元学习——Meta-Amortized Variational Inference and Learning

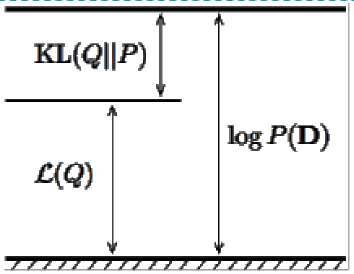
作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

这篇博客是论文“Meta-Amortized Variational Inference and Learning”的阅读笔记。博客中前半部分内容与变分自编码器(VAE)的推导极为类似，所涉及的公式推导如果有不明白的地方，可以提前阅读这篇博客：变分推断与变分自编码器。主要涉及到了概率论的相关知识。这篇博客介绍了精确推理(Exact Inference)，近似变分推理(Approximate Variational Inference)，摊销变分推理(Amortized Variational Inference)，摊销变分自编码器(Amortized Variational Autoencoders)，元摊销变分推理(Meta-Amortized Variational Inference)，以及论文中提出的元变分自编码器(MetaVAE)。主要阅读了论文原理部分，实验部分没有去关注。

尽管最近在概率建模及其应用方面取得了成功，但使用传统推理技术训练的生成模型很难适应新的分布，即使目标分布可能与训练中已见过的分布密切相关。这篇文章提出了一个双重摊销变分推理模型，以解决这一挑战。通过不仅在一组查询输入中共享计算，而且在一组不同但相关的概率模型中共享计算，所提算法学习到了可迁移的潜在表示，这些表示可在多个相关的分布中进行拓展。特别地，给定一组在图像上的分布，所提算法找到了学习表示，以迁移到不同的数据变换。在MNIST(10-50%)和NORB(10-35%)上，通过引入MetaVAE验证了该方法的有效性，并表明该方法在下游图像分类任务中显著优于基准结果。

1. Exact and Approximate Inference (精确与近似推理)

➢ Exact and Approximate Inference (精确与近似推理)



- 精确推理：An *inference query* (即$p(z|x)$) involves computing posterior beliefs after incorporating *evidence* (即$p(x)$) into the prior:

$$p(z\mid x)=\frac{p(x,z)}{p(x)}=\frac{p(x\mid z)p(z)}{\boxed{\int_z p(x,z)dz}}$$

难求解

- 近似推理技术，如马可夫链蒙特卡罗(MCMC)抽样和变分推理(VI)被广泛用于近似求解后验概率$p(z|x)$。
- 变分推理中，找一族易求解的分布Q，其中$q_\psi(z)\in Q$ ，用KL散度来度量该分布$q(z)$与真实后验概率$p(z|x)$之间的近似程度：

$$q_{\psi^*}(z)=\arg\min_{q_\psi\in Q} D_{KL}(q_\psi(z)\,\|\,p(z\mid x)) \tag{1}$$

- 可以看到求解上式依赖于观测变量x，因此将参数ψ改写为ψ_x。
- 设$p_D(x)$为观测变量$x\in X$上的经验分布，变分近似的average quality可以被量化为：

$$\mathbb{E}_{p_D(x)}\left[\max_{\psi_x} ELBO\right],其中ELBO=\mathbb{E}_{q_{\psi_x}(z)}\ln\frac{p(x,z)}{q_{\psi_x}(z)} \tag{2}$$

- 实际中，观测数据的分布$p_D(x)$是未知的，但是我们假设可以访问从$p_D(x)$中独立同分布采样得到的样本，这些样本构成训练数据集D。
- $\ln p(x)=ELBO+D_{KL}(q\|p)$， $p(x)$固定不变，最小化KL散度相当于最大化ELBO。

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

2. Amortized Variational Inference (摊销变分推理)

➢ Amortized Variational Inference (摊销变分推理)

- 另一种替代方法是利用摊销技术，通过将样本优化过程视为监督回归任务，由原先的$q(z)$改为$q(z|x)$。该方法降低了上式(2)的计算成本(加速训练)。KL散度变为：$\min_{\phi} D_{KL}(q_{\phi}(z\,|\,x)\,\|\,p(z\,|\,x))$

