如何有效利用这些无标签数据对模型进行训练，是软测量建模领域亟待解决的关键问题。本文所提出的基于泛函优化的隐变量模型建模方法主要针对监督学习场景，在半监督学习环境下的扩展与应用仍有待探索。未来研究可从两个方向进行拓展：一方面，可结合迁移学习技术，将已训练模型的知识迁移至标签受限的新场景[224]，充分利用有限标签数据的价值；另一方面，可重新设计隐变量模型概率图结构和训练流程[88]，使其能够同时从有标签和无标签数据中学习有效表示，从而提高模型在标签稀缺环境中的性能。

（5） **泛函优化方法在“冷启动”场景中的应用：**工业过程软测量建模中常面临装置“冷启动”（cold start）的挑战[225]。在装置初始运行阶段，由于可收集的数据量有限，构建精确的软测量模型往往十分困难。如何有效利用其他相关装置的现有数据来辅助目标装置的模型构建，是一个具有重要实用价值的研究课题。值得注意的是，在推荐系统领域，泛函优化方法已在解决“冷启动”问题上取得了显著成效[226-227]。因此，探索将这类泛函优化策略迁移至工业场景，构建适用于“冷启动”阶段场景的软测量模型是一个值得思考的问题。

上述研究展望不仅为泛函优化方法在概率隐变量建模理论及工业过程软测量建模中的未来发展指明了方向，也提供了针对实际工业问题的可行解决方案。随着相关技术的不断优化与创新，泛函优化方法有望在更广泛的流程工业数据驱动建模场景中展现出更强大的适应性和应用价值，为复杂工业系统的智能化发展提供新的动力。

参考文献

[1] AMS N. Strategic Plan for the Manufacturing USA Program[J]. 2024.
[2] 赵璐, 宋大伟, 张凤, 等. 欧盟 “工业 5.0” 对我国制造业高质量发展的影响与启示——基于智库双螺旋法的应用探索研究[J]. 中国科学院院刊, 2022, 37(6): 756-764.
[3] 柴天佑, 钱锋, 管晓宏, 等. 面向双碳目标的自动化和智能化理论与技术[J]. 中国科学基金, 2024, 38(4): 560-570.
[4] 金涌, 胡山鹰, 张强. 2060 中国碳中和[M]. 化学工业出版社, 2022.
[5] 柴天佑. 工业智能助力实现 “双碳” 目标[J]. 中国科学基金, 2024, 38(4): 559.
[6] LUO Y, ZHANG X, KANO M, et al. Data-driven Soft Sensors in Blast Furnace Ironmaking: A Survey[J]. Front. Inf. Technol. Electron. Eng., 2023, 24(3): 327-354.
[7] 潘冬, 许川, 龚芃旭, 等. 基于红外与可见光视觉的高炉铁口铁水温度场在线检测[J]. 自动化学报, 2024, 51: 1-13.
[8] YANG Z, MAO L, YE L, et al. AKGNN: When Adaptive Graph Neural Network Meets Kolmogorov-Arnold Network for Industrial Soft Sensors[J]. IEEE Trans. Instrum. Meas., 2025.
[9] YUAN X, HUANG B, WANG Y, et al. Deep Learning-Based Feature Representation and Its Application for Soft Sensor Modeling With Variable-Wise Weighted SAE[J]. IEEE Trans. Ind. Inform., 2018, 14(7): 3235-3243.
[10] HAN Y, LIU X, GUO C, et al. Improved Pearson Correlation Coefficient-Based Graph Neural Network for Dynamic Soft Sensor of Polypropylene Industries[J]. Ind. Eng. Chem. Res., 2024, 64(1): 551-563.
[11] ZHANG M, LIU X, ZHANG Z. A soft sensor for industrial melt index prediction based on evolutionary extreme learning machine[J]. Chin. J. Chem. Eng., 2016, 24(8): 1013-1019.
[12] GE Z. Process Data Analytics via Probabilistic Latent Variable Models: A Tutorial Review[J]. Ind. Eng. Chem. Res., 2018, 57(38): 12646-12661.
[13] SHEN B, YAO L, GE Z. Nonlinear probabilistic latent variable regression models for soft sensor application: From shallow to deep structure[J]. Control Eng. Pract., 2020, 94: 104198.
[14] KONG X, JIANG X, ZHANG B, et al. Latent Variable Models in the Era of Industrial Big Data: Extension and Beyond[J]. Annu. Rev. Control, 2022, 54: 167-199.
[15] SUN Q, GE Z. A Survey on Deep Learning for Data-Driven Soft Sensors[J]. IEEE Trans. Ind. Inform., 2021, 17(9): 5853-5866.
[16] 柴天佑. 大数据与制造流程知识自动化发展战略研究[M]. 北京, 中国: 科学出版社, 2019.
[17] CHEN H, HUANG B. Fault-Tolerant Soft Sensors for Dynamic Systems[J]. IEEE Trans. Control Syst. Technol., 2023, 31(6): 2805-2818.
[18] JOSEPH B, BROSILOW C B. Inferential Control of Processes: Part I. Steady State Analysis and Design[J]. AIChE J., 1978, 24(3): 485-492.
[19] BROSILOW C, TONG M. Inferential Control of Processes: Part II. The Structure and Dynamics of Inferential Control Systems[J]. AIChE J., 1978, 24(3): 492-500.
[20] JOSEPH B, BROSILOW C. Inferential Control of Processes: Part III. Construction of Optimal and Suboptimal Dynamic Estimators[J]. AIChE J., 1978, 24(3): 500-509.
[21] 曾令权. 基于贝叶斯网络的工业过程建模与关键性能指标预测[D]. 杭州, 中国: 浙江大学, 2024.
[22] AL-MALAH K I. Aspen Plus: Chemical Engineering Applications[M]. New York, USA: John Wiley & Sons, 2022.
[23] WINKEL M, ZULLO L, VERHEIJEN P, et al. Modelling and simulation of the operation of an industrial batch plant using gPROMS[J]. Comput. Chem. Eng., 1995, 19: 571-576.
[24] CHEN Y, et al. System Simulation Techniques with MATLAB and Simulink[M]. New York, USA: John Wiley & Sons, 2013.
[25] BUSSIECK M R, MEERAUS A. General Algebraic Modeling System (GAMS)[M]. Berlin, Germany: Springer, 2004: 137-157.
[26] LI M, ALLAN D A, DINH S, et al. Nonlinear Model Predictive Control for Mode-Switching Operation of Reversible Solid Oxide Cell Systems[J]. AIChE J., 2024, 70(11): e18550.

[27] CHANDESRIS M, MÉDEAU V, GUILLET N, et al. Membrane Degradation in PEM Water Electrolyzer: Numerical Modeling and Experimental Evidence of the Influence of Temperature and Current Density[J]. Int. J. Hydrog. Energy, 2015, 40(3): 1353-1366.
[28] HAENLEIN M, KAPLAN A M. A beginner's guide to partial least squares analysis[J]. Understanding statistics, 2004, 3(4): 283-297.
[29] 孔祥印. 基于轻量级深度隐变量模型的工业过程监测[D]. 杭州, 中国: 浙江大学, 2024.
[30] VAPNIK V N. The Support Vector Method[C]//Proc. Int. Jt. Conf. Neural Netw. 1997: 261-271.
[31] HUA L, ZHANG C, SUN W, et al. An evolutionary deep learning soft sensor model based on random forest feature selection technique for penicillin fermentation process[J]. ISA Trans., 2023, 136: 139-151.
[32] WAN Y, LIU D, REN J C. A modeling method of wide random forest multi-output soft sensor with attention mechanism for quality prediction of complex industrial processes[J]. Adv. Eng. Inform., 2024, 59: 102255.
[33] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. science, 2006, 313(5786): 504-507.
[34] SUN Q, GE Z. Gated Stacked Target-Related Autoencoder: A Novel Deep Feature Extraction and Layerwise Ensemble Method for Industrial Soft Sensor Application[J]. IEEE Trans. Cybern., 2022, 52(5): 3457-3468.
[35] LECUN Y, KAVUKCUOGLU K, FARABET C. Convolutional networks and applications in vision [C]//Proc. IEEE. 2010: 253-256.
[36] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Comput., 1997, 9(8): 1735-1780.
[37] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 30. 2017.
[38] YUAN X, LI L, WANG Y. Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network[J]. IEEE Trans. Ind. Inform., 2020, 16(5): 3168-3176.
[39] YUAN X, QI S, WANG Y, et al. Quality Variable Prediction for Nonlinear Dynamic Industrial Processes Based on Temporal Convolutional Networks[J]. IEEE Sens. J., 2021, 21(18): 20493-20503.
[40] PULI V K, CHIPLUNKAR R, HUANG B. Physics-Informed Probabilistic Slow Feature Analysis [J]. Automatica, 2024, 169: 111851.
[41] LOU S, YANG C, ZHANG X, et al. Data/Mechanism Hybrid-Driven Modeling of Blast Furnace Smelting System and Global Sequential Optimization[J]. J. Process Control, 2024, 139: 103235.
[42] BISHOP C M, NASRABADI N M. Pattern Recognition and Machine Learning: vol. 4[M]. Berlin: Springer, 2006.
[43] BARBER D. Bayesian Reasoning and Machine Learning[M]. London, UK: Cambridge University Press, 2012.
[44] BISHOP C M. Latent variable models[G]//Learning in graphical models. Springer, 1998: 371-403.
[45] KONG X, GE Z. Deep PLS: A Lightweight Deep Learning Model for Interpretable and Efficient Data Analytics[J]. IEEE Trans. Neural Netw. Learn. Syst., 2023, 34(11): 8923-8937.
[46] KONG X, HE Y, SONG Z, et al. Deep Probabilistic Principal Component Analysis for Process Monitoring[J]. IEEE Trans. Neural Netw. Learn. Syst., 2025, 36(4): 7422-7436.
[47] ZHU J, GE Z, SONG Z, et al. Review and Big Data Perspectives on Robust Data Mining Approaches for Industrial Process Modeling with Outliers and Missing Data[J]. Annu. Rev. Control, 2018, 46: 107-133.
[48] TIPPING M E, BISHOP C M. Probabilistic Principal Component Analysis[J]. J. R. Stat. Soc., B: Stat. Methodol., 1999, 61(3): 611-622.
[49] YANG Z, YAO L, GE Z. Streaming Parallel Variational Bayesian Supervised Factor Analysis for Adaptive Soft Sensor Modeling with Big Process Data[J]. J. Process Control, 2020, 85: 52-64.
[50] TOMCZAK J M. Deep Generative Modeling: From Probabilistic Framework to Generative AI[Z]. 2025.
[51] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes[C]//Proc. Int. Conf. Learn. Represent. 2014.
[52] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative Adversarial Networks: An Overview

