1、Learning Causal Semantic Representation for OoD prediction

看这篇文章的时候觉得比较吃力，因为对causality、inference这方面的文章了解很少，所以没看完，而且其中还有很多不明白的地方。下周继续看。

1.1 CSG(Causal semantic generative)模型提出

$p :=< p(s,v), p(x|s,v), p(y|s) > s是semantic隐变量，v是variation隐变量.$

1.2 关键假设

The Causal Invariance Principle：在CSG模型中，$p(x|s,v)$和$p(y|s)$在domain间保持不变，$p(s,v)$是domain change的唯一来源。

2、Methods

在这一部分，需要预先理解以下事实：
$p :=< p(s,v), p(x|s,v), p(y|s) >$是我们希望训练的模型，可以把里面这三个东西理解成要优化的参数。q也是我们要学习（优化）的参数。p^*是我们拥有的数据（在ood setting下是有标注的$p^*(x,y)$，DA setting下是无标注的$p^*(x)$）所服从的潜在的分布（我们不知道）。
2.1 Methods for OoD generalization
直接通过maximize $E_{p^*(x,y)}[log\, p(x,y)]$来拟合CSG模型$p :=< p(s,v), p(x|s,v), p(y|s) >$比较棘手，因为$p(s,v,x,y) := p(s,v)p(x|s,v)p(y|s)$是很难估计的。于是引入容易采样的inference model$q(s,v|x,y)$，通过maximize上述联合分布的ELBO来实现对p的估计。（**见附录F.1.1 一旦当q(s,v|x,y)达到了p(s,v|x,y),那么ELBO就变成了$logp(x,y)$的在一个固定的模型p的紧下界。因此再通过优化p来maximize ELBO就是maximize $logp(x,y)$，即进行极大似然估计**）问题转化为：
$\max\limits_{q_{s,v|x,y}} \mathcal{L}_{p,q_{s,v|x,y}}(x,y) = E_{q(s,v|x,y)}[log\frac{p(s,v,x,y)}{q(s,v|x,y)}]$
进一步，由于q(s,v|x,y)不能帮助估计$p(y|x)$，上述方法缺乏预测能力。为了解决这个问题，引入$q(s,v,y|x)$来估计$p(s,v,y|x)$。它既能表示出$q(s,v|x,y)$以获得前面提出的inference model：
$q(s,v|x,y) = q(s,v,y|x)/q(y|x)其中的q(y|x) = \int q(s,v,y|x)dsdv$
又能得到$q(y|x)$进而估计$p(y|x)$，解决预测问题。

将(2)代入(1)，得到$E_{p^*(x,y)}[\mathcal{L}_{p,q_{s,v|x,y}}(x,y)]$的新形式：(**注意这里套上了$E_{p^*(x,y)}$，是因为要对$p(x,y)$做极大似然估计，为什么写成这种形式可以见下面的知识补充。**)
$E_{p^*(x)}E_{p^*(y|x)}[log\, q(y|x)] + E_{p^*(x)}E_{q(s,v,y|x)}[\frac{p^*(y|x)}{p(y|x)}log\, \frac{p(s,v,x,y)}{p(s,v,y|x)}]$
分析一下(4)：第一项是负CE，迫使$q(y|x) \to p^*(y|x)$，在第一项实现的情况下，第二项退化为期望ELBO $E_{p^*(x)}[E_{q(s,v,y|x)}(x)]$，迫使$q(s,v,x|y) \to p(s,v,x|y)$以及$p(x) \to p*(x)$**(因为是对$p(x)$做了MLE)**.进一步，由于$p(s,v,y|x) = p(s,v|x)p(y|s)$，其中$p(y|s)$已知，考虑用q(s,v|x)估计难处理的p(s,v|x)，于是$q(s,v,y|x)$可以表示为$q(s,v|x)p(y|s)$。代入(3)，问题变为：
$\max\limits_{p,q_{s,v|x}} E_{p^*(x,y)}[log\, q(y|x) + \frac{1}{q(y|x)}E_{p(s,v,y|x)}[p(y|s)log\, \frac{p(s,v)p(x|s,v)}{q(s,v|x)}]]$
其中：$q(y|x) = E_{q(s,v|x)}[p(y|s)]$(这样(4)中的变量就全部用要优化的变量$p$以及$q_{s,v|x}$表示了)

