Learning Representations that Support Robust Transfer of Predictors

思路是把模型在不同domain的泛化能力直接作为objective优化。提出了TRM(Transfer Risk Minimization)。所谓的Transfer Risk，也就是将一个环境的最优分类器在其他环境上的泛化能力进行平均得到的risk。可以用下面的loss来衡量：

w是分类器,Φ是特征表示，Ω是环境集合，P、Q是对应环境的empirical distribution。对于给定的Φ，TRM所提出的泛化能力衡量标准如下：
$R(\Phi;\Omega)=\sum_{Q\in\Omega}(\sup_{P\in Conv(\Omega/Q)} E_P[l(w(Q;\Phi)\circ\Phi)])$其中$w(Q;\Phi)=\arg\min_w E_Q[l(w\circ\Phi)]$是$\Phi$在Q上的最优分类器，$Conv(\Omega/Q)$是$\Omega$中除去Q的环境的凸包。

优化：

1、内层优化：需要对抗地找到让$w(Q;\Phi)$性能最差的环境P，要搜索凸包$Conv(\Omega/Q)=\{\sum_{P_i\in\Omega/Q}\alpha_i(Q)P_i|\alpha_i(Q)\geq 0,||\alpha(Q)||_1=1\}$中的元素P，于是需要对$\alpha$进行优化。做法是对$\alpha_i$进行Exponential Gradient Ascent。

2、外层优化：需要通过优化Φ来让$w(Q;\Phi)$在$P(Q)$上的loss最低。将$w(Q;\Phi)$在$P(Q)$上的loss $E_{P(Q)}[l(w(Q;\Phi)\circ\Phi)]$记为$L_P(Q)$。对$\Phi$的全梯度$\frac{dL_P(Q)}{d\Phi}$可以写成两项，$\frac{\partial L_P(Q)}{\partial\Phi}+(\frac{\partial L_P(Q)}{\partial w(Q;\Phi)})^T\frac{dw(Q;\Phi)}{d\Phi}$
第二项（implicit gradient）可以改写为$-\frac{\partial(sg((\frac{\partial L_P(Q)}{w(Q)})^T H_{w(Q)}^{-1})^T(\frac{\partial E_Q[l(w(Q))]}{\partial w(Q)}))}{\partial\Phi}$)其中$H_{W(Q)}=\frac{\partial^2 E_Q[l(w(Q))]}{\partial w(Q)^2}$，$sg$为stop_gradient操作，即让括号内的函数不对Φ求导。
证明：*注意到，$w(Q;\Phi)$是由方程$\frac{\partial E_Q[l(w(Q)\circ\Phi)]}{\partial w(Q)}=0$确定的隐函数。由隐函数存在定理，$\frac{dw(Q)}{d\Phi}=-\frac{\frac{\partial E_Q[l(w(Q)\circ\Phi)]}{\partial\Phi}}{\frac{\partial E_Q[l(w(Q)\circ\Phi)]}{\partial w(Q)}}=-H_{w(Q)}^{-1}\frac{\partial^2 E_Q[l(w(Q)\circ\Phi)]}{\partial w(Q)\partial\Phi}$代入implicit gradient，即可得证。*

将(2)代入(1)并对Φ积分，得到外层优化的损失函数:
$E_P[l(w(Q)\circ\Phi)]-(sg((\frac{\partial L_P(Q)}{w(Q)})^T H_{w(Q)}^{-1})^T(\frac{\partial E_Q[l(w(Q))]}{\partial w(Q)}))+C$

第一项保证环境Q上的最优分类器w(Q)在P上的性能；第二项是环境Q和w(Q)的最差性能环境P(Q)对w(Q)梯度的alignment。

第二项相比于Fish直接align不同domain的梯度：$\mathcal{L}_{idgm}=\mathcal{L}_{erm}(\mathcal{D}_{tr};\theta)-\gamma\frac{2}{S(S-1)}\sum_{i,j\in S}^{i\neq j}G_iG_j$
其中，$G_i=E_{\mathcal{D}_i}\frac{\partial l(x,y;\theta)}{\partial\theta}$，区别在于TRM加入了Hessian Inverse这一项，而且是对最优分类器的导数。

对Hessian Inverse的近似：对H^{-1}泰勒展开到前j项：$H_j^{-1}=\sum_{i=0}^{j}(I-H)^i$可以用递推式计算$H_j^{-1}$：$H_j^{-1}=I+(I-H)H_{j-1}^{-1}$递推式的矩阵乘法用参考Fast Exact Multiplication by the Hessian，可以在线性时间内完成。

算法：
将（3）的损失函数加上ERM的loss $E_Q[l(w_{all}\circ\Phi)]$用于更新总的分类器w_{all}，在给定$P(Q)$的条件下得到环境Q的loss：

$$R(\Phi,w_{all};Q)=E_Q[l(w_{all}\circ\Phi)]+E_P[l(w(Q)\circ\Phi)]-\lambda\frac{\partial(sg((\frac{\partial L_P(Q)}{w(Q)})^T H_{w(Q)}^{-1})^T(\frac{\partial E_Q[l(w(Q))]}{\partial w(Q)}))}{\partial\Phi}$$