[J]. IEEE Signal Process. Mag., 2018, 35(1): 53-65.
[53] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Commun. ACM, 2020, 63(11): 139-144.
[54] GANGULY A, JAIN S, WATCHAREERUETAI U. Amortized Variational Inference: A Systematic Review[J]. J. Artif. Intell. Res., 2023, 78: 167-215.
[55] MARGOSSIAN C C, BLEI D M. Amortized Variational Inference: When and Why?[C]//Proc. Conf. Uncertainty in Artificial Intelligence. 2024: 2434-2449.
[56] GE Z. Supervised Latent Factor Analysis for Process Data Regression Modeling and Soft Sensor Application[J]. IEEE Trans. Control Syst. Technol., 2016, 24(3): 1004-1011.
[57] BEAL M J. Variational Algorithms for Approximate Bayesian Inference[D]. University of London, University College London (United Kingdom), 2003.
[58] GE Z, CHEN X. Dynamic Probabilistic Latent Variable Model for Process Data Modeling and Regression Application[J]. IEEE Trans. Control Syst. Technol., 2019, 27(1): 323-331.
[59] YUAN X, GE Z, HUANG B, et al. A Probabilistic Just-in-Time Learning Framework for Soft Sensor Development With Missing Data[J]. IEEE Trans. Control Syst. Technol., 2017, 25(3): 1124-1132.
[60] YUAN X, WANG Y, YANG C, et al. Weighted Linear Dynamic System for Feature Representation and Soft Sensor Application in Nonlinear Dynamic Industrial Processes[J]. IEEE Trans. Ind. Electron., 2018, 65(2): 1508-1517.
[61] SHAO W, XIAO C, WANG J, et al. Real-time estimation of quality-related variable for dynamic and non-Gaussian process based on semisupervised Bayesian HMM[J]. J. Process Control, 2022, 111: 59-74.
[62] SHAO W, GE Z, SONG Z. Bayesian Just-in-Time Learning and Its Application to Industrial Soft Sensing[J]. IEEE Trans. Ind. Inform., 2020, 16(4): 2787-2798.
[63] YANG Z, ZHENG J, GE Z. Lifelong Bayesian Learning Machines for Streaming Industrial Big Data [J]. IEEE Trans. Syst. Man Cybern.: Syst., 2023, 53(3): 1554-1565.
[64] WEI R, MAHMOOD A. Recent Advances in Variational Autoencoders with Representation Learning for Biomedical Informatics: A Survey[J]. IEEE Access, 2020, 9: 4939-4956.
[65] GIRIN L, LEGLAIVE S, BIE X, et al. Dynamical Variational Autoencoders: A Comprehensive Review[M]. Now Foundations, 2021.
[66] CHAI Z, ZHAO C, HUANG B, et al. A Deep Probabilistic Transfer Learning Framework for Soft Sensor Modeling With Missing Data[J]. IEEE Trans. Neural Netw. Learn. Syst., 2022, 33(12): 7598-7609.
[67] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational Inference: A Review for Statisticians [J]. Journal of the American statistical Association, 2017, 112(518): 859-877.
[68] ZHANG C, BÜTEPAGE J, KJELLSTRÖM H, et al. Advances in Variational Inference[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2019, 41(8): 2008-2026.
[69] 郑煌杰. 变分自编码机 VAE 的信息论理解及其应用[D]. 上海, 中国: 上海交通大学, 2019.
[70] SUGIYAMA M, SUZUKI T, KANAMORI T. Density Ratio Estimation in Machine Learning[M]. London: Cambridge University Press, 2012.
[71] YU S, YU K, TRESP V, et al. Supervised Probabilistic Principal Component Analysis[C]//Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2006: 464-473.
[72] GE Z, GAO F, SONG Z. Mixture Probabilistic PCR Model for Soft Sensing of Multimode Processes [J]. Chemom. Intell. Lab. Syst., 2011, 105(1): 91-105.
[73] GE Z, HUANG B, SONG Z. Mixture Semisupervised Principal Component Regression Model and Soft Sensor Application[J]. AIChE J., 2014, 60(2): 533-545.
[74] GE Z. Mixture Bayesian Regularization of PCR Model and Soft Sensing Application[J]. IEEE Trans. Ind. Electron., 2015, 62(7): 4336-4343.
[75] GHAHRAMANI Z, ROWEIS S. Learning Nonlinear Dynamical Systems Using an EM Algorithm [C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 11. 1998: 1-7.
[76] ZHU J, GE Z, SONG Z. Robust Semi-Supervised Mixture Probabilistic Principal Component Regression Model Development and Application to Soft Sensors[J]. J. Process Control, 2015, 32: 25-37.
[77] GHAHRAMANI Z, BEAL M. Variational Inference for Bayesian Mixtures of Factor Analysers[C] //Proc. Adv. Neural Inf. Process. Syst. Vol. 12. 1999: 1-7.