  代替直接为每个x都求解一个最优化的$q_{\psi_x^*}(z)$，摊销变分推理通过学习单个确定性映射

  $f_{\phi}: X \rightarrow Q$来预测$\psi_x^*$。新定义一种表示形式$q_{\phi}(z\,|\,x) = f_{\phi}(x)(z)$ 别急，后续会用

- 该方法引入了一个摊销差距(gap)，其中推理模型的不太灵活的参数化将等式(2)中的目标替换为如下下界

$$\mathbb{E}_{p_D(x)}\left[\max_{\psi_x} \mathbb{E}_{q_{\psi_x}(z)} \ln \frac{p(x,z)}{q_{\psi_x}(z)}\right] \Rightarrow \max_{\phi} \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_{\phi}(z|x)} \ln \frac{p(x,z)}{q_{\phi}(z\,|\,x)}\right] \tag{3}$$

- 这一差距是指在整个训练集中对变分参数进行摊销而导致的次优性，而不是针对每个样本单独进行优化(从等式(2)中求出最大化期望值的所在参数)。然而，在表现力上的这种折衷可以实现显著的加速。
- 到目前为止，我们假设真实的生成模型$p(x, z)$是已知的。然而，我们通常只拥有一组可能的模型，由θ参数化的$p_{\theta}(x, z)$和观测数据集D。因此，面临的挑战是选择最优的θ，使其模型能更好地解释*evidence* (即$p(x)$)。为此，我们最大化数据的对数边际似然：

$$\max_{\theta} \mathbb{E}_{p_D(x)}\left[\ln p_{\theta}(x)\right] = \mathbb{E}_{p_D(x)}\left[\ln \int_z p_{\theta}(x,z)dz\right] \tag{4}$$

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

3. 近似变分推理 vs 摊销变分推理

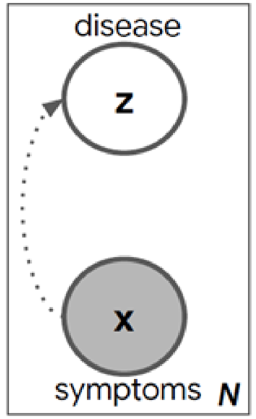
➢ 近似变分推理 vs 摊销变分推理

- Approximate Variational Inference (近似变分推理)

dependence on x: learn new q **per data point**

$$\mathbb{E}_{p_D(x)}\left[\max_{\psi_x} \mathbb{E}_{\boxed{q_{\psi_x}(z)}} \ln \frac{p(x,z)}{q_{\psi_x}(z)}\right]$$

-> turned an intractable inference problem into an optimization problem

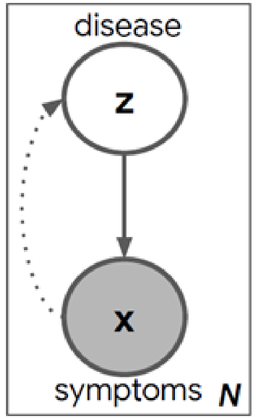


- Amortized Variational Inference (摊销变分推理)

deterministic mapping predicts z as a function of x

$$\max_{\phi} \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{\boxed{q_\phi(z|x)}} \ln \frac{p(x,z)}{q_\phi(z \mid x)}\right]$$

-> scalability: VAE formulation



4. Amortized Variational Autoencoders (摊销变分自编码器)

Meta-Amortized Variational Inference and Learning

➢ Amortized Variational Autoencoders (摊销变分自编码器)

- 如上所述，式(4)很难求解。相反，我们使用$q_\phi(z\mid x)$作为一个易于处理的摊销推理模型来推导式(4)的证据下界(L)

$$\mathbb{E}_{p_D(x)}\left[\ln p_\theta(x)\right] \ge \boxed{\mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln \frac{p_\theta(x,z)}{q_\phi(z\mid x)}\right] = L} \quad (5)$$

- L基础上减去一项常数项

$$\mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln \frac{p_\theta(x,z)}{q_\phi(z\mid x)}\right] - \boxed{\mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln p_D(x)\right]} = -\left\{\mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln \frac{q_\phi(z\mid x)}{p_\theta(x,z)}\right] + \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln p_D(x)\right]\right\}$$

$$= -\mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln \frac{q_\phi(x,z)}{p_\theta(x,z)}\right] = -\mathbb{E}_{q_\phi(x,z)} \ln \frac{q_\phi(x,z)}{p_\theta(x,z)} = -D_{KL}(q_\phi(x,z)\,\|\,p_\theta(x,z)) \quad (6)$$

