首先总结一下这篇文章最本质的observation：

基于H-divergence的表示不变性+源域上的低empirical risk并不足以保证目标域上的low risk。我自己的理解，由**于基于H-divergence的表示不变性并不是在given y条件下的表示不变性，而是整个input space的表示的不变性**（直观理解为通过不变表示g映射之后，整个源域和整个目标域的分布相同），所以在每个class内部，不同domain的表示并不一致，这就导致不能泛化。

PS：相比之下，Ye et al . 2021的工作中将不变性刻画为$d(p(\phi|y,e),p(\phi|y,e')$，就更靠谱。

PS：之前工作为什么提出基于H-divergence的distribution match，是因为之前的DA研究中，假设的是P(X)不同（covariate shift），甚至源域和目标域的X的support都不同，所以得先通过一个表示把他们的support映到同一个集合上，也就是H-divergence的衡量标准。这篇文章指出的就是这种衡量标准的问题。

1. 对H-divergence的理解：

   定义：

   Definition 2.1 (H-divergence). Let $\mathcal{H}$ be a hypothesis class on input space $\mathcal{X}$, and $\mathcal{A}_{\mathcal{H}}$ be the collection of subsets of $\mathcal{X}$ that are the support of some hypothesis in $\mathcal{H}$, i.e., $\mathcal{A}_{\mathcal{H}} := \left\{h^{-1}(1) \mid h \in \mathcal{H}\right\}$. The distance between two distributions $\mathcal{D}$ and $\mathcal{D}'$ based on $\mathcal{H}$ is: $d_{\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right) := \sup_{A \in \mathcal{A}_{\mathcal{H}}} \left|\Pr_{\mathcal{D}}(A) - \Pr_{\mathcal{D}'}(A)\right|$.[1]

   理解：选取一个predictor h，它在source和target domain上将某一部分样本预测为1（这个定义针对二分类问题，定义为预测为1或0没区别），被预测为1的这部分样本就是A（A表示在source和target上的各一部分样本）。H-divergence就是**针对所有可能的预测器集合$\mathcal{H}$，$\mathcal{D}$和$\mathcal{D}'$中被$\mathcal{H}$中的预测器判别为1的样本占比（概率）最大差多少。**

   remark：这个对距离分布的度量，受假设空间$\mathcal{H}$影响。

2. symmetic H-divergence:

   先定义$\mathcal{H}\Delta\mathcal{H}$：$\mathcal{H}\Delta\mathcal{H} = \{h(\mathbf{x}) \oplus h'(\mathbf{x}) \mid h, h' \in \mathcal{H}\}$

   $d_{\mathcal{H}\Delta\mathcal{H}}$就是**针对所有可能的预测器集合$\mathcal{H}$，$\mathcal{D}$和$\mathcal{D}'$中被$\mathcal{H}$中的预测器判别为不同label的样本占比（概率）最大差多少**

3. 本文中对于不变表示g的定义（关于某一假设空间$\mathcal{H}$）：$d_{\mathcal{H}}\left(\mathcal{D}_S^g, \mathcal{D}_T^g\right) = 0$，括号里的两个分布分别表示被$g$提取出来的特征的分布

4. Ben-David 2007的结果中$\lambda^* = \min_{h \in \mathcal{H}} \varepsilon_S(h) + \varepsilon_T(h)$项由于在实际中假设空间$\mathcal{H}$可能不够好（$\mathcal{H}$不足以包含最优的$h$），因此这一项在实际中的值可能较大。

   但本文把Ben的这一项换成了$\min\left\{\mathbb{E}_{\mathcal{D}_S}\left[|f_S - f_T|\right], \mathbb{E}_{\mathcal{D}_T}\left[|f_S - f_T|\right]\right\}$。

   意义：

   1. 消除了对$\mathcal{H}$的依赖，泛化误差取决于表示不变性（基于H-divergence）、source上的低error，和**源域和目标域的labeling function**，也即$P(Y|X)$的差距。

   remark：

   1. labelling function的差距是数据集本身的性质，我们无能为力。但是前面那个分布的差距那一项，可以通过学习一个不变特征提取器g来使得特征的分布很近，是我们可以控制的。
   2. covariate shift是$p(Y|X)$在不同domain不变，但不同domain $P(X)$不同。（现在我感觉这个假设很扯，不同domain由于spurious feature不同，$P(Y|X)$应当不同，$P(Y|\phi_i(X))$才应该不同，ϕ_i是不变特征提取器）

5. **Theorem 4.3，对学习不变表示的批判：**

如果label的marginal distribution在源域和目标域变化很大，那么强行学不变表示，会导致源域和目标域的误差之和较大。

Theorem 4.3. Suppose the condition in Lemma 4.8 holds and $d_{\mathrm{JS}}\left(\mathcal{D}_S^Y, \mathcal{D}_T^Y\right) \geq d_{\mathrm{JS}}\left(\mathcal{D}_S^Z, \mathcal{D}_T^Z\right)$, then:

$$\varepsilon_S(h \circ g)+\varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}\left(\mathcal{D}_S^Y, \mathcal{D}_T^Y\right)-d_{\mathrm{JS}}\left(\mathcal{D}_S^Z, \mathcal{D}_T^Z\right)\right)^2$$