DOI:10.7544/issn1000-1239.20200637
59(3):617-632,2022

概率生成模型变分推理方法综述

陈亚瑞　杨巨成　史艳翠　王　嫄　赵婷婷
(天津科技大学人工智能学院　天津　300457)
(yrchen@tust.edu.cn)

Survey of Variational Inferences in Probabilistic Generative Models

Chen Yarui, Yang Jucheng, Shi Yancui, Wang Yuan, and Zhao Tingting
(*College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin* 300457)

Abstract　Probabilistic generative models are important methods for knowledge representation. Exact probabilistic inference methods are intractable in these models, and various approximate inferences are required. The variational inferences are important deterministic approximate inference methods with rapid convergence and solid theoretical foundations, and they have become the research hot in probabilistic generative models with the development of big data especially. In this paper, we first present a general variational inference framework for probabilistic generative models, and analyze the parameter learning process of the models based on variational inference. Then, we give the framework of analytic optimization of variational inference for the conditionally conjugated exponential family, and introduce the stochastic variational inference based on the framework, which can scale to big data with the stochastic strategy. Furthermore, we provide the framework of black box variational inferences for the general probability generative models, which train the model parameters of variational distributions based on the stochastic gradients; and analyze the realization of different variational inference algorithms under the framework. Finally, we summarize the structured variational inferences, which improve the inference accuracy by enriching the variational distributions with different strategies. In addition, we discuss the further development trends of variational inference for probabilistic generative models.

Key words　probabilistic generative model; variational inference; conditional conjugate exponential family; black box variational inference; structured variational inference

摘　要　概率生成模型是知识表示的重要方法,在该模型上计算似然函数的概率推理问题一般是难解的.变分推理是重要的确定性近似推理方法,具有较快的收敛速度、坚实的理论基础.尤其随着大数据时代的到来,概率生成模型变分推理方法受到工业界和学术界的极大关注.综述了多种概率生成模型变分推理框架及最新进展,具体包括:首先综述了概率生成模型变分推理一般框架及基于变分推理的生成模型参数学习过程;然后对于条件共轭指数族分布,给出了具有解析优化式的变分推理框架及该框架下

收稿日期:2020-08-25;**修回日期**:2021-03-04
基金项目:国家自然科学基金项目(61976156, 61702367);天津市自然科学基金项目(18JCQNJC69800);天津科技大学青年教师基金项目(2017LG10)
This work was supported by the National Natural Science Foundation of China (61976156, 61702367), the Tianjin Natural Science Foundation (18JCQNJC69800), and the Youth Scholars Foundation of Tianjin University of Science and Technology (2017LG10).
通信作者:赵婷婷(tingting@tust.edu.cn)

可扩展的随机化变分推理;进一步,对于一般概率分布,给出了基于随机梯度的黑盒变分推理框架,并简述了该框架下多种变分推理算法的具体实现;最后分析了结构化变分推理,通过不同方式丰富变分分布提高推理精度并改善近似推理一致性.此外,展望了概率生成模型变分推理的发展趋势.

关键词 概率生成模型;变分推理;条件共轭指数族;黑盒变分推理;结构化变分推理

中图法分类号 TP18

概率生成模型(probabilistic generative model)是知识表示的重要方法,拟从数据集中学习数据概率分布的估计形式.该估计可以是显式的概率密度函数形式,也可以是隐式表示,即不具有显式的概率密度函数,只有生成样本数据的能力[1].具有隐式概率密度函数的生成模型的典型范例是生成对抗网络,这类模型虽然可以生成高质量图像,但是无法获得真实分布的多样性,会导致模式崩塌,同时缺少度量生成模型质量的有效指标[2].具有显式概率密度函数的生成模型通过显式分布建模,该类模型以对数似然函数作为优化目标,可以覆盖数据集上的所有模式,避免模式崩塌问题[1-2].但是在该类模型上求解似然函数一般是难解的,发展了各种近似推理方法,主要包括2类方法:变分推理方法(variational inference methods)[3-4]和蒙特卡洛方法(Monte Carlo methods)[5-6].蒙特卡洛方法通过采样对概率分布进行估计,根据大数定律可知,在采样数目足够多时,蒙特卡洛方法可以很好地估计目标函数,但是存在采样效果严重依赖超参数设置、收敛缓慢等缺点[5-6].变分推理方法把变量求和的概率推理问题转化成优化问题,具有坚实的理论基础、较快的收敛速度、紧致的变分下界,且较容易扩展到大规模数据集,受到研究者的关注[3-4].

对于概率生成模型,变分推理方法通过引入变分分布作为后验概率分布的近似分布,把变量求和或求积分的概率推理问题转化为优化问题,再近似求解该变分优化问题[3-4,7].概率生成模型变分推理方法的发展经历了不同的发展阶段,从开始的指数族(exponential family)分布下具有解析优化表示的变分推理框架[8-12],到一般概率分布下基于随机梯度的黑盒变分推理框架[13-23],再到深度模型结构下基于分摊变分推理(amortized variational inference)框架[24-32],以及基于丰富变分分布的各种结构化变分推理方法[33-45].概率生成模型的变分推理已经被广泛应用于计算生物学、计算机视觉、计算神经科学、自然语言处理和语音识别等领域[3-4,7].

变分推理起源于20世纪80年代,最具代表性的是统计物理学中的均值场方法[8],20世纪90年代变分推理被机器学习领域研究者广泛关注,并被应用到概率图模型中[9-11].之后,Wainwright教授团队[4]指出,在指数族分布下,结合均值场变分假设,基于坐标上升(coordinate ascent)优化方法可以给出变分优化式的解析表示形式.Bishop教授团队[3]进一步利用指数族分布的条件共轭(conditional conjugate)性质,给出变分贝叶斯优化问题的解析表示形式.但是经典的基于坐标上升的变分推理方法很难推广到大规模数据集,Hoffman教授团队[12]提出随机化变分推理(stochastic variational inference)方法,利用条件共轭指数族分布的自然梯度性质,并结合随机化方法,可以方便将变分推理应用到大规模数据集.条件共轭指数分布族下,结合均值场假设,不需要指定变分分布的形式,利用坐标上升优化方法可以直接给出变分分布及参数后验概率分布的解析表示形式.该类方法可靠性高、优化更新过程收敛速度快,但是条件共轭指数族分布的条件限制了其应用范围,较难应用于复杂模型.

一般的概率生成模型,尤其是深度复杂模型及复合模型,如贝叶斯深度网络、复杂的层次化结构等,都不满足指数族分布条件.针对这类模型,研究者提出了基于随机梯度(stochastic gradient)的黑盒变分推理(black box variational inference)[13-14],这类方法针对变分优化目标,先计算优化目标的随机梯度,再采用随机梯度方法进行目标优化问题求解.这种基于随机梯度的训练方式可方便应用于复杂概率分布的大规模数据训练中.计算变分优化目标的无偏随机梯度估计是黑盒变分推理的关键任务,同时降低随机梯度的方差使算法有较快的收敛速度是黑盒变分推理算法设计的难点[46].目前最重要的2类梯度计算方法包括评分函数梯度(scoring function gradient)[13]和重参梯度(reparameteration gradient)[15-16].评分函数梯度利用梯度性质及蒙特卡洛采样直接计算变分优化目标的带噪无偏随机梯度,但该方法的方差较大,影响算法性能,后续开展了各种降低随机梯度方差的方法[17-19].重参随机梯度

首先对变分分布进行重参转换,再对重参后的目标函数计算梯度[15-16].实践中重参随机梯度比评分梯度的方差小很多,但并没有直接的理论保证[20].但是重参随机梯度的应用范围有限,只能用于连续隐变量,且变分分布形式可表示成位置尺度类分布、累计分布函数可逆类分布,或可以表示成上述2种类型的分布.研究者后续开展了离散型隐变量的重参方法[21-22],及对于其他类型连续隐变量的隐式重参方法[23]等.基于随机梯度的黑盒变分推理,对概率生成模型分布及结构形式没有限制,但是需要提前建模变分分布结构,通过随机梯度方法求解其变分参数.随机梯度的方差大小直接决定了算法的收敛速度及优化效果,但对于复杂模型,较难直接有效地度量随机梯度的方差,这也是黑盒变分推理算法设计的关键和难点[46].

随着深度神经网络的发展,如何将变分推理应用到深度模型引起了研究者的兴趣.分摊变分推理[24]利用参数化函数,使所有样本点共享变分分布参数,为深度概率生成模型的发展提供了很好的思路[15-16].相比于传统变分推理中每个样本点都具有独立的变分分布参数,分摊变分推理的参数共享方式,可以有效处理深层神经网络,推进了深度概率生成模型的发展.变分自编码(variational autoencoder)模型是深度概率生成模型的典型范例,其概率生成模型采用深度神经网络,其变分分布采用基于分摊变分推理的深度神经网络,并采样基于重参梯度的方法进行模型训练[15-16].之后研究者通过丰富变分分布及生成模型提出了各种改进的变分自编码模型,包括基于隐式分布的变分自编码模型[25-27]、重要加权自编码(important weight autoencoder)模型[28-29]、对抗自编码(adversarial autoencoder)模型[30]、对抗变分贝叶斯(adversarial variational Bayes)模型[31]、像素变分自编码(pixel variational autoencoder)模型[32]等.

基于均值场假设的变分推理极大地简化了优化过程,并得到了广泛应用,上述的指数族分布下具有解析优化式的变分推理,基于随机梯度的黑盒变分推理中的部分算法都利用了均值场假设.但是均值场变分推理存在训练渐进不一致性问题,即当训练样本足够多时,近似后验概率分布仍无法收敛到精确值,使得推理精度受到影响.一直以来各种研究工作通过丰富变分分布来提升变分推理渐进性及推理精度.结构化均值场方法[33-34]结合了传统的均值场方法和精确推理方法,需从模型中识别出易处理子结构,子结构之间采用均值场方法,而子结构内部采用精确推理,这类方法尤其适用于时间序列模型[35-36].标准化流(normalizing flows)方法[37]通过一系列可逆映射将简单的概率密度转换成复杂的概率密度函数,可以提供更紧致的变分界.采用不同的映射函数及结合方法发展出了各种标准化流方法,如平面流(planar flow)、径向流(radial flow)[37]、掩模自回归流(masked autoregressive flow)[38]、可逆自回归流(inverse autoregressive flow)[39]等.层次化变分推理(hierarchical variational inference)模型通过对均值场变分分布的参数引入先验分布增加均值场变量之间的相关性[17].耦合变分推理(copula variational inference)[40-41]通过对均值场变分分布引入耦合分布增加分量之间的相关性.基于混合分布的变分推理利用混合分布的灵活性,提高变分分布结构的灵活性,从而提高变分推理精度[42-43].这类通过丰富变分分布结构提升变分推理精度的方法我们统称为结构化变分推理方法.该类方法通过增加变分分布中隐变量之间的结构信息提升了变分分布的表示能力,尤其对于模型隐变量结构有天然密切相关性模型,如隐Markov模型(hidden Markov models)[35]、动态主题模型等(dynamic topic models)[36],可以明显提高推理精度,但是模型训练过程中需要更多的计算开销.