- 这里令 $q_\phi(x,z) = q_\phi(z\mid x)p_D(x)$，继续推导：

$$D_{KL}(q_\phi(x,z)\,\|\,p_\theta(x,z)) = \int_x\int_z q_\phi(z\mid x)p_D(x)\ln\frac{q_\phi(z\mid x)p_D(x)}{p_\theta(z\mid x)p_\theta(x)}dzdx = \int_x p_D(x)\int_z q_\phi(z\mid x)\ln\frac{q_\phi(z\mid x)}{p_\theta(z\mid x)}dzdx + \int_x\int_z q_\phi(z\mid x)p_D(x)\ln\frac{p_D(x)}{p_\theta(x)}dzdx$$

$$= \mathbb{E}_{p_D(x)}\left[\int_z q_\phi(z\mid x)\ln\frac{q_\phi(z\mid x)}{p_\theta(z\mid x)}dz\right] + \int_x p_D(x)\ln\frac{p_D(x)}{p_\theta(x)}dx = \mathbb{E}_{p_D(x)}\left[D_{KL}(q_\phi(z\mid x)\,\|\,p_\theta(z\mid x))\right] + D_{KL}(p_D(x)\,\|\,p_\theta(x))$$

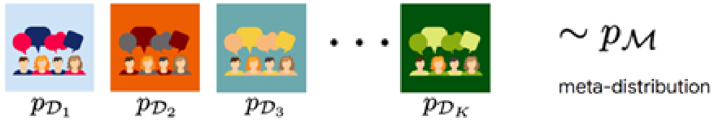
- 因此， $\max L$转化为$\boxed{\min_{\phi,\theta} \mathbb{E}_{p_D(x)}\left[D_{KL}(q_\phi(z\mid x)\,\|\,p_\theta(z\mid x))\right] + D_{KL}(p_D(x)\,\|\,p_\theta(x))}$ (7)

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020. 4

这部分内容推导与直面联合分布很类似，可以参看：https://www.cnblogs.com/kailugaji/p/12463966.html#_lab2_0_2

5. Meta-Amortized Variational Inference (元摊销变分推理)

➢ Meta-Amortized Variational Inference (元摊销变分推理)

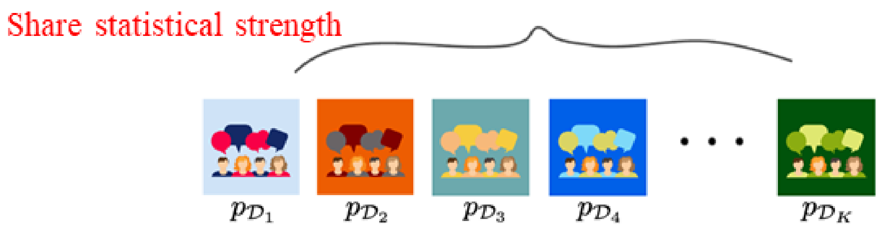


- 假设$p(x,z)$已给定，$q_\phi(z\mid x)=f_\phi(x)(z)$，$L$可进一步写为：

$$\max_\phi \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)} \ln \frac{p_\theta(x,z)}{q_\phi(z\mid x)}\right] \Rightarrow \max_\phi \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{f_\phi(x)} \ln \frac{p_\theta(x,z)}{f_\phi(x)(z)}\right] \tag{8}$$

