文章主要在做的事情：在样本的混合分布信息未知，仅仅是假设是混合的前提下，将loss function对于X的条件期望，根据minorty subgroup样本再做一次期望。

Proof Sketch of Duality (Lemma 1):

1. 我们需要把$\mathbb{E}_{Q_0}[W]$ 转化成$\mathbb{E}_{P_X}[W]$的形式，这一过程其实并不复杂。观察到$P_X = \alpha_0 Q_0 + (1 - \alpha_0 Q_1)$，我们会发现$Q_0$相对于$P_X$是绝对连续的（概念参见概率论）。写出其Radon-Nikodym导数$L = \frac{dQ_0}{dP_X}$，随后重写原函数：

$$\sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathbb{E}_{X \sim Q_0}[W] = \sup_L \left\{ \mathbb{E}_P[LW] \middle| L : \Omega \to [0, 1/\alpha_0],\ \text{measurable},\ \mathbb{E}_p[L] = 1 \right\}. \tag{34}$$

2. 引用其他文章的结论：

Lemma D.1 ([70, Example 6.19]). *For any random variable $W : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}|W| < \infty$,*

$$\begin{aligned} &\sup_L \left\{ \mathbb{E}_P[LW] \middle| L : \Omega \to [0, \frac{1}{\alpha_0}],\ \textit{measurable},\ \mathbb{E}_p[L] = 1 \right\} \\ &= \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \mathbb{E}_{X \sim P_X} \left[(W - \eta)_+ \right] + \eta \right\}. \end{aligned}$$

Variational Approximation

对于条件期望的式子$\mathop{\mathbb{E}}_{X \sim P_X}[(\mathbb{E}[\ell(\theta; (X, Y)) | X] - \eta)_+]$，其内部的条件期望表达方式有很多（random forests, gradient boosted decision trees, kernel methods, nn）。但是我们其实需要一个explicit的表达，因此文章使用了variational的形式，

$$\begin{aligned} &\mathbb{E}[(\mathbb{E}[\ell(\theta; (X, Y)) | X] - \eta)_+] \\ &= \sup_{h: \mathcal{X} \to [0,1]} \mathbb{E}_P[h(X)(\ell(\theta; (X, Y)) - \eta)]. \end{aligned}$$

但是$h : \mathcal{X} \to [0, 1]$其实是一个巨大无比的函数族。通过引入RKHS，我们限制h的范围，得到一个lower bound

$$\left\{ \frac{1}{\alpha_0} \sup_{h \in \mathcal{H}} \mathbb{E}_P[h(X)(\ell(\theta; (X, Y)) - \eta)] + \eta \right\}.$$

KL-Divergence Upper Bound

文章引用Shapiro 2017, Duchi and Namkoong 2021，说明KL散度的对偶实际上就是一个generalized的CVaR对偶。通过适当地去选取Δ参数，KL散度引出的对偶形式为

$$R_p(\theta) = \inf_{\eta \geq 0} \left\{ \frac{1}{\alpha_0} \left(\mathbb{E}_{X \sim P_X}[(\mathbb{E}[\ell(\theta; (X, Y)) | X] - \eta)_+^p] \right)^{1/p} + \eta \right\}$$

和我们真实的performance对比：

$$\inf_{\eta}\left\{\frac{1}{\alpha_0}\mathop{\mathbb{E}}_{X\sim P_X}[(\mathbb{E}[\ell(\theta;(X,Y))|X]-\eta)_+]+\eta\right\}.$$

注：$p \in (1,2]$。你会发现上式是下式的UB（Holder ineq）

Empirical Implementation

我们上文已经给出了LB和UB，文章接下来要做的事是，对于Variational Approximation，给出一个经验式的可计算的表达式。

- 如果我们有replicate的数据，那么计算条件期望其实会比较简单。事实上在语言模型，用重复数据构造$P_{Y|X}$是一个比较常见的事。这种情形下得到一个结论：

Proposition 1. *Let Assumption 1 hold. There exists a universal constant C such that, for any fixed $\theta \in \Theta$ with probability at least $1-\delta$,*

$$\left|\mathcal{R}(\theta)-\inf_{\eta\in[0,M]}\left\{\frac{1}{\alpha_0 n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m}\ell(\theta;(X_i,Y_{i,j}))-\eta\right)_+ +\eta\right\}\right|$$

$$\leq C\frac{M}{\alpha_0}\sqrt{\frac{1+\log\frac{1}{\delta}}{\min\{m,n\}}}.$$

- 但是一些其他的数据中很可能X是各异的，意味着(X_i, Y_i)互不相同，这种情况会比较糟糕一些。
  - 如果我们直接用 $\mathcal{R}(\theta)=\inf_{\eta}\left\{\frac{1}{\alpha_0}\sup_{h:\mathcal{X}\to[0,1]}\mathbb{E}_P[h(X)(\ell(\theta;(X,Y))-\eta)]+\eta\right\}.$ ，那么其实做不了。主要是你怎么去选择一个subset $\mathcal{H}\subset\{h:\mathcal{X}\to[0,1]\}$。其实不管怎么选都是合理的，但是没有一个很统一的标准；
  - 文章的做法：用上面KL诱导的对偶$\mathcal{R}_p(\theta)$来替代。

$$\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))|X]-\eta)_+\right]$$
$$\leq\left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))|X]-\eta)_+^p\right]\right)^{1/p}$$
$$=\sup_{h\mathcal{X}\to\mathbb{R}_+}\left\{\mathbb{E}\left[h(X)(\ell(\theta;(X,Y))-\eta)\right]\middle|\mathbb{E}[h(X)^q]\leq q\right\}. \qquad (9)$$

精度

- 对于上界$\inf_\theta R_p(\theta)$，收敛速度为$O(n^{-\frac{p-1}{d+1}})$，这其实是一个conservative的速度，但是文章给了解释：
  1. 这个速度其实很难被improve(section 5)；
  2. 在实际计算中效果不错(section 6)。
- 对于下界而言，收敛速度为$O(n^{-\frac{1}{4}})$，理论性质不错。