本文首先介绍概率生成模型、贝叶斯概率生成模型,及相应模型上的概率推理问题;并介绍了概率生成模型变分推理的一般框架、变分优化问题求解方法,及基于变分推理的模型参数学习过程;然后对于条件共轭指数族模型,给出了基于坐标上升的均值场变分推理,及推广至大数据集的随机化变分推理方法;进一步,给出了结合随机梯度的黑盒变分推理一般框架,并分析了该框架下多种变分推理算法的具体实现;最后综述了结构化变分推理方法,通过不同策略丰富变分分布结构提高变分推理精度;本文最后讨论了概率生成模型变分推理的发展趋势并进行了总结.

1 概率生成模型

本节介绍概率生成模型、贝叶斯概率生成模型,及相应的概率推理问题.

1.1 概率生成模型

概率生成模型又称隐变量生成模型,是一种重要的生成模型,它假设复杂的、可观测向量由简单

的、不可观测的隐向量(又称为特征、表示)根据某种模式生成[1-4].概率生成模型的目标就是揭示数据集中蕴含的这种模式,并进一步利用该模型进行数据生成、预测等.

概率生成模型的全概率分布表示形式为

$$p(\boldsymbol{x},\boldsymbol{z})=p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta}), \tag{1}$$

其中:$\boldsymbol{x}\in\mathbb{R}^D$表示可观测向量;$\boldsymbol{z}\in\mathbb{R}^M$表示连续隐向量;$p(\boldsymbol{z})$表示隐向量先验概率分布,一般情况下选用高斯分布 $p(\boldsymbol{z})=\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$;$p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$表示条件概率分布,$\boldsymbol{\theta}$ 表示条件概率分布参数.

在概率生成模型下,某样本 $\boldsymbol{x}^{(i)}$ 的生成过程为:首先从隐变量先验分布 $p(\boldsymbol{z})$ 中随机生成一个隐向量样本 $\boldsymbol{z}^{(i)}$,然后根据条件概率分布 $p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$ 生成样本 $\boldsymbol{x}^{(i)}$.概率生成模型中数据生成过程的概率表示形式为

$$\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),$$
$$\boldsymbol{x}|\boldsymbol{z}\sim p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta}).$$

1.2 贝叶斯概率生成模型

对于概率生成模型,若参数 $\boldsymbol{\theta}$ 是随机向量,则称为贝叶斯概率生成模型,此时模型的全概率分布表示形式为

$$p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})=p(\boldsymbol{z})p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta}), \tag{2}$$

其中:$p(\boldsymbol{\theta})$表示参数先验概率分布;$p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta})$表示基于参数 $\boldsymbol{\theta}$ 和隐向量 $\boldsymbol{z}$ 的条件概率分布.此时 $\boldsymbol{\theta}$ 又称为全局隐向量,$\boldsymbol{z}$ 又称为局部隐向量.普通概率生成模型的学习任务之一是求解模型参数 $\boldsymbol{\theta}$,而贝叶斯概率生成模型是求解参数的后验概率分布 $p(\boldsymbol{\theta}|\boldsymbol{x})$.

概率生成模型的图模型结构如图 1 所示,其中每个圆形节点表示随机向量,白色节点表示隐向量,灰色节点表示可观测向量,节点之间的有向线段表示变量之间的依赖关系;黑色实心小方块表示模型参数值;方框表示过程可以重复出现,如概率生成模型中,基于隐向量可以重复生成数据样本,观测到 N 条数据,则该过程重复出现了 N 次.图 1(a)表示一般的概率生成模型,其中参数 $\boldsymbol{\theta}$ 表示参数值,需要通过数据集求解该模型参数;图 1(b)表示贝叶斯概率生成模型,其中参数 $\boldsymbol{\theta}$ 表示随机向量,需要根据数据集求解参数的后验概率分布 $p(\boldsymbol{\theta}|\boldsymbol{x})$.

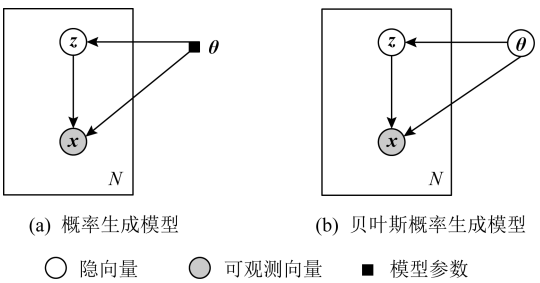

Fig. 1 The graphical models of two models

图 1 2 种模型的图模型结构

1.3 概率推理问题

对于可观测的数据集 $X=\{\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\cdots,\boldsymbol{x}^{(N)}\}$,令 $Z=\{\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\cdots,\boldsymbol{z}^{(N)}\}$ 表示与观测数据对应的隐变量集合.采用生成模型 $p(\boldsymbol{x},\boldsymbol{z})$ 对该数据集进行建模,数据集的联合概率分布形式为

$$p(X,Z)=\prod_{i=1}^{N}p(\boldsymbol{z}^{(i)})p(\boldsymbol{x}^{(i)}\mid\boldsymbol{z}^{(i)};\boldsymbol{\theta}). \tag{3}$$

此时概率推理问题是根据联合概率分布计算观测数据的边缘概率分布 $p(X)$ 及隐变量的后验概率分布 $p(Z|X)$,即:

$$p(X)=\prod_{i=1}^{N}\int p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})\mathrm{d}\boldsymbol{z}^{(i)}, \tag{4}$$

$$p(Z|X)=\prod_{i=1}^{N}\frac{p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})}{p(\boldsymbol{x}^{(i)})}. \tag{5}$$

进一步,利用最大对数似然可以求解模型参数 $\boldsymbol{\theta}$,即:

$$\boldsymbol{\theta}=\arg\max\ln p(X).$$

对于贝叶斯概率生成模型 $p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})$,需要计算随机变量参数 $\boldsymbol{\theta}$ 的后验概率分布 $p(\boldsymbol{\theta}|\boldsymbol{x})$.式(4)(5)中的变量积分是难解的,需要利用变分推理进行近似求解.

2 变分推理一般框架

本节给出概率生成模型的变分推理框架,分析变分优化问题求解方法,并给出基于变分推理的模型参数学习框架.

2.1 变分推理

对于概率生成模型 $p(\boldsymbol{x},\boldsymbol{z})$,数据集 X 的对数似然函数为 $\ln p(X)=\sum_{i=1}^{N}\ln p(\boldsymbol{x}^{(i)})$,其中单样本的对数似然为

$$\ln p(\boldsymbol{x}^{(i)})=\ln\int p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})\mathrm{d}\boldsymbol{z}^{(i)}. \tag{6}$$

式(6)中对隐向量求和是难解的,变分推理通过引入变分分布 $q(\boldsymbol{z}^{(i)})$ 作为后验概率分布 $p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$ 的近似分布,对 $\ln p(\boldsymbol{x}^{(i)})$ 进行变分变换可得

$$\ln p(\boldsymbol{x}^{(i)})=\mathcal{L}(q(\boldsymbol{z}^{(i)}))+$$
$$D_{\mathrm{KL}}(q(\boldsymbol{z}^{(i)})\,\|\,p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})). \tag{7}$$

各部分具体形式为

$$\mathcal{L}(q(\boldsymbol{z}^{(i)}))=E_{q(\boldsymbol{z}^{(i)})}[\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})-\ln q(\boldsymbol{z}^{(i)})],$$

$$D_{\mathrm{KL}}(q(\boldsymbol{z}^{(i)})\parallel p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}))=$$

$$E_{q(\boldsymbol{z}^{(i)})}\left[\ln\frac{q(\boldsymbol{z}^{(i)})}{p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})}\right],$$

其中,$\mathcal{L}(q(\boldsymbol{z}^{(i)}))$称为对数似然 $\ln p(\boldsymbol{x}^{(i)})$的变分下界(variational lower bound);$q(\boldsymbol{z}^{(i)})$表示隐向量 $\boldsymbol{z}^{(i)}$的变分分布;$D_{\mathrm{KL}}(q\parallel p)\geqslant 0$ 表示变分分布 $q(\boldsymbol{z}^{(i)})$与真实后验分布 $p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$之间的 KL 散度距离,当且仅当 $q(\boldsymbol{z}^{(i)})\equiv p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$时等号成立,可以得出 $\ln p(\boldsymbol{x}^{(i)})\geqslant\mathcal{L}(q(\boldsymbol{z}^{(i)}))$.

根据式(7)可知最小化 $D_{\mathrm{KL}}(q(\boldsymbol{z}^{(i)})\parallel p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}))$等价于最大化变分下界 $\mathcal{L}(q(\boldsymbol{z}^{(i)}))$,此时变量求和的概率推理问题转化为变分优化问题:

$$q^{*}(\boldsymbol{z}^{(i)})=\arg\max_{q(\boldsymbol{z}^{(i)})\in D}\mathcal{L}(q(\boldsymbol{z}^{(i)})). \tag{8}$$

通过求解该优化问题可以计算出隐向量后验概率分布及证据似然下界,即以最优变分分布作为隐向量后验概率分布的近似值 $q^{*}(\boldsymbol{z}^{(i)})\approx p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$,同时计算出对数似然的下界:$\mathcal{L}(q^{*}(\boldsymbol{z}^{(i)}))$.

利用不同的概率分布距离度量方式可以得到不同的变分优化问题.基于最小化 KL 散度 $D_{\mathrm{KL}}(q\parallel p)$的优化问题,又称为基于 KL 散度距离的变分推理[47],是最经典的变分转换方式之一,也是本文分析的变分推理的核心.KL 散度距离是非对称的,通过最小化 $D_{\mathrm{KL}}(p\parallel q)$也可以构建相应的变分优化问题,这类方法称为期望传播方法[48-49].α-散度距离 $D_{\alpha}(p\parallel q)$是 KL 散度距离的推广,研究者进一步将 α-散度距离应用到变分优化问题构建中[27,50-51].f-散度距离 $D_{f}(p\parallel q)$是更一般化的概率分布距离度量方式[52-53],其中 f 表示满足条件的凸函数,不同的 f 函数定义对应于不同的距离,α-散度距离是其特殊表示形式,研究者将 f-散度距离应用于变分推理,给出更灵活的一般化变分优化式.

2.2 变分优化问题求解

求解概率生成模型变分优化问题式(8)的方法,主要包括基于坐标上升的优化方法[9,33]和基于随机梯度的优化方法[13].基于坐标上升的优化方法用于变分分布满足均值场假设的模型,对变分分布的各维分布迭代更新.基于随机梯度的优化方法主要用于变分分布模型结构固定,采用梯度上升方法对变分分布模型参数进行迭代更新.