- 其中观测样本$x\sim p_D(x)$。
- 元学习研究对象不再是单个模型，而是一组模型，$\mathcal{J}_\mathcal{I}=\{p_{\theta_i}(x,z), i\in\mathcal{I}\}$，其中$\mathcal{I}$为有限指标集(用于标注来自哪一个模型)
- 为简化模型，给定如下假设：
  1. 每个模型中的随机变量具有相同的域(如$\mathcal{X}$, $\mathcal{Z}$)，但随机变量之间的关系可能是不同的。
  2. 对于每个模型，我们关注相同的*inference query* (即$p_{\theta_i}(z\mid x)$) 。
  3. 在$\mathcal{J}_\mathcal{I}$中每个模型观测变量的代表值都存在某些知识。
- 给定一组观测变量上的边际分布 $\mathcal{M}_\mathcal{I}=\{p_{D_i}(x), i\in\mathcal{I}\}\subseteq\mathcal{M}$，其中$\mathcal{M}$表示$\mathcal{X}$上的所有可能的边际分布的集合.
- $p_\mathcal{M}:\mathcal{M}_\mathcal{I}\to[0,1]$表示$\mathcal{M}_\mathcal{I}$上的一个分布. 由于$p_\mathcal{M}$是所有分布之上的一个分布，因此视为**元分布(meta-distribution)**。
- 在一组模型上摊销的一种简单方法是

$$\mathbb{E}_{p_{D_i}\sim p_\mathcal{M}}\left[\max_\phi \mathbb{E}_{p_{D_i}(x)}\left[\mathbb{E}_{f_\phi(x)} \ln \frac{p_{\theta_i}(x,z)}{f_\phi(x)(z)}\right]\right] \tag{9}$$



- 然而，随着$\mathcal{M}_\mathcal{I}$规模的增加，这种方法的成本太高，而且跨模型的训练是解耦的(分开的)。

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

➢ Meta-Amortized Variational Inference (元摊销变分推理)

shared meta-inference network

- 因此，我们按如下方式进行双重摊销推断过程(再次将最大化max移到外边)

$$\boxed{\max_{\phi}} \mathbb{E}_{p_{D_i} \sim p_{\mathcal{M}}} \left[\mathbb{E}_{p_{D_i}(x)} \left[\mathbb{E}_{\boxed{g_\phi(p_{D_i};x)}} \ln \frac{p_{\theta_i}(x,z)}{g_\phi(p_{D_i},x)(z)} \right] \right] \quad (10)$$

用双重摊销回归器 $g_\phi(p_{D_i},x)$来替代原始的 $f_\phi(x)$，需要边际分布 $p_{D_i}(x)$与观测变量 x来返回后验分布

- 我们称这样的映射 $g_\phi: \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{Q}$ 为元推理模型(meta-inference model)。这种双重摊销推理模型必须在不同的边际 (marginals)和证据(evidence)上具有鲁棒性，并可以在$\mathcal{M}$上泛化到一组足够相似的、以前从未见过的模型上。
- Meta-Amortized Variational Bayes and Learning (元摊销变分贝叶斯与学习)
- 在某些情况下，我们会得到一组生成模型$\{p_{\theta_i^*}(x,z), i \in \mathcal{I}\}, p_i(x) \in \mathcal{M}_{\mathcal{I}}$，然后我们可以立即对式(10)进行优化，得到最优的元推理模型。但在许多情况下，生成模型并非提前知道，因此我们必须联合学习$\{\theta_i, i \in \mathcal{I}\}$与元推理模型参数 ϕ。因此，考虑如下目标：

$$\max_{\phi} \mathbb{E}_{p_{D_i} \sim p_{\mathcal{M}}} \left[\max_{\theta_i} \mathcal{L}_{\phi,\theta_i}(p_{D_i}) \right] \quad (11)$$

- 其中内层损失函数为 $\mathcal{L}_{\phi,\theta_i}(p_{D_i}) = -D_{KL}(p_{D_i}(x) g_\phi(p_{D_i},x) \,\|\, p(z) p_{\theta_i}(x|z))$

$p_{D_i}(x) g_\phi(p_{D_i},x)$表示通过先采样$x \sim p_i(x)$后采样$z \sim g_\phi(p_{D_i},x)$隐式定义的分布。