[78] MA Y, HUANG B. Extracting Dynamic Features with Switching Models for Process Data Analytics and Application in Soft Sensing[J]. AIChE J., 2018, 64(6): 2037-2051.
[79] GHAHRAMANI Z, HINTON G E. Variational Learning for Switching State-Space Models[J]. Neural Comput., 2000, 12(4): 831-864.
[80] BRODERICK T, BOYD N, WIBISONO A, et al. Streaming Variational Bayes[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 26. 2013.
[81] YANG Z, GE Z. Industrial Virtual Sensing for Big Process Data Based on Parallelized Nonlinear Variational Bayesian Factor Regression[J]. IEEE Trans. Instrum. Meas., 2020, 69(10): 8128-8136.
[82] ZHENG J, LIU Y, LIU Y, et al. Semi-Supervised Process Data Regression and Application Based on Latent Factor Analysis Model[J]. IEEE Trans. Instrum. Meas., 2023, 72: 1-11.
[83] 杨泽宇. 基于不同学习范式的工业大数据建模与质量预报[D]. 杭州, 中国: 浙江大学, 2021.
[84] YAO L, GE Z. Scalable semisupervised GMM for big data quality prediction in multimode processes [J]. IEEE Trans. Ind. Electron., 2018, 66(5): 3681-3692.
[85] 李崇轩. 面向多种学习任务的深度生成模型[D]. 北京, 中国: 清华大学, 2019.
[86] 路橙. 可逆生成模型及其高效算法研究[D]. 北京, 中国: 清华大学, 2023.
[87] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C]//. 2015: 1-8.
[88] XIE R, JAN N M, HAO K, et al. Supervised variational autoencoders for soft sensor modeling with missing data[J]. IEEE Trans. Ind. Inform., 2019, 16(4): 2820-2828.
[89] 崔琳琳, 沈冰冰, 葛志强. 基于混合变分自编码器回归模型的软测量建模方法[J]. 自动化学报, 2022, 48: 398.
[90] GUO F, WEI B, HUANG B. A just-in-time modeling approach for multimode soft sensor based on Gaussian mixture variational autoencoder[J]. Comput. Chem. Eng., 2021, 146: 107230.
[91] SHEN B, GE Z. Supervised Nonlinear Dynamic System for Soft Sensor Application Aided by Variational Auto-Encoder[J]. IEEE Trans. Instrum. Meas., 2020, 69(9): 6132-6142.
[92] SHEN B, GE Z. Weighted nonlinear dynamic system for deep extraction of nonlinear dynamic latent variables and industrial application[J]. IEEE Trans. Ind. Inform., 2020, 17(5): 3090-3098.
[93] SHEN B, YAO L, GE Z. Predictive Modeling With Multiresolution Pyramid VAE and Industrial Soft Sensor Applications[J]. IEEE Trans. Cybern., 2023, 53(8): 4867-4879.
[94] SHEN B, JIANG X, YAO L, et al. Gaussian mixture TimeVAE for industrial soft sensing with deep time series decomposition and generation[J]. J. Process Control, 2025, 147: 103355.
[95] CHO K, van MERRIËNBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C]//MOSCHITTI A, PANG B, DAELEMANS W. Conf. Empir. Methods Nat. Lang. Process.: Syst. Demonstr. Proc. Doha, Qatar, 2014: 1724-1734.
[96] LU Y, PENG X, YANG D, et al. The Probabilistic Discriminative Time-series Model with Latent Variables and Its Application to Industrial chemical process modeling[J]. Chem. Eng. J., 2021, 423: 130298.
[97] JIANG C, LU Y, ZHONG W, et al. Deep Bayesian Slow Feature Extraction with Application to Industrial Inferential Modeling[J]. IEEE Trans. Ind. Inform., 2023, 19(1): 40-51.
[98] JIANG C, PENG X, HUANG B, et al. Switching probabilistic slow feature extraction for semisupervised industrial inferential modeling[J]. J. Process Control, 2024, 141: 103277.
[99] YAO L, SHEN B, CUI L, et al. Semi-Supervised Deep Dynamic Probabilistic Latent Variable Model for Multimode Process Soft Sensor Application[J]. IEEE Trans. Ind. Inform., 2023, 19(4): 6056-6068.
[100] WOLPERT D, MACREADY W. No Free Lunch Theorems for Optimization[J]. IEEE Trans. Evol. Comput., 1997, 1(1): 67-82.
[101] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[C]//Proc. Int. Conf. Mach. Learn. 2014: 1278-1286.
[102] FATTORINI H O. Infinite dimensional optimization and control theory: vol. 54[M]. London: Cambridge University Press, 1999.
[103] 老大中. 变分法基础（第三版）[M]. 北京: 国防工业出版社, 2015.
[104] MARSDEN J E, RATIU T S, HERMANN R. Introduction to mechanics and symmetry[J]. SIAM Rev., 1997, 39(1): 152-152.
[105] 傅献彩. 物理化学（第五版）[M]. 北京: 高等教育出版社, 2005.

[106] KANTOROVICH L V. Mathematical methods of organizing and planning production[J]. Manag. Sci., 1960, 6(4): 366-422.
[107] PONTRYAGIN L S. Mathematical Theory of Optimal Processes[M]. CRC press, 1987.
[108] BELLMAN R. Dynamic programming[J]. science, 1966, 153(3731): 34-37.
[109] BIEGLER L T. Nonlinear programming: concepts, algorithms, and applications to chemical processes[M]. Philadelphia: SIAM, 2010.
[110] MURPHY K P. Probabilistic Machine Learning: An Introduction[M]. MIT press, 2022.
[111] Van ERVEN T, HARREMOS P. Rényi Divergence and Kullback-Leibler Divergence[J]. IEEE Trans. Inf. Theory, 2014, 60(7): 3797-3820.
[112] CHUNG J, KASTNER K, DINH L, et al. A Recurrent Latent Variable Model for Sequential Data[C]//CORTES C, LAWRENCE N, LEE D, et al. Proc. Adv. Neural Inf. Process. Syst. Vol. 28. Curran Associates, Inc., 2015: 1-9.
[113] CALUYA K F, HALDER A. Proximal Recursion for Solving the Fokker-Planck Equation[C]// Proc. Am. Control Conf. 2019: 4098-4103.
[114] CALUYA K F, HALDER A. Wasserstein Proximal Algorithms for the Schrödinger Bridge Problem: Density Control With Nonlinear Drift[J]. IEEE Trans. Autom. Control., 2022, 67(3): 1163-1178.
[115] 陈键飞. 表示学习的高效算法[D]. 北京, 中国: 清华大学, 2019.
[116] SASON I, VERDÚ S. f -Divergence Inequalities[J]. IEEE Trans. Inf. Theory., 2016, 62(11): 5973-6006.
[117] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[C]//Proc. Int. Conf. Mach. Learn. 2017: 214-223.
[118] TONG A, FATRAS K, MALKIN N, et al. Improving and generalizing flow-based generative models with minibatch optimal transport[J]. Trans. Mach. Learn. Res., 1-34.
[119] BUNNE C, STARK S G, GUT G, et al. Learning Single-cell Perturbation Responses Using Neural Optimal Transport[J]. Nat. Methods, 2023, 20(11): 1759-1768.
[120] BUNNE C, SCHIEBINGER G, KRAUSE A, et al. Optimal transport for single-cell and spatial omics [J]. Nat Rev Methods Primers., 2024, 4(1): 58.
[121] VILLANI C, et al. Optimal Transport: Old and New: vol. 338[M]. Springer, 2009.
[122] 胡寿松, 王执铨, 胡维礼. 最优控制理论与系统（第二版）[M]. 北京: 科学出版社, 2005.
[123] FORTUNA L, GRAZIANI S, XIBILIA M. Soft Sensors for Product Quality Monitoring in Debutanizer Distillation Columns[J]. Control Eng. Pract., 2005, 13(4): 499-508.
[124] FORTUNA L, GRAZIANI S, RIZZO A, et al. Soft sensors for monitoring and control of industrial processes: vol. 22[M]. Springer, 2007.
[125] FOGLER H S. Essentials of Chemical Reaction Engineering: Essenti Chemica Reaction Engineering [M]. New Jersey, USA: Pearson Education, 2010.
[126] WU P, PEI M, WANG T, et al. A Low-Rank Bayesian Temporal Matrix Factorization for the Transfer Time Prediction Between Metro and Bus Systems[J]. IEEE Trans. Intell. Transp. Syst., 2024, 25(7): 7206-7222.
[127] ZHU J, GE Z, SONG Z, et al. Large-scale Plant-wide Process Modeling and Hierarchical Monitoring: A Distributed Bayesian Network Approach[J]. J. Process Control, 2018, 65: 91-106.
[128] 姚邹静, 赵春晖, 李元龙, 等. 面向工业软测量应用的定制化生成对抗数据填补模型[J]. 控制与决策, 2021, 36(12): 2929-2936.
[129] LIU D, WANG Y, LIU C, et al. Blackout Missing Data Recovery in Industrial Time Series Based on Masked-Former Hierarchical Imputation Framework[J]. IEEE Trans. Autom. Sci. Eng., 2024, 21(2): 1138-1150.
[130] LIU D, WANG Y, LIU C, et al. Scope-free Global Multi-Condition-Aware Industrial Missing Data Imputation Framework via Diffusion Transformer[J]. IEEE Trans. Knowl. Data Eng., 2024, 36(11): 6977-6988.
[131] YUAN X, ZHANG J, WANG K, et al. Missing Data Imputation for Industrial Time Series With Adaptive Median Iteration Based on Generative Adversarial Networks[J]. IEEE Sens. J., 2024, 24(21): 35081-35091.
[132] RUBIN D B. Inference and Missing Data[J]. Biometrika, 1976, 63(3): 581-592.
[133] JARRETT D, CEBERE B C, LIU T, et al. Hyperimpute: Generalized iterative imputation with auto-