CSG-ind
使用$p^\perp(s,v) = p(s)p(v)$作为prior。直觉上，它忽略了s和v之间在training domain上的虚假关联。在ood泛化中，test domain的inference model $q^\perp(s,v|x)$和training domain的inference model $q(s,v|x)$都需要。前者用于prediction：$p^\perp(y|x) = E_{q^\perp(s,v|x)}[p^\perp(y|s)] = E_{q^\perp(s,v|x)}[p(y|s)]$（由1.2知：$p(y|s)$ is invariant across domains, 这里加或不加⊥就代表test domain和training domain）
后者用于在training domain上的学习。

为了将学习上述两个模型简化为一个只学习模型，考虑用其中一个inference model表示另一个：由1.2以及CSG的图结构：$p(s,v|x)=\frac{p(s,v)}{p^{\perp}(s,v)}\frac{p^{\perp}(x)}{p(x)}p^{\perp}(s,v|x)$

进而：

$q(s,v|x)=\frac{p(s,v)}{p^{\perp}(s,v)}\frac{p^{\perp}(x)}{p(x)}q^{\perp}(s,v|x)$

把(5)代入(4)，就得到了：

$\underset{p,q^{\perp}_{s,v|x}}{max}E_{p^*(x,y)}[log\,\pi(y|x)+\frac{1}{\pi(y|x)}E_{p^{\perp}(s,v,y|x)}[\frac{p(s,v)}{p^{\perp}(s,v)}p(y|s)log\,\frac{p(s,v)p(x|s,v)}{q(s,v|x)}]]$

其中，$\pi(y|x)=E_{q^{\perp}(s,v|x)}[\frac{p(s,v)}{p^{\perp}(s,v)}p(y|s)]$

注意：(4)到(6)的推导中，$\frac{p(x)}{p^{\perp}(x)}$由于和优化的变量$p,q^{\perp}_{s,v|x}$无关，因此作为常数项被略去了。

2.2 Method for Domain Adaptation

CSG-DA

在DA的任务中，可以得到test domain的数据分布$\tilde{p}^*(x)$。与CSD-ind推导类似，把$p^{\perp}(s,v)$和$p^{\perp}(s,v|x)$换成$\tilde{p}(s,v)$和$\tilde{p}(s,v|x)$即可，后两者为test domain的利用其数据分布$\tilde{p}^*(x)$得到的新的prior。（和$\tilde{p}^*(x)$的具体关系？）

3、Theory

3.1 Assumption

Additive noise assumption

存在三阶导数有界的非线性函数f、g和独立随机变量μ、ν，使得：$p(x|s,v)=p_{\mu}(x-f(s,v))$，$p(y|s)=p_{\nu}(y-g(s))$对连续的$y$成立，或$p(y|s)=Cat(y|g(s))$对categorical variable y成立。（这个意思大概是说：$p(x|s,v)$这个生成机制可以用$x=f(s,v)+\mu$来表示，μ是噪声。y的生成过程同理）

Bijectivity

假设f是双射的，g是单射的。

Definition 3.2 semantic-identification

一个CSP p是semantic-identification的，如果存在一个$\mathcal{S}\times\mathcal{V}$上的同胚映射$\Phi$，使得：①变换后的output在$s$的维度$\Phi^{\mathcal{S}}(s,v)=\Phi^{\mathcal{S}}(s,v')$ ②它是ground-truth CSG p^*的reparameterization p^*：$\Phi_{\#}[p^*_{s,v}]=p_{s,v},p^*(x|s,v)=p(x|\Phi(s,v))且p^*(y|s)=p(y|\Phi^{\mathcal{S}}(s))$

remark:

3.2.1 ①表示：reparameterization后，output在$\mathcal{S}$空间的维度与v无关。**注**：这里的reparameterization的含义为：把一个CSG p在数据分布$p(x,y)$不变的前提下变换为另一个CSG p'。（见Lemma 9.）

3.2.2 定义3.2实际上说的是p与ground-truth$p*$是**semantic-equivalence**的（见def 10.）

Theorem 4 semantic-identifiability

在假设3.1下，称一个CSG p是semantic-identified，如果它在如下条件下是well-learned，即$p(x,y)=p^*(x,y)$：

①$log\,p(s,v)$和$log\,p^*(s,v)$及其二阶导数有界

②满足$\frac{1}{\sigma_{\mu}^2}\to 0(\sigma_{\mu}^2=E[\mu^{\top}\mu])$或$p_{\mu}$(e.g. 一个高斯变量)有非零的特征函数

remark:

有界意味着s、v之间的correlation是随机的，也即不存在确定性的关联，否则$p^*(s,v)$的概率密度会集中到$\mathcal{S}\times\mathcal{V}$的低维子空间，进而导致其unbounded（？）这里我的理解是：如果s和v之间存在correlation，比如：床总在卧室、桌子总在办公室，那么$p(s,v)$在空间$\mathcal{S}\times\mathcal{V}$中会集中分布在一个点(床，卧室)（或(桌子，办公室)）附近，也就是集中在一个低维子空间上，这样的话对概率密度积分之后$p(s,v)$就会趋向于0，从而使$logp(s,v)$趋向负无穷。

$1/\sigma_{\mu}^2$反映了生成机制$p(x|s,v)$的强度（？）我的理解是$1/\sigma_{\mu}^2$越大，σ_{μ}^2越小，从而p_{μ}的波动程度越小（因为$\sigma_{\mu}^2=E(\mu^{\top}\mu)=D(\mu)+E^2(\mu)=D(\mu)$，所以$\sigma_{\mu}^2$实际上就是噪声$\mu$的方差），从而$\mu=x-f(s,v)$越稳定，也就是生成机制越强。

上面两处（？）的理解受限于个人知识和理解能力，希望有大佬指正。

semantic-identifiability的提出为后续ood generalization的性能提供了保证。

5、 OOD generalization theory
Theorem 5.1（OOD generalization error）
在假设3.1下，对于一个在training domain上semantic-identified且具有semantic-preserving的reparameterization Φ的CSG p，有：
$E_{p^*(x)}||E[y|x]-\tilde{E}^*[y|x]||^2 \leq \sigma_\mu^2 B_{f^{-1}}'^4 B_g'^2 E\tilde{p}_{s,v}||\nabla log(\tilde{p}_{s,v}/p_{s,v})||^2$
remark:
①$\tilde{E}^*[y|x]$是最优分类器。σ_μ^2越小，$p(x|s,v)$越大，生成机制越强，泛化误差就越小

②体现了CSG-ind的优势。p8 remark部分（？）

关键证明：

Learning Causal Semantic Representation for OOD prediction*：Theorem 5'

①预备概念和定理：

def 8.
从p到p'的reparametrization：一个$\mathcal{S}\times\mathcal{V}$上的同胚映射(**连续双射即同胚映射**)，能够使得映射后的三个概率相等。
semantic-preserving：满足$\Phi^{\mathcal{S}}(s,v)$与v无关的reparametrization Φ

def 10. semantic equivalence
是两个CSG p与p'间的二元关系。要求存在一个同胚映射Φ，是semantic-preserving的，且使p与p'的三个概率相等。

prop 14.
semantic equivalence是一个等价关系。该等价类中的CSG，都能够导出相同的p(x,y)，而且s中含有相同的semantic information。

theo 5'.（theo 5.的详细版）
对于两个CSG，p和p'，如果p(x,y)=p'(x,y)且满足三个条件的其中之一，则它们是semantic-equivalence的。也就是说，存在一个semantic-preserving的同胚映射Φ，使得p与p'的三个概率相等。证明过程中提到了：$f'^{-1}(f())$就是一个这样的同胚映射。
这个定理保证了我们学到的模型p可以与ground-truth模型p*实现semantic-equivalence。

② theo 5'.的证明
需要证明：$\Phi():=f'^{-1}(f())$满足以下两条：

[1]是semantic-preserving的，即它与v无关
[2]p与p'的三个概率值相等：$p(x|s,v)=p'(x|\Phi(s,v))$且$p(y|s)=p'(y|\Phi^{\mathcal{S}}(s))$且$\Phi_{\#}[p_{s,v}]=p'_{s,v}$

由于g(s)是p(y|s)的一个单射的充分统计量，由lemma 11.立即得到[1]成立。

下面证明[2]的第一条：由于$\Phi():=f'^{-1}(f())$是同胚映射，立即得到$p(x|s,v)=p'(x|\Phi(s,v))$（equ（7））。

因此只要证明[2]的后面两条。

######（1）假设1下的证明:
比较平凡。

大概就是把p(x)和E(y|x)写成了卷积的形式，然后match p(x,y)=p'(x,y)得到p(x)=p'(x)和E(y|x)=E'(y|x)，之后**由特征函数为0这个假设**可以立即得到$f_\#[p_z]=f'_\#[p'_z]$和$f_\#[gp_z]=f'_\#[g'p'_z]$.后面的推导就比较简单了，**唯一需要注意的一处是：** $f_\#[p_z]=f'_\#[p'_z]$ indicates $\Phi_\#[p_z]=p'_z$的证明。由于f'是双射，因此将等号右边的$f'_\#$取逆，移到左边，即可得到要证明的目标。