- 我们将这个下界称为MetaELBO，将此目标训练的VAE称为**MetaVAE**。

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

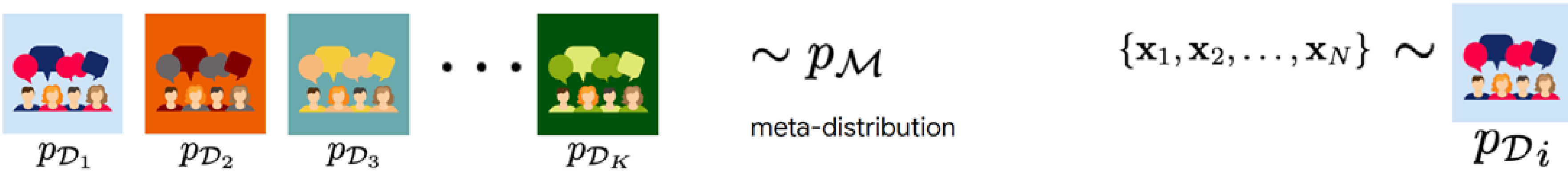
➢ Meta-Amortized Variational Inference (元摊销变分推理)

- 最后，正如我们在式(7)中所做的，我们可以将MetaELBO重写为更易于解释的形式。与$f_\phi(x)$类似，我们的回归器$g_\phi(p_{D_i},x)$可以表示为条件分布，表示为$q_\phi(z \mid p_{D_i},x)=g_\phi(p_{D_i},x)(z)$，然后

$$\mathcal{L}_{\phi,\theta_i}(p_{D_i}) = -D_{KL}(p_{D_i}(x)q_\phi(z \mid p_{D_i},x) \,\|\, p(z)p_{\theta_i}(x \mid z))$$

$$= -D_{KL}(p_{D_i}(x) \,\|\, p_{\theta_i}(x)) - \mathbb{E}_{x\sim p_{D_i}(x)}\left[D_{KL}(q_\phi(z \mid p_{D_i},x) \,\|\, p_{\theta_i}(z \mid x))\right]$$

- 这种形式对每个分布$p_{D_i}(x)$都有一个惩罚项，鼓励元摊销推理模型在从元分布$p_{\mathcal{M}}$中抽样的$p_{D_i}(x)$中表现更好。若$\mathcal{M}=\{p_D\}, g_\phi(p_{D_i},x)=f_\phi(x)$，则MetaELBO等价于ELBO.
- 有趣的是，本文发现MetaVAE的学习表示在测试时能够很好地迁移到从未见过的下游任务上。来自相应边际分布p_{D_i}的样本有助于减少元推理网络对每个查询点x的推断z的方差，使模型的行为正则化，从而产生更鲁棒的表示。
- Representing the Meta-Distribution (表示元分布)
- 将边际分布表示为一个有限的样本集$D_i=\{x_j \sim p_{D_i}(x) \mid j=1,\ldots,N\}$ (12)
  然后用D_i定义$g_\phi(p_{D_i},x)$，$\hat{g}_\phi:\mathcal{X}^N\times\mathcal{X}\to Q$，它将一个有$N$个样本和一个观察的数据集映射到一个后验。
- 然后，在式(11)中，用数据集D_i替换边际$p_{D_i}(x)$。



Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

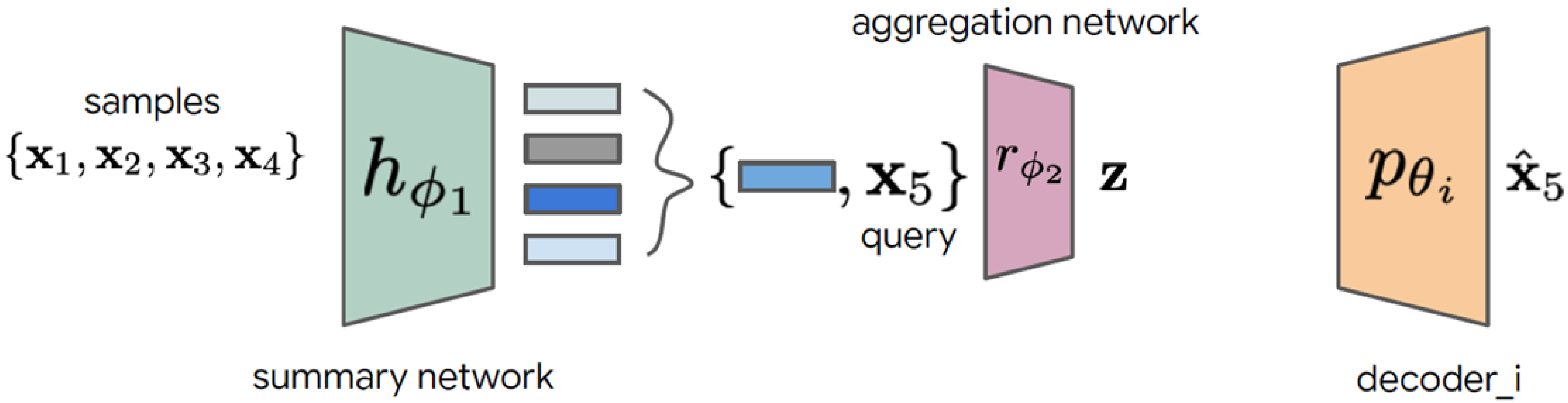
➢ Meta-Amortized Variational Inference (元摊销变分推理)