matic model selection[C]//Proc. Int. Conf. Mach. Learn. 2022: 9916-9937.
[134] MUZELLEC B, JOSSE J, BOYER C, et al. Missing Data Imputation Using Optimal Transport[C]//Proc. Int. Conf. Mach. Learn. 2020: 7130-7140.
[135] WANG H, CHEN Z, LIU Z, et al. Entire Space Counterfactual Learning for Reliable Content Recommendations[J]. IEEE Trans. Inf. Forensics Secur., 2025, 20: 1755-1764.
[136] Chapter 3 Necessary Conditions for Singular Optimal Control[G]//BELL D J, JACOBSON D H. Mathematics in Science and Engineering: Singular Optimal Control Problems: vol. 117. Elsevier, 1975: 61-100.
[137] SANTAMBROGIO F. {Euclidean, metric, and Wasserstein} gradient flows: an overview[J]. Bull. Math. Sci., 2017, 7: 87-154.
[138] PARK M S, KIM C, SON H, et al. The deep minimizing movement scheme[J]. J. Comput. Phys., 2023, 494: 112518.
[139] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 32. 2019: 1-12.
[140] BRADBURY J, FROSTIG R, HAWKINS P, et al. JAX: Autograd and XLA[J]. Astrophysics Source Code Library, 2021: ascl-2111.
[141] LIU Q, WANG D. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 29. 2016: 1-13.
[142] LIU Q. Stein Variational Gradient Descent as Gradient Flow[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 30. 2017: 1-15.
[143] HYVÄRINEN A. Estimation of Non-Normalized Statistical Models by Score Matching[J]. J. Mach. Learn. Res., 2005, 6(24): 695-709.
[144] VINCENT P. A Connection Between Score Matching and Denoising Autoencoders[J]. Neural Comput., 2011, 23(7): 1661-1674.
[145] FOLLAND G B. Real Analysis: Modern Techniques and Their Applications: vol. 40[M]. New York, USA: John Wiley & Sons, 1999.
[146] TASHIRO Y, SONG J, SONG Y, et al. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 34. 2021: 24804-24816.
[147] ZHENG S, CHAROENPHAKDEE N. Diffusion Models for Missing Value Imputation in Tabular Data[C]//Proc. Adv. Neural Inf. Process. Syst. Workshop on First Table Representation. 2022.
[148] OUYANG Y, XIE L, LI C, et al. MissDiff: Training Diffusion Models on Tabular Data with Missing Values[C]//Proc. Int. Conf. Mach. Learn. Workshop on Structured Probabilistic Inference& Generative Modeling. 2023.
[149] MATTEI P A, FRELLSEN J. MIWAE: Deep Generative Modelling and Imputation of Incomplete Datasets[C]//Proc. Int. Conf. Mach. Learn. 2019: 4413-4423.
[150] KYONO T, ZHANG Y, BELLOT A, et al. MIRACLE: Causally-aware imputation via learning missing data mechanisms[J]. Proc. Adv. Neural Inf. Process. Syst., 2021: 23806-23817.
[151] ZHAO H, SUN K, DEZFOULI A, et al. Transformed Distribution Matching for Missing Value Imputation[C]//Proc. Int. Conf. Mach. Learn. 2023: 42159-42186.
[152] WANG H, LIU X, LIU Z, et al. LSPT-D: Local Similarity Preserved Transport for Direct Industrial Data Imputation[J]. IEEE Trans. Autom. Sci. Eng., 2025, 22: 9438-9448.
[153] ZHU Q. Latent Variable Regression for Supervised Modeling and Monitoring[J]. IEEE/CAA J. Autom. Sin., 2020, 7(3): 800-811.
[154] YU W, WU M, HUANG B, et al. A Generalized Probabilistic Monitoring Model with Both Random and Sequential data[J]. Automatica, 2022, 144: 110468.
[155] RAVEENDRAN R, KODAMANA H, HUANG B. Process Monitoring Using a Generalized Probabilistic Linear Latent Variable Model[J]. Automatica, 2018, 96: 73-83.
[156] TANG B, MATTESON D S. Probabilistic Transformer for Time Series Analysis[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 34. 2021: 23592-23608.
[157] KNOBLAUCH J, JEWSON J, DAMOULAS T. An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference[J]. J. Mach. Learn. Res., 2022, 23(132): 1-109.
[158] ELDESOUKEY A, MIANGOLARRA O M, GEORGIOU T T. An Excursion Onto Schrödinger's

Bridges: Stochastic Flows With Spatio-Temporal Marginals[J]. IEEE Control Syst. Lett., 2024, 8: 1138-1143.
[159] CHEN X, DAI H, SONG L. Particle Flow Bayes' Rule[C]//Proc. Int. Conf. Mach. Learn. 2019: 1022-1031.
[160] CHEN Y, GEORGIOU T T, PAVON M. Optimal Transport Over a Linear Dynamical System[J]. IEEE Trans. Autom. Control., 2017, 62(5): 2137-2152.
[161] CHEN Y, GEORGIOU T T, PAVON M. Optimal Steering of a Linear Stochastic System to a Final Probability Distribution, Part I[J]. IEEE Trans. Autom. Control., 2016, 61(5): 1158-1169.
[162] CHEN Y, GEORGIOU T T, PAVON M. Optimal Steering of a Linear Stochastic System to a Final Probability Distribution, Part II[J]. IEEE Trans. Autom. Control., 2016, 61(5): 1170-1180.
[163] CHEN Y, GEORGIOU T T, PAVON M. On the Relation Between Optimal Transport and Schrödinger Bridges: A Stochastic Control Viewpoint[J]. J. Optim. Theory Appl., 2016, 169: 671-691.
[164] YANG L, ZHANG Z, SONG Y, et al. Diffusion Models: A Comprehensive Survey of Methods and Applications[J]. ACM Comput. Surv., 2023, 56(4): 1-39.
[165] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-Based Generative Modeling through Stochastic Differential Equations[C]//Proc. Int. Conf. Learn. Represent. 2020: 1-36.
[166] ZHANG Q, CHEN Y. Path Integral Sampler: A Stochastic Control Approach For Sampling[C]// International Conference on Learning Representations. 2021: 1-13.
[167] TAGHVAEI A, MEHTA P G. A Survey of Feedback Particle Filter and Related Controlled Interacting Particle Systems[J]. Annu. Rev. Control., 2023, 55: 356-378.
[168] HO J, JAIN A, ABBEEL P. Denoising Diffusion Probabilistic Models[J]. Proc. Adv. Neural Inf. Process. Syst., 2020, 33: 6840-6851.
[169] CHEN Y, DENG W, FANG S, et al. Provably Convergent Schrödinger Bridge with Applications to Probabilistic Time-series Imputation[C]//Proc. Int. Conf. Mach. Learn. 2023: 4485-4513.
[170] 胡寿松. 自动控制原理（第三版）[M]. 北京: 科学出版社, 2019.
[171] LEWIS F L, VRABIE D, SYRMOS V L. Optimal Control[M]. New York, USA: John Wiley & Sons, 2012.
[172] EVANS L C. Partial differential equations: vol. 19[M]. Providence, RI: American Mathematical Society, 2022.
[173] 同济大学数学系. 高等数学（第六版）[M]. 北京: 高等教育出版社, 2007.
[174] PFAU D, SPENCER J, de G. MATTHEWS A, et al. Ab-Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks[J]. Phys. Rev. Res., 2020, 2: 033429.
[175] BUTCHER J C. Numerical Methods for Ordinary Differential Equations[M]. New York, USA: John Wiley & Sons, 2016.
[176] LIU Q, LEE J, JORDAN M. A Kernelized Stein Discrepancy for Goodness-of-fit Tests[C]//Proc. Int. Conf. Mach. Learn. 2016: 276-284.
[177] 刘畅. 利用流形结构的高效贝叶斯推理方法研究[D]. 北京, 中国: 清华大学, 2019.
[178] MA Y. Extracting Dynamic Latent Feature with Bayesian Approaches for Process Data Analysis[D]. Edmonton, Canada: Alberta University, 2019.
[179] 崔琳琳. 基于变分自编码器的工业过程软测量建模研究[D]. 杭州, 中国: 浙江大学, 2022.
[180] 王静波. 基于贝叶斯混合模型的鲁棒软测量建模及应用[D]. 杭州, 中国: 浙江大学, 2022.
[181] MA Y, TANG J. Deep Learning on Graphs[M]. London: Cambridge University Press, 2020.
[182] ZHANG R, LIU Q, TONG X. Sampling in Constrained Domains with Orthogonal-Space Variational Gradient descent[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 35. 2022: 37108-37120.
[183] SHI J, LIU C, MACKEY L. Sampling with Mirrored Stein Operators[J]. Proc. Int. Conf. Learn. Represent., 2022: 1-26.
[184] YAO L, SHAO W, GE Z. Hierarchical Quality Monitoring for Large-scale Industrial Plants with Big Process Data[J]. IEEE Trans. Neural Netw. Learn. Syst., 2019, 32(8): 3330-3341.
[185] SHAO W, GE Z, SONG Z, et al. Semisupervised Robust Modeling of Multimode Industrial Processes for Quality Variable Prediction based on Student's t Mixture Model[J]. IEEE Trans. Ind. Inform., 2019, 16(5): 2965-2976.
[186] WANG J, SHAO W, ZHANG X, et al. Nonlinear Variational Bayesian Student'st Mixture Regression and Inferential Sensor Application with S emisupervised Data[J]. J. Process Control, 2021, 105: 141-