1) 基于坐标上升的优化方法

均值场变分推理的基本思想是通过假设隐向量各分量之间是相互独立的,从而限定了变分分布的取值空间,简化了优化训练[9,33].具体地,将隐向量 $\boldsymbol{z}$ 分解成不相交的子向量,记为$\{\boldsymbol{z}_{j}\mid j=1,2,\cdots,J\}$,此时隐向量的变分分布关于子集是边缘独立的,变分分布可以表示成分解形式:

$$q(\boldsymbol{z})=\prod_{j=1}^{J}q_{j}(\boldsymbol{z}_{j}).$$

完全均值场变分是最特殊分解形式,即变分分布按各分量进行分解 $q(\boldsymbol{z})=\prod_{j=1}^{M}q_{j}(\boldsymbol{z}_{j})$. 均值场变分推理中自由分布的取值空间为

$$\mathcal{M}=\left\{q(\boldsymbol{z})\,\middle|\,0\leqslant q(\boldsymbol{z})\leqslant 1,q(\boldsymbol{z})=\prod_{j=1}^{J}q_{j}(\boldsymbol{z}_{j})\right\}.$$

均值场变分优化问题可表示为

$$q^{*}(\boldsymbol{z})=\arg\max_{q(\boldsymbol{z})\in\mathcal{M}}\mathcal{L}(q(\boldsymbol{z})). \tag{9}$$

对于均值场变分推理,可以采用坐标上升算法求解变分优化问题.该算法的基本思想是,对于变分分布的 J 个分量,每次迭代优化其中的一个分量而固定其他分量.对于分量 $\boldsymbol{z}_{j}$,$j=1,2,\cdots,J$,通过坐标上升算法可以计算出分量的迭代式为

$$q_{j}^{*}(\boldsymbol{z}_{j})\propto\int\ln p(\boldsymbol{x},\boldsymbol{z})\prod_{k\neq j}q_{k}(\boldsymbol{z}_{k})\mathrm{d}\boldsymbol{z}_{k}. \tag{10}$$

基于坐标上升的均值场变分推理算法优化过程为:首先初始化均值场各分量变分分布 $q_{j}(\boldsymbol{z}_{j})$;然后对每个分量采用迭代式(10)进行迭代更新,直到收敛;最后输出变分分布 $q(\boldsymbol{z})=\prod_{j=1}^{J}q_{j}(\boldsymbol{z}_{j})$,即隐变量后验概率分布的近似分布.

均值场变分推理中只假设隐向量分量是边缘独立的,没有设定其他假设,包括变分分布的表示形式等.基于坐标上升的均值场变分推理实现简单,研究者进一步提出变分消息传播算法[54],具体地,每个随机隐向量的变分参数利用其 Markov 毯邻域内隐变量的变分参数进行迭代更新.变分消息传播更新方式结合了变分推理与概率图模型,通过因子图给出了更一般表示形式[49],并扩展到了非共轭模型[55].

2) 基于随机梯度的优化方法

对于概率生成模型的变分优化式(8),如果变分分布 $q(\boldsymbol{z}^{(i)})$的结构形式已知,如高斯分布、基于神经网络结构的高斯分布等,此时变分分布可表示为 $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)})$,其中 $\boldsymbol{\phi}$ 表示变分分布的参数.此时可以采用梯度上升方法求解变分参数 $\boldsymbol{\phi}$,即:

$$\boldsymbol{\phi}\leftarrow\boldsymbol{\phi}+\alpha\ \nabla_{\boldsymbol{\phi}}\mathcal{L}(q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)})),$$

其中,α 表示更新步长.此时变分推理的核心是如何有效计算目标函数针对变分参数的梯度$\nabla_{\boldsymbol{\phi}}\mathcal{L}(q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}))$[13].

2.3 基于变分推理的模型参数学习

对于概率生成模型 $p(\boldsymbol{x},\boldsymbol{z})$,通过最大化数据集的对数似然函数 $\ln p(X)$,计算模型参数 $\boldsymbol{\theta}$ 是概率生成模型的重要学习任务[56].期望最大化(expectation maximization, EM)算法是求解含有隐向量模型参数的基本方法,通过迭代执行 E 步(expectation step)和 M 步(maximization step)计算模型参数,其中 E 步是利用上一步的参数计算隐向量后验概率分布,M 步是利用隐向量知识求解模型参数.变分推理与 EM 算法有着天然密切的联系[1],当 E 步无法精确计算隐向量后验概率分布时,需要借助变分推理进行近似求解.

对于概率生成模型 $p(\boldsymbol{x},\boldsymbol{z})$,如果模型参数 $\boldsymbol{\theta}$ 为未知值,此时样本点 $\boldsymbol{x}^{(i)}$ 的变分下界可记为 $\mathcal{L}(q(\boldsymbol{z}^{(i)}),\boldsymbol{\theta})$,说明变分推理模型 $q(\boldsymbol{z}^{(i)})$ 及模型参数 $\boldsymbol{\theta}$ 都是需要求解的未知量.此时,数据集 X 的对数似然函数的变分优化问题为

$$\max \sum_{i=1}^{N}\mathcal{L}(q(\boldsymbol{z}^{(i)}),\boldsymbol{\theta}).$$

对于该优化问题,分别关于参数 $\boldsymbol{\theta}$ 和变分分布 $q(\boldsymbol{z}^{(i)})$ 迭代求解直到收敛,可以计算出模型参数 $\boldsymbol{\theta}$ 及自由分布 $q(\boldsymbol{z}^{(i)})$.该求解过程称为变分 EM 算法,其中变分 E 步用于求解变分分布 $q(\boldsymbol{z}^{(i)})$,变分 M 步用于求解模型参数 $\boldsymbol{\theta}$,迭代执行变分 E 步和 M 步直到收敛.变分 EM 算法的具体过程:

变分 E 步为

$$q(\boldsymbol{z}^{(i)})=\arg\max \mathcal{L}(q(\boldsymbol{z}^{(i)}),\boldsymbol{\theta}),i=1,2,\cdots,N;$$

变分 M 步为

$$\boldsymbol{\theta}=\arg\max \sum_{i=1}^{N}\mathcal{L}(q(\boldsymbol{z}^{(i)}),\boldsymbol{\theta}).$$

对于贝叶斯概率生成模型 $p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})$,需要计算参数的后验概率分布,在均值场变分假设下,变分分布形式为 $q(\boldsymbol{z},\boldsymbol{\theta})=q(\boldsymbol{z})q(\boldsymbol{\theta})$,观测数据的证据下界为

$$\mathcal{L}(q(\boldsymbol{z}),q(\boldsymbol{\theta}))=E_{q(\boldsymbol{z})q(\boldsymbol{\theta})}\{\ln p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta})-\ln q(\boldsymbol{z},\boldsymbol{\theta})\}.$$

此时通过变分 EM 算法交替更新自由分布 $q(\boldsymbol{z})$,$q(\boldsymbol{\theta})$,具体过程:

变分 E 步为

$$q(\boldsymbol{z}^{(i)})=\arg\max \mathcal{L}(q(\boldsymbol{z}^{(i)}),q(\boldsymbol{\theta})),\ i=1,2,\cdots,N;$$

变分 M 步为

$$q(\boldsymbol{\theta})=\arg\max \sum_{i=1}^{N}\mathcal{L}(q(\boldsymbol{z}^{(i)}),q(\boldsymbol{\theta})).$$

最终得到隐变量后验概率分布的近似分布 $q(\boldsymbol{z})$,及隐变量后验概率分布的近似分布 $q(\boldsymbol{\theta})$.

3 条件共轭指数族下的变分推理

指数族分布是一类重要的概率分布表示形式,包含高斯、伯努利分布、多项式分布等多种分布.条件共轭指数族是指对指数族分布参数引入与似然函数共轭的先验分布,从而使得参数后验分布具有与先验相同的分布形式[4].对于条件共轭指数族分布,基于坐标上升算法可以给出变分优化问题的解析表示形式,并且结合随机优化策略可以将变分推理推广到大规模数据中.

3.1 条件共轭指数族分布

指数族分布下概率生成模型的概率分布为

$$p(\boldsymbol{x},\boldsymbol{z}\mid\boldsymbol{\theta})=h(\boldsymbol{x},\boldsymbol{z})g(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^{\mathrm{T}}u(\boldsymbol{x},\boldsymbol{z})\},\tag{11}$$

其中:$\boldsymbol{\theta}$ 称为自然参数;$u(\boldsymbol{x},\boldsymbol{z})$ 为充分统计量;$g(\boldsymbol{\theta})$ 为归一化项;$h(\boldsymbol{x},\boldsymbol{z})$ 为基测度.参数 $\boldsymbol{\theta}$ 的共轭先验为

$$p(\boldsymbol{\theta};\boldsymbol{\alpha}_1,\alpha_2)=f(\boldsymbol{\alpha}_1,\alpha_2)g(\boldsymbol{\theta})^{\alpha_2}\exp\{\alpha_2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\alpha}_1\},\tag{12}$$

其中,$f(\boldsymbol{\alpha}_1,\alpha_2)$ 表示归一化量.观测数据集 X 的联合概率分布为

$$p(X,Z,\boldsymbol{\theta})=\prod_{i=1}^{N}p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}\mid\boldsymbol{\theta})\ p(\boldsymbol{\theta};\boldsymbol{\alpha}_1,\alpha_2).\tag{13}$$

对隐向量及模型参数引入均值场变分分布 $q(Z,\boldsymbol{\theta})$,即:

$$q(Z,\boldsymbol{\theta})=q(\boldsymbol{\theta})\prod_{i=1}^{N}q(\boldsymbol{z}^{(i)}).\tag{14}$$

条件共轭指数族分布下的变分优化式为

$$\mathcal{L}(q(Z,\boldsymbol{\theta}))=E_{q(\boldsymbol{\theta})}\{\ln p(\boldsymbol{\theta};\boldsymbol{\alpha}_1,\alpha_2)-\ln q(\boldsymbol{\theta})\}+\sum_{i=1}^{N}E_{q(\boldsymbol{z}^{(i)})}\{\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)}\mid\boldsymbol{\theta})-\ln q(\boldsymbol{z}^{(i)})\}.\tag{15}$$

3.2 基于坐标上升的优化问题求解

对条件共轭指数族分布的变分优化式(15),采用坐标上升算法进行优化求解,隐向量的变分分布 $q(\boldsymbol{z}^{(i)})$ 的迭代更新式为

$$q(\boldsymbol{z}^{(i)})=h(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})g(E[\boldsymbol{\theta}])\exp\{E_{q(\boldsymbol{\theta})}[\boldsymbol{\theta}^{\mathrm{T}}]\times u(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})\},\ i=1,2,\cdots,N.\tag{16}$$

参数的变分分布 $q(\boldsymbol{\theta})$ 的迭代更新式为

$$q(\boldsymbol{\theta};\boldsymbol{\lambda}_1,\lambda_2)=f(\boldsymbol{\lambda}_1,\lambda_2)g(\boldsymbol{\theta})^{\lambda_2}\exp\{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\lambda}_1\},\tag{17}$$

其中:

$$\boldsymbol{\lambda}_1=\boldsymbol{\alpha}_1+\sum_{i=1}^{N}E_{q(\boldsymbol{z}^{(i)})}[u(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})];$$

$$\lambda_2=\alpha_2+N.$$

对于条件共轭指数族分布,在引入隐变量自由分布过程中,不需要给出自由分布的具体形式,利用坐标上升方法可以直接给出自由分布的分布形式及参数更新式.对比式(12)(17)可知,参数变分分布 $q(\boldsymbol{\theta};\boldsymbol{\lambda}_1,\lambda_2)$ 与先验分布 $p(\boldsymbol{\theta};\boldsymbol{\alpha}_1,\alpha_2)$ 具有相同的分布表示形式.通过迭代执行局部隐变量变分分布更新式(16)和全局参数变分分布更新式(17),直到算法收敛,计算出目标函数.对于条件共轭指数族分布,基于坐标上升的变分推理的执行过程如算法 1 所示:

算法 1. 基于坐标上升的变分推理.