######（2）假设2下的证明
(**注：下面过程中的泰勒展开式都是$f(x-u)$对μ在$\mu=0$时的展开，因此那些函数比如$\overline{p}_zV$，实际上是**$\overline{p}_zV(x)$)

首先利用（1）中的结果：$p(x)=E_\mu[(\overline{p}_zV)(x-\mu)]$和$E(y|x)=\frac{1}{p(x)}E_\mu[(\overline{gp}_zV)(x-\mu)]$，然后对$\mu$麦克劳林展开（因为**假设2：**$E(\mu^\top\mu)$**是无穷小**，所以在$\mu=0$展开），得到：
$p(x)=\overline{p}_zV+\frac{1}{2}E_\mu[\mu^\top\nabla\nabla^\top(\overline{p}_zV)\mu]+O(\sigma_\mu^3)$
之后对$\frac{1}{p(x)}$使用$\frac{1}{x+\epsilon}=\frac{1}{x}-\frac{\epsilon}{x^2}+O(\epsilon)$，对$E_\mu[\mu^\top\nabla\nabla^\top(\overline{p}_zV)\mu]$进行麦克劳林展开（**这一步的合理性？**），并把得到的结果与$E_\mu[(\overline{gp}_zV)(x-\mu)]$对$\mu$的麦克劳林展开相乘，从而得到$E(y|x)$的展开形式。（**这里相乘之后怎么得到的原文式（14）我没看明白**）

之后对$p(x)$和$E(y|x)$的展开形式进行一些比较初等的放缩，能够证明出：
$|p(x)-(\overline{p}_zV)(x)|=O(E[\mu^\top\mu])|E(y|x)-\overline{g}(x)|=O(E[\mu^\top\mu])$，**由假设2，**$E[\mu^\top\mu]\to 0$可以知道$p(x)$和$E(y|x)$分别收敛到$(\overline{p}_zV)(x)=f_\#[p_z](x)$和$\overline{g}(x)$。从而由
$p(x,y)=p'(x,y)\Rightarrow p(x)=p'(x)$且$E(y|x)=E'(y|x)$，我们能得到$f_\#[p_z]=f'_\#[p'_z]$和
$\overline{g}f_\#[p_z]=\overline{g}'f'_\#[p'_z]$，之后证明步骤就和假设1下一样了，得到$p(y|s)=p'(y|\Phi^{\mathcal{S}}(s))$且
$\Phi_\#[p_{s,v}]=p'_{s,v}$

######（3）假设3下的证明:
这一部分大量的矩阵运算，太复杂了，一开始尝试每个式子都推一遍，后来发现太费时间了，于是这里这里只写出一个证明的框架。

这一部分的证明思路还是去bound原文(15)(16)的$|p(x)-(\overline{p}_zV)(x)|$和$|E(y|x)-\overline{g}(x)|$。由于现在没有$E[\mu^\top\mu]\to 0$这个条件了，因此需要用$f/g/p_z$和它们的导数去bound。

①首先bound$|p(x)-(\overline{p}_zV)(x)|$:

从原文(16)可知，我们需要分别bound以下几项：$\nabla log\overline{p}_z$、$\nabla\overline{g}$、$\nabla logV(x)$和$||\nabla^\top\nabla\overline{g}||_2$.

前两项：$\nabla log\overline{p}_z=J_{f^{-1}}\nabla logp_z$和$\nabla\overline{g}=J_{f^{-1}}\nabla_z g$.第三项经过一顿非常复杂的矩阵运算后，可以证出$||\nabla logV(x)||_2\le dB'^2_{f^{-1}}B''_f$.与$\nabla log\overline{p}_z=J_{f^{-1}}\nabla logp_z$和$\nabla\overline{g}=J_{f^{-1}}\nabla_z g$组合之后得到原文(23)的bound.

$||\nabla^\top\nabla\overline{g}||_2$可以由原文(24)bound住。

组合(23)(24)，得到(25)，$|p(x)-(\overline{p}_zV)(x)|$的upper bound全部由假设3中的常数项表示了。证毕。

②然后bound$|E(y|x)-\overline{g}(x)|$:
分别bound原文(15)的几个展开项，再组合起来。太复杂，就不写了。