159.
[187] YANG Z, ZHENG J, GE Z. Lifelong Bayesian learning machines for streaming industrial big data [J]. IEEE Trans. Syst. Man Cybern.: Syst., 2022, 53(3): 1554-1565.
[188] LU C, ZENG J, DONG Y, et al. Streaming Variational Probabilistic Principal Component Analysis for Monitoring of Nonstationary Process[J]. J. Process Control, 2024, 133: 103134.
[189] WANG J, YAO L, XIONG W. Novel Semi-Supervised Deep Probabilistic Slow Feature Extraction for Online Chemical Process Soft Sensing Application[J]. IEEE Trans. Instrum. Meas., 2024, 73: 1-11.
[190] JANG E, GU S, POOLE B. Categorical Reparameterization with Gumbel-Softmax[C]//Proc. Int. Conf. Learn. Represent. 2017: 1-12.
[191] LE H, HØIER R K, LIN C T, et al. AdaSTE: An Adaptive Straight-Through Estimator to Train Binary Neural Networks[C]//Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2022: 460-469.
[192] MADDISON C J, TARLOW D, MINKA T. A* sampling[C]//Proc. Adv. Neural Inf. Process. Syst. Montreal, Canada: MIT Press, 2014: 3086-3094.
[193] KIPF T N, WELLING M. Semi-supervised Classification with Graph Convolutional Networks[C]//Proc. Int. Conf. Learn. Represent. 2016: 1-14.
[194] HAMILTON W L, YING R, LESKOVEC J. Inductive Representation Learning on Large Graphs[C]//Proc. Adv. Neural Inf. Process. Syst. 2017: 1025-1035.
[195] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks[C]//Proc. Int. Conf. Learn. Represent. 2018: 1-12.
[196] DAO T, FU D Y, ERMON S, et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness[C]//Proc. Conf. Neural Inf. Process. Syst. 2022: 1-16.
[197] LIU Y, HU T, ZHANG H, et al. iTransformer: Inverted Transformers are Effective for Time Series Forecasting[C]//Proc. Int. Conf. Learn. Represent. 2024: 1-22.
[198] TAY Y, BAHRI D, METZLER D, et al. Synthesizer: Rethinking Self-attention for Transformer Models[C]//Proc. Int. Conf. Mach. Learn. 2021: 10183-10192.
[199] ZHU K, ZHAO C. Dynamic Graph-Based Adaptive Learning for Online Industrial Soft Sensor With Mutable Spatial Coupling Relations[J]. IEEE Trans. Ind. Electron., 2023, 70(9): 9614-9622.
[200] WANG Y, SUI Q, LIU C, et al. Interpretable Prediction Modeling for Froth Flotation via Stacked Graph Convolutional Network[J]. IEEE Trans. Artif. Intell., 2024, 5(1): 334-345.
[201] BI F, HE T, XIE Y, et al. Two-Stream Graph Convolutional Network-Incorporated Latent Feature Analysis[J]. IEEE Trans. Serv. Comput., 2023, 16(4): 3027-3042.
[202] CHEN J, YUAN Y, LUO X. SDGNN: Symmetry-Preserving Dual-Stream Graph Neural Networks [J]. IEEE/CAA J. Autom. Sin., 2024, 11(7): 1717-1719.
[203] HUANG C, LI M, CAO F, et al. Are Graph Convolutional Networks With Random Weights Feasible? [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2023, 45(3): 2751-2768.
[204] SÄRKKÄ S, SOLIN A. Applied Stochastic Differential Equations: vol. 10[M]. London: Cambridge University Press, 2019.
[205] SÄRKKÄ S, SVENSSON L. Bayesian Filtering and Smoothing: vol. 17[M]. London: Cambridge university press, 2023.
[206] BINGHAM E, CHEN J P, JANKOWIAK M, et al. PyRO: Deep Universal Probabilistic Programming [J]. J. Mach. Learn. Res., 2019, 20(28): 1-6.
[207] WEN Q, ZHOU T, ZHANG C, et al. Transformers in Time Series: A Survey[C]//ELKIND E. Proc. Int. Joint Conf. Artif. Intell. International Joint Conferences on Artificial Intelligence Organization, 2023: 6778-6786.
[208] BILLINGSLEY P. Probability and Measure[M]. 3rd ed. New York, USA: John Wiley & Sons, 1995.
[209] CHEN Z, WANG H, CHEN G, et al. Analyzing and Improving Supervised Nonlinear Dynamical Probabilistic Latent Variable Model for Inferential Sensors[J]. IEEE Trans. Ind. Inform., 2024, 20(11): 13296-13307.
[210] BERTSEKAS D. A Course in Reinforcement Learning[M]. Nashua: Athena Scientific, 2023.
[211] CHEN Y C. A Tutorial on Kernel Density Estimation and Recent Advances[J]. Biostat. Epidemiol., 2017, 1(1): 161-187.

[212] RASMUSSEN C E. Gaussian Processes in Machine Learning[M]. 2003: 63-71.
[213] FENG L, ZHAO C, SUN Y. Dual Attention-Based Encoder–Decoder: A Customized Sequence-to-Sequence Learning for Soft Sensor Development[J]. IEEE Trans. Neural Netw. Learn. Syst., 2021, 32(8): 3306-3317.
[214] ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting[C]//Proc. AAAI Conf. Artif. Intell. Vol. 35: 12. 2021: 11106-11115.
[215] ANTONIAK C E. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems[J]. Ann. Stat., 1974: 1152-1174.
[216] DONG H, WANG X, YONG L, et al. Particle-based Variational Inference with Preconditioned Functional Gradient Flow[C]//Proc. Int. Conf. Learn. Represent. 2022: 1-26.
[217] WANG F, ZHU H, ZHANG C, et al. GAD-PVI: A General Accelerated Dynamic-Weight Particle-Based Variational Inference Framework[C]//Proc. AAAI Conf. Artif. Intell. 2024: 15466-15473.
[218] WANG Y, LI W. Accelerated Information Gradient Flow[J]. J. Sci. Comput., 2022, 90: 1-47.
[219] TAGHVAEI A, MEHTA P. Accelerated Flow for Probability Distributions[C]//Proc. Int. Conf. Mach. Learn. 2019: 6076-6085.
[220] NEKLYUDOV K, NYS J, THIEDE L, et al. Wasserstein Quantum Monte Carlo: A Novel Approach for Solving the Quantum Many-body Schrödinger Equation[J]. Proc. Adv. Neural Inf. Process. Syst., 2023, 36: 63461-63482.
[221] ZHANG C, LI Z, DU X, et al. DPVI: A Dynamic-Weight Particle-Based Variational Inference Framework[C]//Proc. Int. Joint Conf. Artif. Intell. 2022: 4900-4906.
[222] PHAM K, LE K, HO N, et al. On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm [C]//Proc. Int. Conf. Mach. Learn. 2020: 7673-7682.
[223] HE B, ZHANG X, SONG Z. Deep Learning of Partially Labeled Data for Quality Prediction Based on Stacked Target-Related Laplacian Autoencoder[J]. IEEE Trans. Neural Netw. Learn. Syst., 2024, 35(3): 2927-2941.
[224] YANG D, PENG X, JIANG C, et al. Transferable Deep Slow Feature Network With Target Feature Attention for Few-Shot Time-Series Prediction[J]. IEEE Trans. Ind. Inform., 2024, 20(5): 7292-7302.
[225] XIE J, HUANG B, DUBLJEVIC S. Transfer Learning for Dynamic Feature Extraction Using Variational Bayesian Inference[J]. IEEE Trans. Knowl. Data Eng., 2022, 34(11): 5524-5535.
[226] LIU W, SU J, CHEN C, et al. Leveraging Distribution Alignment via Stein Path for Cross-Domain Cold-Start Recommendation[C]//Proc. Adv. Neural Inf. Process. Syst. Vol. 34. 2021: 19223-19234.
[227] LIU W, ZHENG X, SU J, et al. Contrastive Proxy Kernel Stein Path Alignment for Cross-Domain Cold-Start Recommendation[J]. IEEE Trans. Knowl. Data Eng., 2023, 35(11): 11216-11230.
[228] FISHER R A. Statistical Methods for Research Workers[G]//Breakthroughs in statistics: Methodology and distribution. 1925.
[229] NEYMAN J, PEARSON E S. IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses [J]. Philos. Trans. R. Soc. A, 1933, 231(694-706): 289-337.
[230] HE K, ZHANG X, REN S, et al. Delving Deep Into Rectifiers: Surpassing Human-level Performance on Imagenet Classification[C]//Proc. IEEE Int. Conf. Comput. Vis. 2015: 1026-1034.