输入:数据集 X、模型 $p(X,Z,\boldsymbol{\theta})$、变分分布 $q(Z,\boldsymbol{\theta})$;

输出:自由分布 $q(\boldsymbol{\theta};\boldsymbol{\lambda})$,$q(\boldsymbol{z}^{(i)})$,$i=1,2,\cdots,N$.

repeat:

for 数据点 $\boldsymbol{x}^{(i)}\in X$

根据式(16)更新变分分布 $q(\boldsymbol{z}^{(i)})$;

end for

根据式(17)更新自由分布 $q(\boldsymbol{\theta})$;

until 满足收敛条件.

3.3 随机化变分推理

对于条件共轭指数族分布,采用基于坐标上升的变分推理进行模型训练,每次需要遍历整个数据集才能更新一次模型参数,模型训练效率不高,较难推广至大规模数据集上.针对该问题,Hoffman 教授团队[12]提出了随机化变分推理方法,将随机优化技术应用到变分优化目标,并利用指数族自然梯度的性质,得到迭代更新式,使变分推理可以应用于大规模数据集[12,57].

针对条件共轭指数族分布的变分下界 $\mathcal{L}(q(Z,\boldsymbol{\theta}))$,如式(15)所示,对数据集 X 进行均匀随机采样得到批处理子集 X^S,可以构建该变分下界的随机估计,即:

$$\hat{\mathcal{L}}(q(Z,\boldsymbol{\theta}))=E_{q(\boldsymbol{\theta};\boldsymbol{\lambda})}\{\ln p(\boldsymbol{\theta};\boldsymbol{\alpha}_1,\alpha_2)-\ln q(\boldsymbol{\theta};\boldsymbol{\lambda})\}+\frac{N}{S}\sum_{j=1}^{S}E_{q(\boldsymbol{z}^{(j)})}\{\ln p(\boldsymbol{x}^{(j)},\boldsymbol{z}^{(j)}\mid\boldsymbol{\theta})-\ln q(\boldsymbol{z}^{(j)})\}. \tag{18}$$

根据条件共轭指数族的自然梯度的性质,详细公式推导可参考文献[12],可求解该变分优化式关于超参 $\boldsymbol{\lambda}$ 的带噪无偏自然梯度为

$$\hat{\nabla}_{\boldsymbol{\lambda}}^{\mathrm{nat}}\mathcal{L}=[\boldsymbol{\alpha}_1,\alpha_2]^{\mathrm{T}}+\frac{N}{S}\sum_{j=1}^{S}[E_{q(\boldsymbol{z}^{(j)})}[u(\boldsymbol{x}^{(j)},\boldsymbol{z}^{(j)})],1]^{\mathrm{T}}-\boldsymbol{\lambda}.$$

再根据随机梯度上升方法得到自由分布超参数的迭代更新式,即:

$$\boldsymbol{\lambda}\leftarrow\boldsymbol{\lambda}+\rho_t\hat{\nabla}_{\boldsymbol{\lambda}}^{\mathrm{nat}}\mathcal{L}, \tag{19}$$

其中,ρ_t 表示第 t 迭代时的学习率,需满足 Robbins-Monro 条件,即 $\sum_{t=1}^{\infty}\rho_t=\infty$,$\sum_{t=1}^{\infty}\rho_t^2<\infty$.

随机化变分推理中的批处理数据集 X^S,其中 S 表示批处理的规模,一般满足 $1\leqslant S\ll N$.当 S 值较大时可以降低随机自然梯度的方差,但是为了更容易扩展到大规模数据集,必须让批量处理规模远远小于数据集规模,即 $S\ll N$.随机化变分推理算法更新过程如算法 2 所示:

算法 2. 随机化变分推理.

输入:数据集 X、模型 $p(X,Z,\boldsymbol{\theta})$、变分分布 $q(Z,\boldsymbol{\theta})$;

输出:自由分布 $q(\boldsymbol{\theta};\boldsymbol{\lambda})$.

repeat:

采样 $j\sim\mathrm{Uniform}(1,2,\cdots,N)$;

根据式(16)更新变分分布 $q(\boldsymbol{z}^{(j)})$;

根据式(19)更新变分分布 $q(\boldsymbol{\theta};\boldsymbol{\lambda})$;

$$\boldsymbol{\lambda}\leftarrow\boldsymbol{\lambda}+\rho_t\hat{\nabla}_{\boldsymbol{\lambda}}^{\mathrm{nat}}\mathcal{L};$$

until 满足收敛条件.

在线变分推理(online variational infernce)[57]与随机化变分推理具有相同的参数更新方式,其中随机化变分推理的批处理样本集 X^S 是从样本集 X 中均匀随机采样,在线变分推理假设批量数据集 X^S 是已知的,比如在流应用中,假设某数据源可以顺序产生批处理数据集.随机化变分推理的学习率及批处理大小会影响算法的收敛速度.根据大数定律可知,增加批处理规模,可以降低随机梯度噪音,允许较大的学习率.提高随机化变分推理收敛速度,可以通过固定学习率自适应选择批处理数据规模[58],或者通过固定批处理规模自适应选择学习率[59-60].同时,随机化变分推理的梯度方差决定了算法的收敛速度,较小的梯度方差使算法具有较快的收敛速度.可以通过降低随机梯度方差来提高随机化变分推理的收敛速度,包括通过引入控制变量降低随机梯度方差[61-62],通过非均匀采样,如重要采样(important sampling)[63]、分层采样(stratified sampling)[64]、基于聚类的采样(clustering-based sampling)[65]、多元批采样(diversified mini-batch sampling)[66]等,选择具有较小方差的批处理数据来降低梯度方差.

4 黑盒变分推理

本节针对一般概率生成模型,综述基于随机梯度的黑盒变分推理方法,该方法首先通过计算变分优化式的随机梯度估计,再结合随机梯度更新变分分布参数,具有广泛的应用范围,并可以方便应用到深度复杂模型及大规模数据[13-14].

4.1 黑盒变分推理框架

对于一般的概率生成模型,假设变分分布 $q(\boldsymbol{z})$ 的分布形式是固定的,训练过程只需求解变分分布参数 $\boldsymbol{\phi}$.此时变分推理模型可表示为 $q(\boldsymbol{z};\boldsymbol{\phi})$,样本点 $\boldsymbol{x}^{(i)}$ 的变分下界为

$$\mathcal{L}(\boldsymbol{\phi}^{(i)})=E_{\boldsymbol{\phi}^{(i)}}[\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})-\ln q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})],$$

其中,期望 $E_{\boldsymbol{\phi}^{(i)}}[\cdot]$ 表示关于自由分布 $q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})$ 的期望.数据集 X 的对数似然函数 $\ln p(X)$ 的变分优化问题为

$$\max \sum_{i=1}^{N}\mathcal{L}(\boldsymbol{\phi}^{(i)}).$$

由于变分下界 $\mathcal{L}(\boldsymbol{\phi}^{(i)})$ 一般是期望表示形式,很难直接计算关于变分参数的梯度.黑盒变分推理首先计算变分下界的随机梯度估计 $\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$,再采用梯度上升方法更新变分参数 $\boldsymbol{\phi}^{(i)}\leftarrow\boldsymbol{\phi}^{(i)}+\rho_t\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$.黑盒变分推理的算法描述如算法 3 所示:

算法 3. 黑盒变分推理框架.

输入:X,$p(\boldsymbol{x},\boldsymbol{z})$;

输出:自由分布 $q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})$,$i\in\{1,2,\cdots,N\}$.

初始化参数 $\boldsymbol{\phi}^{(i)}$,$i\in\{1,2,\cdots,N\}$;

repeat:

学习率 ρ_t;

for $\boldsymbol{x}^{(i)}\in X$

$\boldsymbol{\phi}^{(i)}\leftarrow\boldsymbol{\phi}^{(i)}+\rho_t\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$;

end for

until 满足收敛条件.

其中,ρ_t 表示满足 Robbins-Monro 条件的学习率.黑盒变分推理的核心是计算变分下界的无偏随机梯度 $\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$.

随机梯度估计方法主要是基于蒙特卡洛的估计方法,典型方法包括基于评分函数的随机梯度[13]和基于重参的随机梯度[15-16].随机梯度的噪音或方差控制是这类方法的难点,会直接影响算法的收敛及收敛速度.基于随机梯度下降的方法对步长很敏感,步长太大会使算法不收敛,步长太小会使算法收敛速度特别慢.针对这些问题研究者提出各种随机优化框架,如 Adagrad[67],Adam[68],RMDProg[69] 等,可以根据当前或过去的梯度值自适应地确定更新步长.

4.2 基于评分函数的随机梯度

对于变分下界 $\mathcal{L}(\boldsymbol{\phi}^{(i)})$,利用梯度性质 $\nabla_{\boldsymbol{\phi}}q(\boldsymbol{z};\boldsymbol{\phi})=q(\boldsymbol{z})\nabla_{\boldsymbol{\phi}}\ln q(\boldsymbol{z};\boldsymbol{\phi})$,可以计算出变分下界 $\mathcal{L}(\boldsymbol{\phi}^{(i)})$ 关于参数 $\boldsymbol{\phi}^{(i)}$ 的梯度为[13]

$$\nabla_{\boldsymbol{\phi}^{(i)}}\mathcal{L}(\boldsymbol{\phi}^{(i)})=E_{\boldsymbol{\phi}^{(i)}}[(\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})-\ln q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)}))\nabla_{\boldsymbol{\phi}^{(i)}}\ln q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})].$$

梯度 $\nabla_{\boldsymbol{\phi}^{(i)}}\mathcal{L}(\boldsymbol{\phi}^{(i)})$ 是期望表示形式,无法进行精确求解,采用蒙特卡洛采样可以得到变分下界的带噪无偏随机梯度估计:

$$\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})=\frac{1}{L}\sum_{l=1}^{L}(\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i,l)})-\ln q(\boldsymbol{z}^{(i,l)};\boldsymbol{\phi}^{(i)}))\nabla_{\boldsymbol{\phi}}\ln q(\boldsymbol{z}^{(i,l)};\boldsymbol{\phi}^{(i)}),$$

其中,$\boldsymbol{z}^{(i,l)}\sim q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})$ 表示从变分分布 $q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})$ 进行的蒙特卡洛采样.在统计学中,梯度 $\nabla_{\boldsymbol{\phi}^{(i)}}\ln q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})$ 称为评分函数梯度,又称似然比(likelihood ratio)或强化梯度(reinforce gradient)[70].