在实践中，对于某一数据集 D_i 与输入 x，我们实现元推理模型 $g_\phi(D_i, x) = r_{\phi_2}(\mathrm{CONCAT}(x, h_{\phi_1}(D_i)))$

其中 $\phi = \{\phi_1, \phi_2\}$

$h(\cdot)$ is a *summary* neural network that ingests the elements in D,

$r(\cdot)$ is an *aggregation* neural network that ingests the input and the summary.



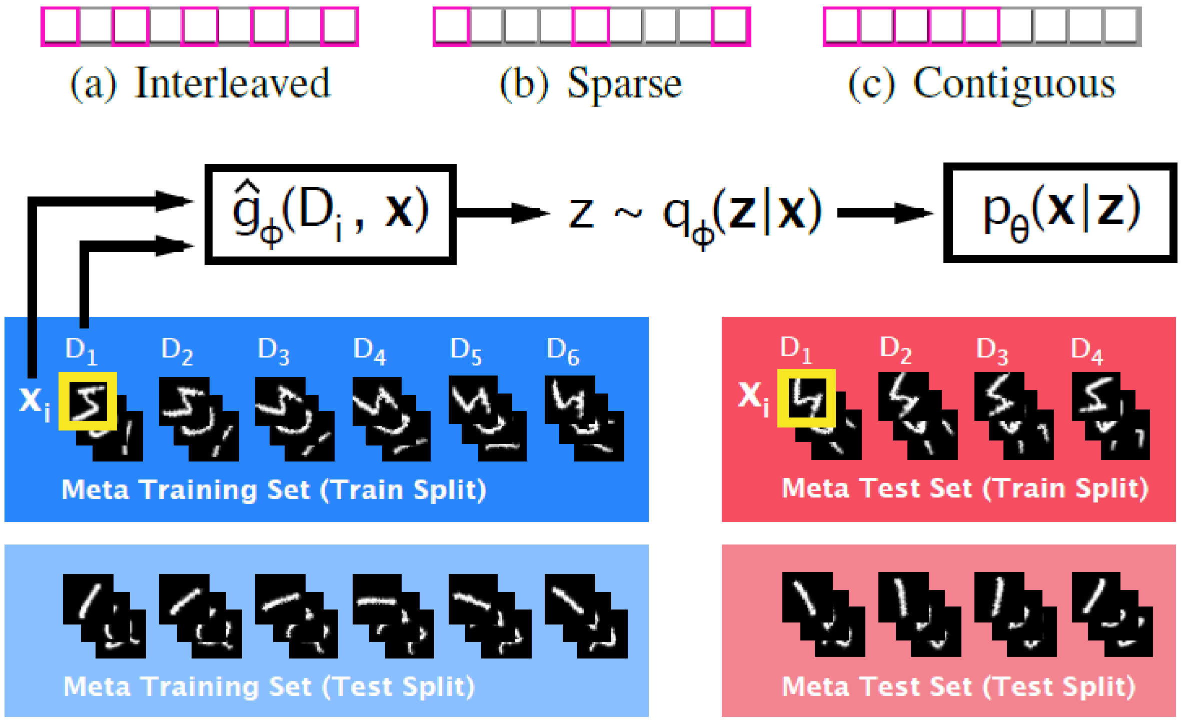
MetaVAE

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

➢ Meta-Amortized Variational Inference (元摊销变分推理)

Learning translation-invariant representations:

- (a-c)定义元训练和元测试的三种方法; (b,c)提出了更困难的泛化挑战。
- (d)为双重摊销推理程序概述。
- **元训练集**用于训练MetaVAE（其中验证集用于选择最佳参数)。
- **元测试集**用于评估学习的特征(其中训练部分用于拟合线性分类器，测试部分用于计算准确性)。



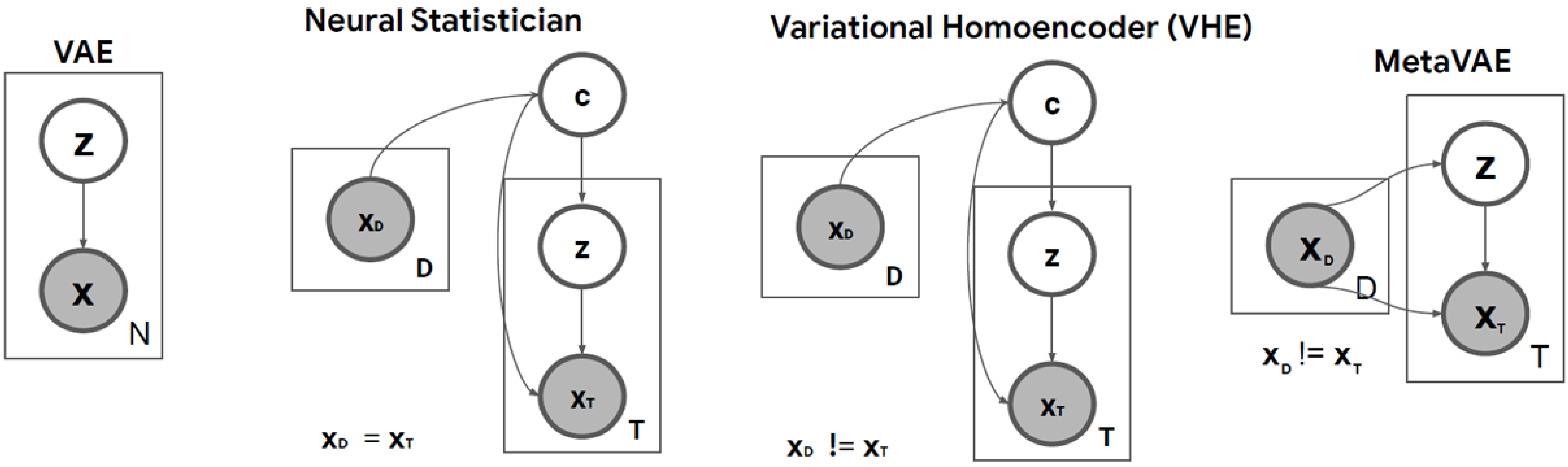
(d) Meta-Inference Pipeline

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

6. Related Works (VAE, Neural Statistician, VHE, and MetaVAE)

Meta-Amortized Variational Inference and Learning

➢ Related Works (VAE, Neural Statistician, VHE, and MetaVAE)



Avoid restrictive assumption on global prior over datasets $p(c)$

Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020. 10

7. 参考文献

[1] Mike Wu, Kristy Choi, Noah Goodman, and Stefano Ermon. Meta-Amortized Variational Inference and Learning. AAAI, 2020.

Paper: https://ojs.aaai.org//index.php/AAAI/article/view/6111

Code: https://github.com/mhw32/meta-inference-public

[2] CS236, Meta-Amortized Variational Inference and Learning. https://deepgenerativemodels.github.io/assets/slides/meta_amortized.pdf

[3] Variational Inference: Foundations and Modern Methods, P100 Amortizing Inference, 2016, https://media.nips.cc/Conferences/2016/Slides/6199-Slides.pdf

[4] 2021 Pyro Fundamentals, Amortized Inference, and Variational Autoencoders, https://robsalomone.com/wp-content/uploads/2021/07/L4_VAE.pdf

[5] Rui Shu, Hung H. Bui, Shengjia Zhao, Mykel J. Kochenderfer, Stefano Ermon. Amortized Inference Regularization. NeurIPS 2018, https://papers.nips.cc/paper/2018/hash/1819932ff5cf474f4f19e7c7024640c2-Abstract.html

[6] Amortized Optimization - Rui Shu http://ruishu.io/2017/11/07/amortized-optimization/