附录

A 假设性检验的流程

为了更好地理解实验结果中涉及的统计假设及其验证方法，本节将介绍本文采用的两种关键的统计检验方法：KSD 拟合优度检验方法和成对样本 t-检验方法。

A.1 基于 KSD 的拟合优度检验流程

KSD 的定义如下式所示：

$$\mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) \coloneqq \mathbb{E}_{z,z' \sim \mathcal{Q}(z)}\left[\mathcal{V}_{\mathcal{P}(z|x)}(z, z')\right] \approx \frac{1}{\mathrm{M}^2} \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} \mathcal{V}_{\mathcal{P}(z|x)}\left(z_i, z_j\right), \tag{A-1}$$

其中冯 • 米塞斯统计量 $\mathcal{V}_{\mathcal{P}(z|x)}\left(z_i, z_j\right)$ 根据下式进行定义：

$$\begin{aligned}\mathcal{V}_{\mathcal{P}(z|x)}(z, z') \coloneqq & \left[\nabla_z \log \mathcal{P}(z|x)\right]^{\top} K(z, z')\left[\nabla_{z'} \log \mathcal{P}(z'|x)\right] + \left[\nabla_z \log \mathcal{P}(z|x)\right]^{\top} \nabla_{z'} K(z, z') \\ & + \left[\nabla_z K(z, z')\right]^{\top}\left[\nabla_{z'} \log \mathcal{P}(z'|x)\right] + \operatorname{Trace}\left(\nabla_{z,z'} K(z, z')\right).\end{aligned} \tag{A-2}$$

在此基础上，根据 Liu 等人的论文[176]，重复 L 次调用下式（L 又被称为自举样本数），进行自举样本（bootstrap sample）$\mathbb{S}^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right)$ 的计算：

$$\mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) = \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} (w_{i,l} - \frac{1}{\mathrm{M}})(w_{j,l} - \frac{1}{\mathrm{M}}) \mathcal{V}_{\mathcal{P}(z|x)}\left(z_i, z_j\right), \tag{A-3}$$

这个重复调用过程被称为“自举迭代”（bootstrap iteration），其中 $l \in \{1, 2, \cdots, \mathrm{L}\}$ 是的自举迭代轮次的索引，而

$$(w_1, \cdots, w_{\mathrm{M}}) \sim \text{Mult}(\mathrm{M}; \frac{1}{\mathrm{M}}, \cdots, \frac{1}{\mathrm{M}}), \tag{A-4}$$

其中 Mult 表示多项分布（multinominal distribution），M 为从分布 $\mathcal{Q}(z)$ 采出的样本大小。

基于上述结果，按照下式计算满足符合 $\mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) > \mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right)$ 条件的自举样本比例 $\widehat{\alpha}$：

$$\widehat{\alpha} = \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \mathbb{I}\left[\mathbb{S}_l^*\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right) > \mathbb{S}\left(\mathcal{Q}(z), \mathcal{P}(z|x)\right)\right], \tag{A-5}$$

其中指示函数（indicator function）$\mathbb{I}$ 定义如下：

$$\mathbb{I}\left[\mathbb{S}_l^*\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right)>\mathbb{S}\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right)\right]\coloneqq\begin{cases}1 & \text{if } \mathbb{S}_l^*\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right)>\mathbb{S}\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right),\\ 0 & \text{if } \mathbb{S}_l^*\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right)\leq\mathbb{S}\left(\mathcal{Q}(z),\mathcal{P}(z|x)\right).\end{cases} \tag{A-6}$$

最后，对于预定义的置信水平 α 和以下假设[1]：

- 原假设 H_0: 样本 $z_i|_{i=1}^{\mathrm{M}}$ 来自于分布 $\mathcal{P}(z|x)$。
- 备择假设 H_1: 样本 $z_i|_{i=1}^{\mathrm{M}}$ 不来自于分布 $\mathcal{P}(z|x)$。

相应的决策规则如下：

- 如果 $\widehat{\alpha}>\alpha$，则拒绝 H_0。
- 如果 $\widehat{\alpha}\leq\alpha$，则接受 H_0。

综上所述，利用 KSD 进行拟合优度检验的流程可以总结为算法A.1：

算法 A.1 基于 KSD 的自举拟合优度检验算法伪代码

输入： 样本 $z_i|_{i=1}^{\mathrm{M}}\overset{\text{i.i.d.}}{\sim}\mathcal{Q}(z)$，分布 $\mathcal{P}(z|x)$ 的得分函数：$\nabla_z\log\mathcal{P}(z|x)$，显著性水平 α 和自举样本数 L。
输出： 原假设 H_0: $z_i|_{i=1}^{\mathrm{M}}$ 来自于分布 $\mathcal{P}(z|x)$；或备择假设 H_1：$z_i|_{i=1}^{\mathrm{M}}$ 不来自于分布 $\mathcal{P}(z|x)$。

```
1: 𝕊(𝒬(z), 𝒫(z|x)) ← 式(A-1)
2: α̂ ← 式(A-5)
3: if α̂ > α then
4:     return H_1
5: else
6:     return H_0
7: end if
```

A.2 成对样本 t-检验

本小节系统阐述成对样本 t-检验的实施方法。该方法通过比较在相同实验条件下两组模型的性能指标（如 RMSE），以评估其均值差异是否具有统计显著性。该分析对验证本文提出的软测量模型相较基线模型的性能提升是否来源于算法改进而非随机因素具有重要意义。

[1]置信水平 α 表示第一类错误的概率，即错误拒绝零假设 H_0 的风险。Fisher 在其研究中首次提出了 $\alpha=0.05$ 作为经验标准[228]，这一标准在平衡第一类和第二类错误方面表现良好，已被广泛接受并成为学术界的普遍共识[229]

接下来以 RMSE 为性能指标进行成对样本 t-检验的计算流程说明。设有两个工业过程软测量模型（模型 A 与模型 B），在相同随机种子集合 $\{s_1, s_2, \ldots, s_n\}$ 控制下进行训练与测试，建立配对样本数据集如下：

- 模型 A 的 RMSE 序列：$\mathcal{D}_{\mathrm{A}} = \{\mathrm{RMSE}_{\mathrm{A},1}, \mathrm{RMSE}_{\mathrm{A},2}, \ldots, \mathrm{RMSE}_{\mathrm{A},n}\}$
- 模型 B 的 RMSE 序列：$\mathcal{D}_{\mathrm{B}} = \{\mathrm{RMSE}_{\mathrm{B},1}, \mathrm{RMSE}_{\mathrm{B},2}, \ldots, \mathrm{RMSE}_{\mathrm{B},n}\}$

对于每个随机种子 s_i 对应的实验结果，定义性能差异量：

$$\Delta_i = \mathrm{RMSE}_{\mathrm{A},i} - \mathrm{RMSE}_{\mathrm{B},i}, \quad i \in \{1, 2, \ldots, n\} \tag{A-7}$$

建立如下统计假设以进行显著性检验：

- 零假设 H_0：$\mathbb{E}[\Delta_i] = 0$（模型间无显著性能差异）
- 备择假设 H_1：$\mathbb{E}[\Delta_i] \neq 0$（模型间存在统计显著性差异）

基于差异量序列 $\{\Delta_1, \Delta_2, \ldots, \Delta_n\}$，按以下步骤计算检验统计量：

1） 计算均值差异：

$$\bar{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \tag{A-8}$$

2） 估计差异量标准差：

$$s_\Delta = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\Delta_i - \bar{\Delta})^2} \tag{A-9}$$

3） 计算 t-统计量：

$$t = \frac{\bar{\Delta}}{s_\Delta / \sqrt{n}} \tag{A-10}$$

4） 确定 p-值：

$$p = 2\mathcal{P}(\mathrm{L} \geq |t|), \quad \mathrm{L} \sim t(n-1) \tag{A-11}$$

5） 统计决策：设定显著性水平 $\alpha = 0.05$，决策规则为：

- 若 $p < \alpha$，拒绝 H_0，表明差异具有统计显著性（置信度 $1 - \alpha$）
- 若 $p \geq \alpha$，接受 H_0，认为差异源于随机波动