基于评分函数的随机梯度估计 $\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$ 只需要计算对于对数变分分布的梯度,对自由分布结构形式没有其他限制条件,可以应用于贝叶斯深层神经网络、复杂的层次化结构等深度复杂模型.但是随机梯度估计 $\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$ 的方差较大,这会导致收敛速度较慢,一般通过 Rao-Blackwellization 和控制变量法(control variates)来降低方差[13],其中评分函数控制变量是最重要的控制变量法[13].相关研究进一步指出好的控制变量的选择依赖于模型结构,并提出局部期望梯度方法来降低期望方差[18].进一步通过引入重要采样进行非均匀采样降低梯度方差[19].

4.3 基于重参的随机梯度

基于重参的随机梯度首先对变分分布进行变量变换,此时变分分布由噪音分布经过带参数的确定性函数转换得到,然后对重参后的目标函数求梯度[15-16].对于变分分布 $q(\boldsymbol{z};\boldsymbol{\phi})$,通过引入噪音分布 $\boldsymbol{\varepsilon}\sim p(\boldsymbol{\varepsilon})$,并基于参数 $\boldsymbol{\phi}$ 设计从 $\boldsymbol{\varepsilon}$ 到 $\boldsymbol{z}$ 的确定性映射函数 S,对变分分布实现重参,即:

$$\boldsymbol{\varepsilon}\sim p(\boldsymbol{\varepsilon}),$$

$$\boldsymbol{z}=S(\boldsymbol{\varepsilon};\boldsymbol{\phi}),$$

其中:$p(\boldsymbol{\varepsilon})$表示与参数 $\boldsymbol{\phi}$ 无关的噪音分布.例如对于高斯自由分布,若 $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu},\sigma^2\boldsymbol{I})$,则重参过程[15-16]为 $\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{\varepsilon};\boldsymbol{0},\boldsymbol{I})$,$\boldsymbol{z}=\boldsymbol{\mu}+\sigma\otimes\boldsymbol{\varepsilon}$,其中,符号$\otimes$表示点乘操作.

利用重参映射函数对变分优化式进行变量替换,并通过蒙特卡洛采样,可计算出变分下界的带噪无偏估计:

$$\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})=\frac{1}{L}\sum_{l=1}^{L}[\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i,l)};\boldsymbol{\theta})-q(\boldsymbol{z}^{(i,l)})],$$

其中:L 表示采样的个数;$\boldsymbol{z}^{(i,l)}$表示通过映射函数 S 获得的样本,即:

$$\boldsymbol{z}^{(i,l)}=S(\boldsymbol{\varepsilon}^{(l)};\boldsymbol{\phi}^{(i)}),\boldsymbol{\varepsilon}^{(l)}\sim p(\boldsymbol{\varepsilon}),$$

其中:$\boldsymbol{\varepsilon}^{(l)}$表示从自由分布 $p(\boldsymbol{\varepsilon})$进行的蒙特卡洛采样.此时再计算该无偏估计 $\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})$的梯度,得到变分优化式的带噪无偏随机梯度估计:

$$\nabla_{\boldsymbol{\phi}^{(i)}}\hat{\mathcal{L}}(\boldsymbol{\phi}^{(i)})=\frac{1}{L}\sum_{l=1}^{L}[\nabla_{\boldsymbol{z}^{(i,l)}}(\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i,l)})-\ln q(\boldsymbol{z}^{(i,l)}))\nabla_{\boldsymbol{\phi}^{(i)}}\boldsymbol{z}^{(i,l)}],$$

该梯度称为基于重参的随机梯度.

相比基于评分函数的随机梯度,基于重参的随机梯度在实践中具有较小的梯度方差,但是并没有严格的理论证明[20],研究者开展了特定结构下重参梯度的方差分析研究[46,71].基于重参的方法需要找到一个从噪音向量 $\boldsymbol{\varepsilon}$ 到隐向量 $\boldsymbol{z}$ 之间的映射 S,重参方法适用于 3 类概率分布:1)位置尺度类概率分布;2)具有可逆累计分布函数的概率分布;3)通过变换可以表示成上述 2 类的概率分布分布[15].对于不满足上述条件的连续隐向量,研究者提出隐式重参梯度(implicit reparameterization gradients)方法,利用从 $\boldsymbol{z}$ 到 $\boldsymbol{\varepsilon}$ 累计分布函数的梯度,结合链式规则求变分下界的随机梯度[23].对于离散隐向量,研究者采用 gumbel-max 技巧,并利用 softmax 操作代替 argmax 操作,实现离散变量重参[21-22].

4.4 分摊变分推理方法

在第 3 节和 4.1~4.3 节变分推理的变分分布 $q(\boldsymbol{z}^{(i)};\boldsymbol{\phi}^{(i)})$中,每个隐向量 $\boldsymbol{z}^{(i)}$ 都有自己的变分参数 $\boldsymbol{\phi}^{(i)}$.分摊变分推理利用参数化函数,使所有隐向量 $\{\boldsymbol{z}^{(i)}\}_{i=1}^{N}$的变分分布都共享相同模型参数[24].对于单样本 $\boldsymbol{x}^{(i)}$,分摊变分推理的变分下界可表示为

$$\mathcal{L}(\boldsymbol{\phi})=E_{\boldsymbol{\phi}}[\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)})-\ln q(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)};f_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)}))],$$

其中,$f_{\boldsymbol{\phi}}(\boldsymbol{x})$表示基于样本的参数化函数;$\boldsymbol{\phi}$ 表示变分分布模型的共享参数.分摊变分推理的共享变分参数训练是基于所有样本累计训练的结果,相比传统每个样本独立的变分参数方式,该训练方式又称为有记忆的训练模式.传统变分推理中独立变分参数如图 2(a)所示,分摊变分推理中分摊变分推理参数如图 2(b)所示,其中,实线箭头表示生成过程,虚线箭头表示推理过程.对于分摊变分推理对应的优化问题,可以采用基于评分函数的随机梯度或基于重参的随机梯度进行迭代求解.

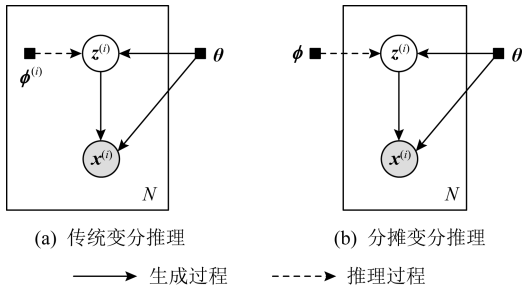

Fig. 2 The graphical models of two inferences

图 2 2 种推理的图模型结构

4.5 范例:变分自编码模型

变分自编码模型由 Kingma[15] 和 Rezende[16] 这 2 个研究团队于 2014 年分别独立提出的,是当前最经典的深度生成模型之一,它有效结合了概率生成模型和深层神经网络.具体地,变分自编码的生成模型 $p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$结合了深层神经网络,变分推理模型采用分摊方式 $q(\boldsymbol{z}|\boldsymbol{x};f_{\boldsymbol{\phi}}(\boldsymbol{x}))$,其中函数 $f_{\boldsymbol{\phi}}(\boldsymbol{x})$也采用深层神经网络.在模型训练中,采用基于重参的随机梯度对生成模型参数 $\boldsymbol{\theta}$ 及变分推理模型参数 $\boldsymbol{\phi}$ 同时进行训练学习.

对于单样本 $\boldsymbol{x}^{(i)}$,变分自编码的变分下界为

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta};\boldsymbol{x}^{(i)})=E_{\boldsymbol{\phi}}[\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i)};\boldsymbol{\theta})-\ln q(\boldsymbol{z}^{(i)};f_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)}))],$$

其中,$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta};\boldsymbol{x}^{(i)})$表示此时的未知量包括变分参数 $\boldsymbol{\phi}$ 和模型参数 $\boldsymbol{\theta}$.利用重参映射可计算变分下界的带噪无偏估计,即:

$$\hat{\mathcal{L}}(\boldsymbol{\phi},\boldsymbol{\theta};\boldsymbol{x}^{(i)})=\frac{1}{L}\sum_{l=1}^{L}[\ln p(\boldsymbol{x}^{(i)},\boldsymbol{z}^{(i,l)};\boldsymbol{\theta})-q(\boldsymbol{z}^{(i,l)};f_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)}))],$$

其中:$\boldsymbol{z}^{(i,l)}$表示根据蒙特卡洛采样及映射函数得到的样本,即 $\boldsymbol{z}^{(i,l)}=S(\boldsymbol{\varepsilon}^{(l)};\boldsymbol{\phi}),\boldsymbol{\varepsilon}^{(l)}\sim p(\boldsymbol{\varepsilon})$.对于数据集 X,基于批处理的数据集 X^R 的变分下界为

$$\hat{\mathcal{L}}^R(\boldsymbol{\phi},\boldsymbol{\theta};X^R)=\frac{N}{R}\sum_{i=1}^{R}\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta};\boldsymbol{x}^{(i)}).$$

此时可以方便计算出变分下界关于变分参数 $\boldsymbol{\phi}$ 和模型参数 $\boldsymbol{\theta}$ 的随机梯度$\nabla_{\boldsymbol{\phi}}\hat{\mathcal{L}}^R(\boldsymbol{\phi},\boldsymbol{\theta};X^R)$和$\nabla_{\theta}\hat{\mathcal{L}}^R(\boldsymbol{\phi},\boldsymbol{\theta};X^R)$,再利用随机梯度上升方法进行参数迭代更新.

变分自编码模型以其简单的训练方式和较好的实践效果很快成为深度生成模型的研究焦点[44].相关研究工作主要集中在丰富变分推理结构、提高变分界的紧致性、提高生成图像质量、研究变分隐空间的解耦特性等方面.研究者提出将隐式分布应用到变分分布中增强变分分布表示能力[25-27].进一步针对隐式变分分布中 KL 散度项难以求解的问题,提出结合对抗网络的判别函数进行优化求解的方法[27,31].为了提高变分下界的紧致度,研究者提出重要加权变分自编码[28-29],通过从变分分布中进行多次采样构建变分下界,采样次数越多,变分界越紧致.同时,研究指出通过丰富生成模型,如引入丰富的隐变量先验分布,提高深度生成模型变分下界的紧致性[72],对先验分布引入高斯混合丰富模型结构[73].针对似然函数在各维度上的分解形式引起的模型性能的降低,提出像素变分自编码[32],通过条件分布保留像素之间的相关性,提高变分界精度.针对变分自编码模型生成图像模糊的问题,很多研究团队开展了研究.其中,英伟达研究团队提出矢量量化的变分自编码(vector quantized variational auto-encoder)[74],将 PixelCNN 引入隐变量先验分布,并利用多尺度层次化建模方法生成高质量图像;进一步对于变分自编码模型,通过使用残差块和批正则化精心设计神经网络结构生成高质量图像,同时保留了变分下界[75].针对变分自编码模型的解耦统计性分析,研究者开展了相关分析,通过加入不同的正则化方法对变分隐空间进行解耦,提升可解释性[76-79].