完整计算流程的伪代码实现详见算法A.2。

算法 A.2 成对样本 t-检验的算法伪代码（RMSE 为例）

输入： 模型 A 的 RMSE 序列：$\mathrm{RMSE_A} = \{\mathrm{RMSE_{A,1}}, \mathrm{RMSE_{A,2}}, \ldots, \mathrm{RMSE_{A,}}_n\}$、模型 B 的 RMSE 序列：$\mathrm{RMSE_B} = \{\mathrm{RMSE_{B,1}}, \mathrm{RMSE_{B,2}}, \ldots, \mathrm{RMSE_{B,}}_n\}$ 和显著性水平：α。

输出： 零假设 H_0：模型间无显著性能差异或备择假设 H_1：模型间存在统计显著性差异。

$\Delta_i \leftarrow \mathrm{RMSE_{A,i}} - \mathrm{RMSE_{B,i}} \forall i \in \{1, 2, \ldots, n\}$
$\bar{\Delta} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \Delta_i$
$s_\Delta \leftarrow \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\Delta_i - \bar{\Delta})^2}$
$t \leftarrow \frac{\bar{\Delta}}{s_\Delta/\sqrt{n}}$
$p \leftarrow 2\mathcal{P}(\mathrm{L} \geq |t|), \quad \mathrm{L} \sim t(n-1)$
if $p < \alpha$ **then**
    return H_1
else
    return H_0
end if

B 第3章的实验设置

B.1 缺失数据的模拟方法

第3章采用文献[133]提出的框架模拟缺失场景，具体实现如下：

1）. **MAR：**

- 从特征空间中随机选取一个子集作为完全观测变量。
- 剩余特征的缺失模式通过逻辑斯蒂模型（logistic model）生成，其中模型以已选完全观测变量作为协变量，逻辑斯蒂回归的系数通过随机初始化得到，并通过调节逻辑斯蒂回归的截距项（bias term）精确控制总体缺失率 p_{miss}。

2）. **MCAR：**

- 对每个观测样本独立生成掩码变量，其中掩码变量服从参数为 p_{miss} 的伯努利分布（Bernoulli distribution）。
- 样本的缺失概率为预设 p_{miss}，以确保所有样本同概率缺失。

3）. **MNAR：**

- 在第一阶段采用 MAR 场景的模拟方法选择待缺失特征。
- 在第二阶段对选定的特征施加 MCAR 场景的模拟方法以构建缺失值。

B.2 超参数配置和实验方案

KnewImp 算法的带宽 v 设置为 0.5，离散步长设置为 1.0×10^{-1}，$\mathcal{L}^{\text{DSM}}$ 中的神经网络学习率 η 为 1.0×10^{-3}，隐层单元数 HU_{score} 为 256，外层迭代轮次 $\mathcal{T}$ 和内层泛函优化轮次 T 分别被设置为 1 和 200；而基线算法的超参数配置如下所示：

- **批次大小：**基线方法的批次统一设置为 512。
- **隐藏单元数：**
  - MIWAE 的特征数和隐层单元数分别为 16 和 32。
  - TDM 算法中的神经网络隐藏层数为 16，其神经网络层数设置为 2。
  - MIRACLE 的隐层单元设置为 32。
  - MissDiff 和 CSDI_T 的频道数设置为 16，嵌入层（embedding layer）维度设置为 128，神经网络隐藏层层数设置为 2。
- **微分方程离散步数和粒子数：**MissDiff 和 CSDI_T 的扩散步数设置为 100，粒子数设置为 50。

所有实验均在配备 Intel Xeon E5×4 处理器、Nvidia GTX 1080×8 显卡和 128 GB 内存的工作站上进行；所有涉及到深度学习模型的算法均采用 PyTorch 后端默认的 Kaiming Uniform 初始化方法[230]以及 Adam 优化器[87]进行训练。所有算法均采用 k-近邻（k-Nearest Neighbor）算法（$k = 20$）进行初始化补全。模型训练与推理均在 Python 3.8 环境和 PyTorch 1.20 后端中完成，所有实验均在至少 4 个不同的随机种子下进行独立实验。根据文献[133-134]，考虑到数据补全任务本身未划分训练集、验证集和测试集，第3章实验不进行数据集划分，统一采用整体数据进行评估。

C 第4章的超参数配置和实验方案

第4章使用多层感知器对 $p_\theta(x|z)$ 进行参数化。根据文献[13]，第4章将多层感知器的隐藏单元设置为 $[10, 7, 5]$。在此基础上，设置 InfO-PLVM 的粒子数 M、带宽 v、离散步长 ε 和最优控制模拟时间 T 分别为 20、1.0、0.01 和 200。其余的超参数如下所示：

- **批次大小 $\mathcal{B}$ 的设置：**DBPSFA、PDTM、MUDVAE-SDVAE、NPLVR、GMVAE 和 InfO-PLVM 的批次大小 $\mathcal{B}$ 分别为 32、128、32、64、128、64。

- **学习率 η 的设置：**DBPSFA、PDTM、MUDVAE-SDVAE、NPLVR、GMVAE 和 InfO-PLVM 的学习率 η 分别为 0.005、0.01、0.01、0.01、0.0001、0.1。

- **隐变量维度 $\mathrm{D_{LV}}$ 的设置：**DBPSFA、PDTM、MUDVAE-SDVAE、GMVAE 和 InfO-PLVM 隐变量维度 $\mathrm{D_{LV}}$ 均设置为 5；由于 NPLVR 由多个 VAE 堆叠而成，因此其隐变量维度列表设为 $[10, 7, 5]$。

所有实验均在配备 Intel Xeon E5×4 处理器、Nvidia GTX 1080×8 显卡和 128 GB 内存的工作站上进行；所有深度学习模型均采用 PyTorch 后端默认的 Kaiming Uniform 初始化方法[230]以及 Adam 优化器[87]进行训练。模型训练与推理均在 Python 3.8 环境和 PyTorch 1.20 的深度学习后端中完成。所有实验均在至少 4 个不同的随机种子下进行独立实验。对于所有数据集，本文按照时间戳升序对数据进行排序。在此基础上，前 60% 的数据作为训练集，60% 至 80% 区间的数据作为验证集，剩余数据作为测试集。所有深度学习模型的训练流程均设置为 200 轮，每 5 轮进行一次验证，并在验证集效果最佳的模型参数上进行最终测试。

D 第5章的超参数配置和实验方案

第5章的超参数如下所示：

- **批次大小 $\mathcal{B}$ 的设置：**在水煤气变换单元数据集上，Flashformer、iTransformer、Synthesizer (Random)、Synthesizer (Dense)、S-GCN、TGLFA、SDGNN、GCLAda、UniFilter、DGDL 和 $\mathrm{E^2AG}$ 的批次大小 $\mathcal{B}$ 分别被设置为 128、128、128、512、32、128、64、64、64 和 128；在二氧化碳吸收单元数据集上，Flashformer、iTransformer、Synthesizer (Random)、Synthesizer (Dense)、S-GCN、TGLFA、SDGNN、GCLAda、UniFilter、DGDL 和 $\mathrm{E^2AG}$ 的批次大小 $\mathcal{B}$ 分别被设置为 128、128、128、512、32、128、64、64、64、256 和 128。

- **学习率 η 的设置：**在水煤气变换单元数据集上，Flashformer、iTransformer、Synthesizer (Random)、Synthesizer (Dense)、S-GCN、TGLFA、SDGNN、GCLAda、UniFilter、DGDL 和 $\mathrm{E^2AG}$ 的学习率 η 分别被设置为 0.02、0.02、0.02、0.01、0.001、0.01、0.001、0.01、0.01、

0.001 和 0.02；在二氧化碳吸收单元数据集上，Flashformer、iTransformer、Synthesizer (Random)、Synthesizer (Dense)、S-GCN、TGLFA、SDGNN、GCLAda、UniFilter、DGDL 和 E^{2}AG 的学习率 η 分别被设置为 0.02、0.01、0.02、0.01、0.001、0.01、0.001、0.001、0.01、0.001 和 0.02。

- **隐藏层单元数的设置：**在水煤气变换单元数据集上，Flashformer、iTransformer、Synthesizer (Random)、Synthesizer (Dense)、S-GCN、TGLFA、SDGNN、GCLAda、UniFilter、DGDL 和 E^{2}AG 的藏层单元数分别被设置为 16、16、16、16、16、32、16、32、16、32 和 1；在二氧化碳吸收单元数据集上，Flashformer、iTransformer、Synthesizer (Random)、Synthesizer (Dense)、S-GCN、TGLFA、SDGNN、GCLAda、UniFilter、DGDL 和 E^{2}AG 的藏层单元数分别被设置为 16、16、16、16、16、32、16、16、16、32 和 1。

- **注意力头数/粒子数：**在水煤气变换单元数据集和二氧化碳吸收单元数据集上，Flashformer、iTransformer、Synthesizer (Random) 和 Synthesizer (Dense) 的注意力头数都被设置为 2；在水煤气变换单元数据集和二氧化碳吸收单元数据集上 E^{2}AG 的粒子数都被设置为 6。

所有实验均在配备 Intel Xeon E5×4 处理器、Nvidia GTX 1080×8 显卡和 128 GB 内存的工作站上进行；所有深度学习模型均采用 PyTorch 后端默认的 Kaiming Uniform 初始化方法[230]以及 Adam 优化器[87]进行训练。模型训练与推理均在 Python 3.8 环境和 PyTorch 1.20 的深度学习后端中完成，所有实验均在至少 4 个不同的随机种子下进行独立实验。对于所有数据集，本文按照时间戳升序对数据进行排序。在此基础上，前 60% 的数据作为训练集，60% 至 80% 区间的数据作为验证集，剩余数据作为测试集。所有深度学习模型的训练流程均设置为 200 轮，每 5 轮进行一次验证，并在验证集效果最佳的模型参数上进行最终测试。

E 第6章的超参数配置和实验方案

在第6章的实验中，历史序列的长度 T 被设置为 5，而各模型在水煤气变换单元数据集和二氧化碳吸收单元数据集的超参数配置如下：

- **批次大小 $\mathcal{B}$ 设置：**概率隐变量类模型、变压器网络类模型、循环神经网络类模型、DMVAER 和 OC-NDPLVM 分别设为 128、64、32、16 和 32。DPMM 和 DMVAER 的混合组件数经网格搜索分别设为 2 和 3。

- **学习率 η 设置：**概率隐变量类模型、变压器网络类模型、循环神经网络类模型、DMVAER 和 OC-NDPLVM 分别设为 0.001、0.001、0.001、0.001 和 0.005。

- **变压器网络类模型的其他参数：**编码器层数 3 层、解码器层数 1 层、注意力头数 4 个、随机丢弃率（dropout rate）0.3、隐层维度 16。Informer 模型的蒸馏层参数设为 $[3, 2, 1]$。

- **循环神经网络的记忆单元的选择：**NDPLVMs、OC-NDPLVM 和 DMVAER 均采用 2 层 GRU 结构。

- **隐变量维度 $\mathrm{D_{LV}}$：**NPLVR 之外的概率隐变量类模型的隐变量维度 $\mathrm{D_{LV}}$ 均设为 4。

- **其他模型参数：**

  – AR-TCN 通道列表设为 $[64, 32, 16]$。

  – 由于 NPLVR 由多个 VAE 堆叠而成，因此其隐变量维度列表设为 $[10, 7, 5]$。

  – DA-LSTM 编码器层数 2 层、解码器层数 1 层、隐藏维度 4。

所有实验均在配备 Intel Xeon E5×4 处理器、Nvidia GTX 1080×8 显卡和 128 GB 内存的工作站上进行。为确保公平性，所有深度学习模型均采用 PyTorch 后端默认的 Kaiming Uniform 初始化方法[230]以及 Adam 优化器[87]进行训练。模型训练与推理均在 Python 3.8 环境和 PyTorch 1.20 的深度学习后端中完成。所有实验均在至少 4 个不同的随机种子下进行独立实验。对于所有数据集，本文按照时间戳升序对数据进行排序。在此基础上，前 60% 的数据作为训练集，60% 至 80% 区间的数据作为验证集，剩余数据作为测试集。所有深度学习模型的训练流程均设置为 200 轮，每 5 轮进行一次验证，并在验证集效果最佳的模型参数上进行最终测试。

攻读博士学位期间发表的科研成果

以第一作者身份发表论文

1. Improving Data-Driven Inferential Sensor Modeling by Industrial Knowledge: A Bayesian Perspective, IEEE Transactions on Systems Man and Cybernetics: Systems. [中科院 1 区 Top 期刊，CAA-A 类期刊，CCF-B 类期刊，IEEE 汇刊，IF:8.7，对应第5章].

2. Rethinking the Diffusion Models for Missing Data Imputation: A Gradient Flow Perspective, The 38th Neural Information Processing Systems Conference（NeurIPS 2024），Main Track. [CCF-A 类会议，对应第3章]

3. E^2AG: Entropy-Regularized Ensemble Adaptive Graph for Industrial Soft Sensor Modeling, IEEE/CAA Journal of Automatica Sinica. [中科院 1 区 Top 期刊，CCF-T1 类期刊，CAA-A^+ 类期刊，IF:19.2，对应第5章].

4. Analyzing and improving supervised nonlinear dynamical probabilistic latent variable model for inferential sensors, IEEE Transactions on Industrial Informatics. [中科院 1 区 Top 期刊，CAA-A^+ 类期刊，CCF-C 类期刊，IEEE 汇刊，IF:9.9，对应第6章].

5. Variational Inference Over Graph: Knowledge Representation for Deep Process Data Analytics, IEEE Transactions on Knowledge and Data Engineering. [中科院 1 区 Top 期刊，JCR-Q1 期刊，CCF-A 类期刊，CAA-A 类期刊，IF:10.4，对应第 5 章].

6. Monotonic neural ordinary differential equation: Time-series forecasting for cumulative data, The 32nd ACM International Conference on Information and Knowledge Management（CIKM 2023），Applied Research Track. [CCF-B 类会议].

7. Directed Acyclic Graphs With Tears, IEEE Transactions on Artificial Intelligence. [IEEE 汇刊].

8. Knowledge Automation Through Graph Mining, Convolution, and Explanation Framework: A Soft Sensor Practice, IEEE Transactions on Industrial Informatics. [中科院 1

区 Top 期刊，JCR-Q1 期刊，CAA-A$^+$ 类期刊，CCF-C 类期刊，IEEE 汇刊，IF:9.9].

9. Stochastic Optimization-based Approach for Simultaneous Process Design and HEN Synthesis of Tightly-coupled RO-ORC-HI Systems Under Seasonal Uncertainty, Chemical Engineering Science. [中科院 2 区期刊，JCR-Q2 期刊，CIESC-T1 类期刊，IF:4.3].

10. 耦合化工过程的废热回收废水脱盐系统的优化, 高校化学工程学报. [EI 期刊，CIESC-T2 类期刊].

以第一作者身份在投论文

1. Relaxing Latent Variable Specification for Probabilistic Latent Variable Models: An Infinite-Time Optimal Control Approach, IEEE Transactions on Automatic Control. [中科院 2 区 Top 期刊，JCR-Q1 期刊，CAA-A$^+$ 类期刊，IF:6.2，对应第4章，一审].

以第一作者身份出版专著章节

1. Integrating Incomplete Prior Knowledge into Data-Driven Inferential Sensor Models Under Variational Bayesian Framework, Applied AI Techniques in the Process Industry, Chapter 7. [Wiely 出版社出版]

以合作作者身份发表论文

1. Optimal transport for treatment effect estimation, The 37th Neural Information Processing Systems Conference （NeurIPS 2023）, Main Track. [CAA-A 类会议，CCF-A 类会议，CAAI-A 类会议]

2. Entire Space Counterfactual Learning for Reliable Content Recommendations, IEEE Transactions on Information Forensics and Security. [中科院 1 区 Top 期刊，CCF-A 类期刊，IEEE 汇刊，IF=8.0]

3. Debiased Recommendation via Wasserstein Causal Balancing，ACM Transactions on Information Systems. [JCR-Q1 期刊，CCF-A 类期刊，ACM 汇刊，IF=9.1]

4. Heat Equation Stein Variational Ensemble: Rethinking and Advancing Uncertainty-Aware Soft Sensor Modeling, IEEE Transactions on Industrial Informatics. [中科院 1 区 Top 期刊，CAA-A$^+$ 类期刊，CCF-C 类期刊，IEEE 汇刊，IF:9.9].

5. When Adaptive Graph Neural Network Meets Kolmogorov-Arnold Network for Industrial Soft Sensors, IEEE Transactions on Instrumentation & Measurement. [中科院 2 区期刊，CAA-A$^+$ 类期刊，CIS-T1 类期刊，IF:5.9]

6. LSPT-D: Local Similarity Preserved Transport for Direct Industrial Data Imputation, IEEE Transactions on Automation Science and Engineering. [中科院 2 区 Top 期刊，CAA-A$^+$ 类期刊，CCF-B 类期刊，IEEE 汇刊，IF:6.4].

7. SPOT-I: Similarity Preserved Optimal Transport for Industrial IoT Data Imputation, IEEE Transactions on Industrial Informatics. [中科院 1 区 Top 期刊，CAA-A$^+$ 类期刊，CCF-C 类期刊，IEEE 汇刊，IF:9.9].

8. Optimal Transport for Time Series Imputation, The 13th International Conference on Learning Representations（ICLR 2025）, Main Track. [CAA-A 类会议，CAAI-A 类会议]