5 结构化的变分推理

均值场变分推理假设变分分布各分量是边缘独立的,这种假设可以简化推理,但是损失了推理精度.结构化的变分推理是指通过各种方法增加变分分布分量之间的相关性,提高推理精度.本节综述通过各种不同方式增加变分分布分量之间相关度的结构化变分推理方法.

5.1 结构化均值场方法

结构化均值场方法(structured mean field approach)首先由 Saul 和 Jordan 等人提出[11],该方法结合了传统的均值场方法和精确推理方法,首先从概率生成模型中识别出易处理的子结构,如链式结构、树型结构等,然后在各子结构之间执行均值场变分推理方法,而在每个子结构内部执行精确推理,如 junction tree 方法[34].该方法需要根据模型的具体情况人工选取易处理的子结构.如对于概率生成模型,选择的 J 个易处理子结构可记作$\{\boldsymbol{z}_j \mid j=1,2,\cdots,J\}$,此时变分分布表示为

$$q(\boldsymbol{z})=\prod_{j=1}^{J} q_j(\boldsymbol{z}_j),$$

其中,变分子分布$\{q_1(\boldsymbol{z}_1),q_2(\boldsymbol{z}_2),\cdots,q_J(\boldsymbol{z}_J)\}$之间是条件独立的,对于每个子结构 $q_j(\boldsymbol{z}_j)$内可以精确计算单变量边缘概率分布.

之后,结构化均值场方法得到了进一步发展[34,80-81],并结合黑盒变分推理方法应用于大规模数据的更新[82].该方法尤其适合于时间序列模型,如隐 Markov 模型[35]、动态主题模型[36]等,此时采用的变分分布显式保留了时间点之间的结构信息,而其他变量之间仍旧是边缘独立的.

5.2 标准化流方法

标准化流方法[37]是构建丰富变分分布的一类重要方法,可以给出更紧致的变分界.该类方法通过一系列可逆函数将一个简单的变分分布,如均值场变分分布结构,转换成结构更丰富的变分分布.

令 $\boldsymbol{z}$ 表示连续随机向量,其概率分布为 $q(\boldsymbol{z})$,$f:\mathbb{R}^d\to\mathbb{R}^d$ 表示映射函数,通过 f 对隐向量 $\boldsymbol{z}$ 进行变量变换 $\boldsymbol{z}'=f(\boldsymbol{z})$,变换后的随机向量 $\boldsymbol{z}'$ 的概率分布为

$$q(\boldsymbol{z}')=q(\boldsymbol{z})\left|\det\frac{\partial f^{-1}}{\partial \boldsymbol{z}'}\right|=q(\boldsymbol{z})\left|\det\frac{\partial f}{\partial \boldsymbol{z}}\right|^{-1},$$

其中,$\frac{\partial f}{\partial \boldsymbol{z}}$表示函数的雅可比矩阵,操作 det 表示行列式.利用映射函数的复合操作,可以对概率分布进行多次映射,如对初始随机向量 $\boldsymbol{z}_0$,其概率分布为 $q_0(\boldsymbol{z}_0)$,利用 K 个可逆函数 $f_1,f_2,\cdots,f_K$ 进行 K 次转换,得到的随机向量 $\boldsymbol{z}_K$ 及对应的概率分布 $q_K(\boldsymbol{z}_K)$为

$$\boldsymbol{z}_K=f_K\circ\cdots\circ f_2\circ f_1(\boldsymbol{z}_0),$$

$$\ln q_K(\boldsymbol{z}_K)=\ln q_0(\boldsymbol{z}_0)-\sum_{k=1}^{K}\ln\left|\det\frac{\partial f_k}{\partial \boldsymbol{z}_{k-1}}\right|^{-1}.$$

利用变换后的分布作为变分分布,即 $q(\boldsymbol{z}^{(i)}))=q_K(\boldsymbol{z}_K)$,此时的变分下界为

$$L(\boldsymbol{\phi})=E_{q_0(\boldsymbol{z}_0)}[\ln p(\boldsymbol{x},\boldsymbol{z}_K)-\ln q_K(\boldsymbol{z}_K)].$$

求解该优化问题只需要计算映射函数雅可比行列式行列式$\left|\frac{\partial f}{\partial \boldsymbol{z}}\right|$.

选用不同的映射函数可以得到不同的标准化流方法,平面流和径向流是 2 种最经典的标准化流,其中每次映射可以看作一个隐层单元,其雅可比行列式可以很方便计算,一般应用于低维隐空间[37].Sylvester 标准化流是平面流的一般化表示,克服了

单一隐层的缺点,丰富了变分分布的表示能力[83].自回归流是一种结构丰富同时雅可比行列式容易计算的映射函数,成为标准化流的研究特点之一,不同的结合方式形成不同的算法,包括 real non-volume preserving flows[84]、掩模自回归流[38]、可逆自回归流[39]等.

5.3 其他结构化变分推理

除了结构化均值场方法、标准化流方法之外,研究者还提出了各种其他策略,通过丰富变分分布结构提升变分推理精度.

1) 层次化变分模型.均值场方法假设变分推理模型各维度之间是相互独立的,层次化变分模型[17](hierachical variational models)通过引入贝叶斯方法丰富模型结构,对变分参数引入先验分布,增加变量之间的相关性.该方法在保持隐变量之间条件独立性的同时,丰富了变分推理模型结构关系.对于观测向量 $\boldsymbol{x}$,基于完全均值场变分推理中变分分布为

$$q(\boldsymbol{z};\boldsymbol{\phi})=\prod_{i=1}^{d}q(\boldsymbol{z}_i;\boldsymbol{\phi}_i).$$

层次化变分推理通过对变分参数 $\boldsymbol{\phi}$ 引入先验分布 $q(\boldsymbol{\phi};\boldsymbol{\lambda})$,此时层次化变分推理模型为

$$q_{\mathrm{HVM}}(\boldsymbol{z};\boldsymbol{\phi})=\int q(\boldsymbol{\phi};\boldsymbol{\lambda})\prod_{i=1}^{d}q(z_i;\boldsymbol{\phi}_i)\mathrm{d}\boldsymbol{\lambda}.$$

2) 耦合变分推理.耦合变分推理[40-41]通过对均值场变分分布引入耦合分布增加分量之间的相关性,此时变分分布形式为

$$q(\boldsymbol{z};\boldsymbol{\phi})=\Big[\prod_{i=1}^{d}q(z_i;\boldsymbol{\phi}_i)\Big]c(Q(z_1),Q(z_2),\cdots,Q(z_d)),$$

其中,c 表示耦合函数,即基于各分量边缘累计分布函数 $Q(z_1),Q(z_2),\cdots,Q(z_d)$ 的联合概率分布.

3) 基于辅助变量的变分推理模型.在隐变量生成模型中引入附加向量 $\boldsymbol{a}$,此时生成模型概率分布形式为

$$p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{a})=p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{z}).$$

变分推理模型的形式为

$$q(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})=q(\boldsymbol{z}|\boldsymbol{a},\boldsymbol{x})q(\boldsymbol{a}|\boldsymbol{x}).$$

此时,观测样本 $\boldsymbol{x}$ 对应的变分下界为

$$\mathcal{L}(\boldsymbol{\phi})=E_{q(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})}[\ln p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{a})-\ln q(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})].$$

4) 基于混合分布的变分推理.混合模型具有灵活的表示形式,从 20 世纪末开始已经被应用到变分推理模型中[42-43].混合模型灵活的结构形式也意味着复杂的变分推理,研究者提出了各种训练方式,包括基于辅助界的方法[17]、固定点更新方法[85]等.受到 boosting 方法的启发,最近发展了 boosting 变分推理方法[86]及变分 boosting 方法[87],通过每次只更新其中一个分量,而固定其他分量的方式进行训练.

6 研究展望

概率生成模型变分推理已经成为人工智能领域的研究热点,该方向进一步研究工作主要包括 4 个方面:

变分推理的理论研究工作还非常有限,已开展的部分理论研究工作主要集中在变分推理统计性质分析方面,包括对变分贝叶斯推理中模型参数的训练一致性分析[88],如贝叶斯线性模型参数训练一致性的分析[89-90]、泊松混合效应模型参数渐进正态性的分析[91]、随机区块模型参数渐进正态性的分析[92]等.相比蒙特卡洛近似推理理论研究,变分近似推理的理论研究还有很多工作有待深入,包括变分推理近似误差的量化、基于变分推理的预测分布的统计性质等.

变分推理通过概率分布距离度量来构建变分优化式,根据不同的距离度量方式可以得到不同的变分优化式,当前变分推理主流是基于 $D_{\mathrm{KL}}(q\|p)$ 散度的变分优化问题开展研究.研究者也开展了基于其他概率分布度量的变分推理研究,包括 $D_{\mathrm{KL}}(p\|q)$ 散度距离[48,93]、α-散度距离 $D_\alpha(p\|q)$[94]、f-散度距离 $D_f(p\|q)$[53,95]等.但是这方面研究的深度不够,包括如何基于这些概率分布距离度量设计收敛速度更快的优化算法、如何丰富变分分布结构、如何有效地同深度神经网络相结合处理大规模数据等,这些研究问题都是有待深入探讨分析的.

基于采样和基于变分的推理方法是两大类重要的近似推理方法,基于采样的近似推理方法,精度较高且有理论保证,但是收敛速度慢,且受先验参数影响;基于变分的近似推理方法收敛速度快,但是推理精确不易量化.研究如何将 2 类近似方法进行有效地结合,实现近似推理精度与计算速度的权衡折中是一个重要的研究方向.已开展的相关研究包括:将 Metropolis-Hastings 采样引入已训练的变分分布[96]、利用 MCMC 进行近似求解变分推理的坐标上升优化成果[97-98]、引入变分近似方法到 MCMC 链中[99]等.进一步深入研究采样方法和变分方法的结合方式及结合场景对机器学习领域都有重要的理论价值和实践意义.

深度概率生成模型结合了概率生成模型和深度神经网络,是深度模型的重要构成部分.得益于黑盒

变分推理方法的发展,可以方便地开展深度概率生成模型在大规模数据集上的训练.基于变分推理的深度概率生成模型也是深度学习方面的研究热点.随着大规模数据集规模及算力的进步,深度学习在实际应用中取得了很多瞩目成果,但是理论研究方面发展缓慢.可以以深度概率生成模型为切入点,基于变分推理开展相关理论研究,如深度特征表示、隐变量可解释性等.深度概率生成模型的2类典型范例是变分自编码和生成对抗网络.变分自编码是针对显式深度概率生成模型进行训练,模型具有生成数据及特征表示能力;而生成对抗网络是针对隐式深度生成模型进行训练,模型仅具有生成高质量数据的能力.针对不同的应用场景或模型结构,如何扬长避短将2类方法有效结合起来是一个重要的研究方向.已有的相关研究工作包括对抗自编码模型[30]、对抗变分贝叶斯[31]等,是将对抗策略引入到变分优化式的求解中.但是这方面仍旧有很大的研究空间.

7 总 结

本文给出了概率生成模型变分推理的一般框架及基于变分推理的概率生成模型参数学习过程.并从条件共轭指数族的变分推理、基于随机梯度的黑盒变分推理及结构化变分推理3方面综述了变分推理的最新进展及相应框架下算法的优缺点.最终对概率生成模型变分推理的进一步工作进行了讨论分析.

作者贡献声明:陈亚瑞负责论文整体思路框架设计、撰写及修改;杨巨成对论文框架提出指导意见;史艳翠与王嫄负责相关研究现状的补充、研究展望的完善;赵婷婷对论文框架提出指导意见并修改论文.

参 考 文 献

[1] Goodfellow I, Bengio Y, Courville A. Deep Learning [M]. Cambridge, MA: MIT Press, 2016

[2] Goodfellow I. Tutorial: Generative adversarial networks [J]. arXiv preprint, arXiv:1701.00160, 2016

[3] Bishop C M. Pattern Recognition and Machine Learning [M]. Berlin: Springer, 2016

[4] Wainwright M J, Jordan M I. Graphical Models, Exponential Families, and Variational Inference [M]. Hanover, MA: Now Publishers Inc, 2008

[5] Liu Jun. Monte Carlo Strategies in Scientific Computing [M]. Berlin: Springer, 2008

[6] Robert C P, Castlla G. Monte Carlo Statistical Methods [M]. Berlin: Springer, 1999

[7] 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

[8] Peterson C, Anderson J R. A mean field theory learning algorithm for neural networks [J]. Complex Systems, 1987, 1: 995-1019

[9] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An introduction to variational methods for graphical models [J]. Machine Learning, 1999, 37(2): 183-233

[10] Parisi G. Statistical Field Theory [M]. New York: Perseus Books, 1998

[11] Saul L K, Jaakkola T, Jordan M I. Mean field theory for sigmoid belief networks [J]. Journal of Artificial Intelligence Research, 1996, 4(1): 61-76

[12] Hoffman M, Blei D M, Wang Chong, et al. Stochastic variational inference [J]. Journal of Machine Learning Research, 2013, 14: 1303-1347

[13] Ranganath R, Gerrish S, Blei D M. Black box variational inference [C/OL] //Proc of the 17th Int Conf on Artificial Intelligence and Statistics (AISTATS). 2014 [2020-01-25]. http://proceedings.mlr.press/v33/ranganath14.pdf

[14] Paisley J, Blei D M, Jordan M I. Variational Bayesian inference with stochastic search [C/OL] //Proc of the 29th Int Conf on Machine Learning (ICML). Madison, WI: Omnipress, 2012 [2020-05-04]. https://arxiv.org/ftp/arxiv/papers/1206/1206.6430.pdf

[15] Kingma D P, Welling M. Auto-encoding variational Bayes [C/OL] //Proc of the 2nd Int Conf on Learning Representations (ICLR). 2014 [2020-05-04]. https://iclr.cc/archive/2014/conference-proceedings

[16] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models [C/OL] //Proc of the 31st Int Conf on Machine Learning (ICML). 2014 [2020-05-04]. http://proceedings.mlr.press/v32/rezende14.pdf

[17] Ranganath R, Tran D, Blei D M. Hierarchical variational models [C/OL] //Proc of the 33rd Int Conf on Machine Learning (ICML). 2016 [2020-05-04]. http://proceedings.mlr.press/v48/ranganath16.pdf

[18] Titsias M, Lázaro-Gredilla M. Local expectation gradients for black box variational inference [C/OL] //Advances in Neural Information Processing Systems 28. Red Hook, NY: Curran Associates, Inc, 2015 [2020-05-04]. https://proceedings.neurips.cc/paper/2015/file/1373b284bc381890049e92d324f56de0-Paper.pdf

[19] Ruiz F J R, Titsias M K, Blei D M. Overdispersed black-box variational inference [C/OL] //Proc of the 32nd Conf on Uncertainty in Artificial Intelligence (UAI). 2016 [2020-05-04]. http://www.auai.org/uai2016/proceedings/papers/15.pdf

[20] Gal Y. Uncertainty in deep learning [D]. Cambridge, UK: University of Cambridge, 2016

[21] Jang E, Gu Shixiang, Poole B. Categorical reparameterization with gumbel-softmax [C/OL] //Proc of the 5th Int Conf on Learning Representations (ICLR). 2017 [2020-05-04]. https://research.google/pubs/pub45822

[22] Maddison C J, Mnih A, Teh Y W. The concrete distribution: A continuous relaxation of discrete random variables [J]. arXiv preprint, arXiv:1611.00712, 2016

[23] Figurnov M, Mohamed S, Mnih A. Implicit reparameterization gradients [C] //Advances in Neural Information Processing Systems 31. North Miami Beach, FL: Curran Associates, Inc, 2018: 441-452

[24] Marino J, Yue Yisong, Mandt S. Iterative amortized inference [C/OL] //Proc of the 35th Int Conf on Machine Learning (ICML). 2018 [2020-05-04]. http://proceedings.mlr.press/v80/marino18a/marino18a.pdf

[25] Huszár F. Variational inference using implicit distributions [J]. arXiv preprint, arXiv:1702.08235, 2017

[26] Karaletsos T. Adversarial message passing for graphical models [J]. arXiv preprint, arXiv:1612.05048, 2016

[27] Li Yingzhen, Liu Qiang. Wild variational approximations [C/OL] //Proc of the 30th Int Conf on Neural Information Processing System: Approximate Inference Workshops. 2016 [2020-05-04]. http://approximateinference.org/accepted/LiLiu2016.pdf

[28] Burda Y, Grosse R, Salakhutdinov R. Importance weighted autoencoders [J]. arXiv preprint, arXiv:1509.00519, 2015

[29] Cremer C, Morris Q, Duvenaud D. Reinterpreting importance-weighted autoencoders [C/OL] //Proc of the 5th Int Conf on Learning Representations Workshop(ICLR). 2017 [2020-05-30]. https://paperswithcode.com/paper/reinterpreting-importance-weighted

[30] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders [J]. arXiv preprint, arXiv 1511.05644, 2015

[31] Mescheder L, Nowozin S, Geiger A. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks [C/OL] //Proc of the 34th Int Conf on Machine Learning (ICML). 2017 [2020-05-04]. http://proceedings.mlr.press/v70/mescheder17a/mescheder17a.pdf

[32] Gulrajani I, Kumar K, Ahmed F, et al. PixelVAE: A latent variable model for natural images [C/OL] //Proc of the 5th Int Conf on Learning Representations (ICLR). 2017 [2020-05-30]. https://arxiv.org/abs/1611.05013

[33] Saul L K, Jordan M I. Exploiting tractable substructures in intractable networks [C] //Advances in Neural Information Processing Systems 9. Cambridge, MA: MIT Press, 1995: 486-492

[34] Opper M, Saad D. Advanced Mean Field Methods: Theory and Practice [M]. Cambridge, MA: MIT Press, 2001: 129-160

[35] Ghahramani Z, Jordan M I. Factorial hidden Markov models [C/OL] //Advances in Neural Information Processing Systems 8. Cambridge, MA: MIT Press, 1995 [2020-05-04]. https://proceedings.neurips.cc/paper/1995/file/4588e674d3f0faf985047d4c3f13ed0d-Paper.pdf

[36] Blei D M, Lafferty J D. Dynamic topic models [C/OL] //Proc of the 23rd Int Conf on Machine Learning (ICML). New York: ACM, 2006 [2020-05-04]. https://dl.acm.org/doi/10.1145/1143844.1143859

[37] Rezende D, Mohamed S. Variational inference with normalizing flows [C/OL] //Proc of the 32nd Int Conf on Machine Learning (ICML). 2015 [2020-05-04]. http://proceedings.mlr.press/v37/rezende15.pdf

[38] Papamakarios G, Pavlakou T, Iain M. Masked autoregressive flow for density estimation [C] //Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, Inc, 2017: 2338-2347

[39] Kingma D P, Salimans T, Jozefowicz R, et al. Improving variational autoencoders with inverse autoregressive flow [C] //Advances in Neural Information Processing Systems 29. Red Hook, NY: Curran Associates, Inc, 2016: 4736-4744

[40] Han Shaobo, Liao Xuejun, Dunson D, et al. Variational Gaussian copula inference [C/OL] //Proc of the 19th Int Conf on Artificial Intelligence and Statistics (AISTATS). 2016 [2020-01-25]. http://proceedings.mlr.press/v51/han16.pdf

[41] Tran D, Blei D M, Airoldi E M. Copula variational inference [C] //Advances in Neural Information Processing Systems 28. Red Hook, NY: Curran Associates, Inc, 2015: 3564-3572

[42] Jordan M I. Learning in Graphical Models [M]. Berlin: Springer, 1998: 163-173

[43] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An introduction to variational methods for graphical models [J]. Machine Learning, 1999, 37(2): 183-233

[44] Doersch C. Tutorial on variational autoencoders [J]. arXiv preprint, arXiv:1606.05908, 2016

[45] Tschannen M, Olivier B, Mario L. Recent advances in autoencoder-based representation learning [J]. arXiv preprint, arXiv:1812.05069, 2018

[46] Domke J. Provable gradient variance guarantees for black-box variational inference [C] //Advances in Neural Information Processing Systems 32. Red Hook, NY: Curran Associates, Inc, 2019: 329-338

[47] Barber D. Bayesian reasoning and machine learning [D]. Cambridge, UK: University of Cambridge, 2012

[48] Minka T P. Expectation propagation for approximate Bayesian inference [C/OL] //Proc of the 17th Conf on Uncertainty in Artificial Intelligence(UAI). 2001 [2020-05-04]. https://arxiv.org/ftp/arxiv/papers/1301/1301.2294.pdf

[49] Minka T P. Divergence measures and message passing, MSR-TR-2005-173 [R/OL]. 2005 [2020-05-04]. https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/

[50] Minka T P. Power EP, MSR-TR-2004-149 [R/OL]. 2004 [2020-05-04]. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2004-149.pdf

[51] Hernández-Lobato J M, Li Yingzhen, Rowland M, et al. Black-box α-divergence minimization [J]. arXiv preprint, arXiv:1511.03243, 2015

[52] Bamler R, Zhang Cheng, Opper M, et al. Perturbative black box variational inference [C] //Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, Inc, 2017: 5079-5088

[53] Wan Neng, Li Dapeng, Hovakimyan N. f-divergence variational inference [C] //Advances in Neural Information Processing Systems 33. Red Hook, NY: Curran Associates, Inc, 2020: 17370-17379

[54] Winn J, Bishop C. Variational message passing [J]. Journal of Machine Learning Research, 2005, 6: 661-694

[55] Knowles D, Minka T. Non-conjugate variational message passing for multinomial and binary regression [C] //Advances in Neural Information Processing Systems 24. Red Hook, NY: Curran Associates, Inc, 2011: 1701-1709

[56] Zhu Jun, Hu Wenbo. Recent advances in Bayesian machine learning [J]. Journal of Computer Research and Development, 2015, 52(1): 16-26 (in Chinese)
(朱军, 胡文波. 贝叶斯机器学习前沿进展综述[J]. 计算机研究与发展, 2015, 52(1): 16-26)

[57] Wang Chong, Paisley J, Blei D M. Online variational inference for the hierarchical Dirichlet process [C/OL] //Proc of the 14th Int Conf on Artificial Intelligence and Statistics (AISTATS). 2011 [2020-01-25]. http://proceedings.mlr.press/v15/wang11a/wang11a.pdf

[58] Ranganath R, Wang Chong, Blei D, et al. An adaptive learning rate for stochastic variational inference [C] //Proc of the 30th Int Conf on Machine Learning (ICML). 2013 [2020-05-04]. http://proceedings.mlr.press/v28/ranganath13.pdf

[59] Tan L S L. Stochastic variational inference for large scale discrete choice models using adaptive batch sizes [J]. Statistics and Computing, 2017, 27(1): 237-257

[60] De S, Yadav A, Jacobs D, et al. Automated inference with adaptive batches [C/OL] //Proc of the 20th Int Conf on Artificial Intelligence and Statistics (AISTATS). 2017 [2020-01-25]. http://proceedings.mlr.press/v54/de17a/de17a.pdf

[61] Johnson R, Zhang Tong. Accelerating stochastic gradient descent using predictive variance reduction [C] //Advances in Neural Information Processing Systems 26. Red Hook, NY: Curran Associates, Inc, 2013: 315-323

[62] Wang Chong, Chen Xi, Smola A J, et al. Variance reduction for stochastic gradient optimization [C] //Advances in Neural Information Processing Systems 26. Red Hook, NY: Curran Associates, Inc, 2013: 181-189

[63] Zhao Peilin, Zhang Tong. Stochastic optimization with importance sampling for regularized loss minimization [C/OL] //Proc of the 32nd Int Conf on Machine Learning (ICML). 2015 [2020-05-04]. http://proceedings.mlr.press/v37/zhaoa15.pdf

[64] Zhao Peilin L, Zhang Tong. Accelerating minibatch stochastic gradient descent using stratified sampling [J]. arXiv preprint, arXiv:1405.3080, 2014

[65] Fu Tianfan, Zhang Zhihua. CPSG-MCMC: Clustering-based preprocessing method for stochastic gradient MCMC [C/OL] //Proc of the 20th Int Conf on Artificial Intelligence and Statistics (AISTATS). 2017 [2020-01-25]. http://proceedings.mlr.press/v54/fu17a/fu17a.pdf

[66] Zhang Cheng, Kjellstrom H, Mandt S. Determinantal point processes for mini-batch diversification [C/OL] //Proc of the 33rd Conf on Uncertainty in Artificial Intelligence (UAI). 2017 [2020-05-04]. http://auai.org/uai2017/proceedings/papers/69.pdf

[67] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. Journal of Machine Learning Research, 2011, 12: 2121-2159

[68] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C/OL] //Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, Inc, 2017 [2020-05-04]. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[69] Hinton G, Srivastava N, Swersky K. Neural networks for machine learning: Lecture 6a overview of mini-batch gradient descent [R/OL]. 2012 [2020-05-04]. http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

[70] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. Machine Learning, 1992, 8(3/4): 229-256

[71] Xu Ming, Quiroz M, Kohn R, et al. Variance reduction properties of the reparameterization trick [C/OL] //Proc of the 22nd Int Conf on Artificial Intelligence and Statistics (AISTATS). 2019 [2020-01-25]. http://proceedings.mlr.press/v89/xu19a/xu19a.pdf

[72] Johnson M J, Duvenaud D, Wiltschko A B, et al. Structured VAEs: Composing probabilistic graphical models and variational autoencoders [J]. arXiv preprint, arXiv: 1603.06277v5, 2016

[73] Chen Yarui, Jiang Shuoran, Yang Jucheng, et al. Mixture of variational autoencode [J]. Journal of Computer Research and Development, 2020, 57(1): 136-144 (in Chinese)
(陈亚瑞, 蒋硕然, 杨巨成, 等. 混合变分自编码[J]. 计算机研究与发展, 2020, 57(1): 136-144)

[74] Razavi A, Oord A, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2 [C] //Advances in Neural Information Processing Systems 32. Red Hook, NY: Curran Associates, Inc, 2019: 14837-14847

[75] Vahdat A, Kautz J. NVAE: A deep hierarchical variational autoencoder [C] //Advances in Neural Information Processing Systems 33. Red Hook, NY: Curran Associates, Inc, 2020: 19667-19679

[76] Kumar A, Poole B. On implicit regularization in β-VAEs [C/OL] //Proc of the 37th Int Conf on Machine Learning (ICML). 2020 [2021-02-14]. http://proceedings.mlr.press/v119/kumar20d/kumar20d.pdf

[77] Mathieu E, Rainforth T, Siddharth N, et al. Disentangling disentanglement in variational autoencoders [C/OL] //Proc of the 36th Int Conf on Machine Learning (ICML). 2019 [2020-04-05]. http://proceedings.mlr.press/v97/mathieu19a/mathieu19a.pdf

[78] Kim H, Mnih A. Disentangling by factorising [C/OL] //Proc of the 35th Int Conf on Machine Learning (ICML). 2018 [2020-05-04]. http://proceedings.mlr.press/v80/kim18b/kim18b.pdf

[79] Tschannen M, Bachem O, Lucic M. Recent advances in autoencoder-based representation learning [C/OL] //Proc of the 32nd Conf on Neural Information Processing Systems: Bayesian Deep Learning Workshop. 2018 [2020-05-04]. http://bayesiandeeplearning.org/2018/papers/151.pdf

[80] Wiegerinck W. Variational approximations between mean field theory and the junction tree algorithm [C/OL] //Proc of the 16th Conf on Uncertainty in Artificial Intelligence(UAI). 2000 [2020-05-04]. https://arxiv.org/ftp/arxiv/papers/1301/1301.3901.pdf

[81] Barber D, Wiegerinck W. Tractable variational structures for approximating graphical models [C] //Advances in Neural Information Processing Systems 11. Cambridge, MA: MIT Press, 1999: 183-189

[82] Bamler R, Mandt S. Structured black box variational inference for latent time series models [J]. arXiv preprint, arXiv:1707.01069, 2017

[83] Berg R, Hasenclever L, Tomczak J M, et al. Sylvester normalizing flows for variational inference [C/OL] //Proc of the 34th Conf on Uncertainty in Artificial Intelligence(UAI). 2018 [2020-05-04]. http://auai.org/uai2018/proceedings/papers/156.pdf

[84] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation usingreal NVP [C/OL] //Proc of the 4th Int Conf on Learning Representations (ICLR). 2016 [2020-05-30]. https://arxiv.org/pdf/1605.08803.pdf

[85] Salimans T, Knowles D A. Fixed-form variational posterior approximation through stochastic linear regression [J]. Bayesian Analysis, 2013, 8(4): 837-882

[86] Guo Fangjian, Wang Xiangyu, Fan Kai, et al. Boosting variational inference [J]. arXiv preprint, arXiv:1611.05559, 2016

[87] Miller A C, Foti N, Adams R P. Variational boosting: Iteratively refining posterior approximations [C/OL] //Proc of the 34th Int Conf on Machine Learning (ICML). 2017 [2020-05-04]. http://proceedings.mlr.press/v70/miller17a/miller17a.pdf

[88] Wang Yixin, Blei D M. Frequentist consistency of variational Bayes [J]. arXiv preprint, arXiv:1705.03439v2, 2018

[89] You Chong, Ormerod J, Muller S. On variational Bayes estimation and variational information criteria for linear regression models [J]. Australian & New Zealand Journal of Statistics, 2014, 56(1): 73-87

[90] Ormerod J, You Chong, Muller S. A variational Bayes approach to variable selection [J]. Electronic Journal of Statistics, 2017, 11(2): 3549-3594

[91] Hall P, Pham T, Wand M, et al. Asymptotic normality and valid inference for Gaussian variational approximation [J]. Annals of Statistics, 2011, 39(5): 2502-2532

[92] Celisse A, Daudin J J, Pierre L. Consistency of maximum-likelihood and variational estimators in the stochastic block model [J]. Electronic Journal of Statistics, 2012, 6: 1847-1899

[93] Naesseth C A, Lindsten F. Markovian score climbing: Variational inference with $KL(p \parallel q)$ [C] //Advances in Neural Information Processing Systems 33. Red Hook, NY: Curran Associates, Inc, 2020: 15499-15510

[94] Li Yingzhen, Turner R E. Rényi divergence variational inference [J]. arXiv preprint, arXiv:1602.02311, 2016

[95] Zhang Cheng, Butepage J, Kjellstrom H, et al. Advances in variational inference [J]. IEEE Translation Pattern Analysis Machine Intelligence, 2019, 41(8): 2008-2026

[96] Freitas N D, Højen-Sørensen P, Jordan M, et al. Variational MCMC [C/OL] //Proc of the 17th Conf on Uncertainty in Artificial Intelligence (UAI). 2001 [2020-05-04]. https://arxiv.org/ftp/arxiv/papers/1301/1301.2266.pdf

[97] Hoffman M D, Blei D M. Stochastic structured variational inference [C/OL] //Proc of the 18th Int Conf on Artificial Intelligence and Statistics (AISTATS). 2015 [2020-01-25]. http://proceedings.mlr.press/v38/hoffman15.pdf

[98] Mimno D, Hoffman M, Blei D. Sparse stochastic inference for latent Dirichlet allocation [C] //Proc of the 29th Int Conf on Machine Learning (ICML). Madison, WI: Omnipress, 2012

[99] Salimans T, Kingma D, Welling M. Markov chain Monte Carlo and variational inference: Bridging the gap [C/OL] //Proc of the 32nd Int Conf on Machine Learning(ICML). 2015 [2020-05-04]. http://proceedings.mlr.press/v37/salimans15.pdf

Chen Yarui, born in 1982. PhD, professor, MS supervisor. Member of CCF. Her main research interests include probabilistic graphical models, machine learning algorithms, and approximate inference algorithms.

陈亚瑞,1982 年生,博士,教授,硕士生导师.CCF 会员.主要研究方向为概率图模型、机器学习算法及近似推理算法.

Yang Jucheng, born in 1980. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include image processing, biometrics, pattern recognition, and neural networks.

杨巨成,1980 年生.博士,教授,博士生导师.CCF 高级会员.主要研究方向为图像处理、生物识别、模式识别及神经网络.

Shi Yancui, born in 1982. PhD, associate professor, MS supervisor. Member of CCF. Her main research interests include recommendation system, social network, and data mining.

史艳翠,1982 年生.博士,副教授,硕士生导师.CCF 会员.主要研究方向为推荐系统、社会网络及数据挖掘.

Wang Yuan, born in 1989. PhD, associate professor, MS supervisor. Member of CCF. Her main research interests include data mining, machine learning, and natural language processing.

王　媛,1989 年生.博士,副教授,硕士生导师.CCF 会员.主要研究方向为数据挖掘、机器学习及自然语言处理.

Zhao Tingting, born in 1986. PhD, associate professor, MS supervisor. Member of CCF. Her main research interests include machine learning, reinforcement learning and robot control.

赵婷婷,1986 年生.博士,副教授,硕士生导师.CCF 会员.主要研究方向为机器学习、强化学习及机器人控制.