Domain Generalization via Entropy Regularization -Supplementary Materials-

Shanshan Zhao
The University of Sydney
Australia
szha4333@uni.sydney.edu.au

Mingming Gong
University of Melbourne
Australia
mingming.gong@unimelb.edu.au

Tongliang Liu
The University of Sydney
Australia
tongliang.liu@sydney.edu.au

Huan Fu
Alibaba Group
China
fuhuan.fh@alibaba-inc.com

Dacheng Tao
The University of Sydney
Australia
dacheng.tao@sydney.edu.au

Here, we provide the proofs and the illustration of our framework supporting the contents in the submission.

S. 1 Proof of Theorem 1

Proof. According to the definition of mutual information and under the assumption that all classes are equally likely, we have:

$$\begin{aligned}
&- H_{P_i}(Y|F(X)) \\
=&I_{P_i}(Y, F(X)) - H(Y) \\
=&H_{P_i}(F(X)) - H_{P_i}(F(X)|Y) - H(Y) \\
=&-\frac{1}{C}\sum_{c=1}^{C} \mathop{\mathbb{E}}_{X'\sim P_i^F(X|Y)} \log P_i(X') + \frac{1}{C}\sum_{c=1}^{C} \mathop{\mathbb{E}}_{X'\sim P_i^F(X|Y)} \log P_i(X'|Y=c) - H(Y) \\
=&\frac{1}{C}\sum_{c=1}^{C} \mathop{\mathbb{E}}_{X'\sim P_i^F(X|Y)} \log \frac{P_i(X'|Y=c)}{P_i(X')} - H(Y) \\
=&\frac{1}{C}\sum_{c=1}^{C} KL(P_i(X'|Y=c)||P_i(X')) - H(Y) \\
=&\frac{1}{C}\sum_{c=1}^{C} KL(P_i(F(X)|Y=c)||P_i(F(X))) - H(Y) \\
=&JSD(P_i(F(X)|Y=1), P_i(F(X)|Y=2), \cdots, P_i(F(X)|Y=C)) - H(Y).
\end{aligned} \tag{S. 1}$$

Since $H(Y)$ is a constant, then minimizing $-H_{P_i}(Y|F(X))$ is equivalent to minimizing $JSD(P_i(F(X)|Y=1), P_i(F(X)|Y=2), \cdots, P_i(F(X)|Y=C))$, the global minimum of which is achieved at $P_i(F(X)|Y=1) = P_i(F(X)|Y=2) = \cdots = P_i(F(X)|Y=C)$. □

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

S. 2 Proof of Theorem 2

S. Proposition 1. *Let* $V(F,\{T_i'\}) = \sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)}[\log Q_i^{T_i'}(Y|F(X))]$*. Then the optimal prediction probabilities of* T_i' *are*

$$\langle T_i'^{*}(\mathbf{x}_i')\rangle_c = Q_i^{T_i'^{*}}(Y=c|\mathbf{x}_i') = \frac{P_i(\mathbf{x}_i'|Y=c)}{\sum_{c=1}^{C} P_i(\mathbf{x}_i'|Y=c)}, \tag{S. 2}$$

where $\langle \mathbf{z}\rangle_i$ *denotes the* i^{th} *element of* $\mathbf{z}$*, and* $\mathbf{x}_i' = F(\mathbf{x}_i)$*.*

Proof. For a fixed F, $\min_F \max_{\{T_i'\}} V(F,\{T_i'\})$ reduces to maximizing $V(F,\{T_i'\}_{i=1}^{K})$ w.r.t. $\{T_1', T_2', \cdots, T_K'\}$[1]:

$$\begin{aligned}
&\{\langle T_i'^{*}(\mathbf{x}')\rangle_1, \langle T_i'^{*}(\mathbf{x}')\rangle_2, \cdots, \langle T_i'^{*}(\mathbf{x}')\rangle_C\}\\
=&arg \max_{\{\langle T_i'(\mathbf{x})\rangle_c\}_{c=1}^{C}} \sum_{c=1}^{C} \int_{\mathbf{x}_i'} P_i(\mathbf{x}_i'|Y=c)\log(\langle T_i'(\mathbf{x}_i')\rangle_c) d\mathbf{x}_i',\\
&s.t. \sum_{c=1}^{C} \langle T_i'(\mathbf{x}_i')\rangle_c = 1.
\end{aligned} \tag{S. 3}$$

Maximizing the value function point-wisely and applying Lagrange multipliers, we obtain the following problem:

$$\begin{aligned}
&\{\langle T_i'^{*}(\mathbf{x}')\rangle_1, \langle T_i'^{*}(\mathbf{x}')\rangle_2, \cdots, \langle T_i'^{*}(\mathbf{x}')\rangle_C\}\\
=&arg \max_{\{\langle T_i'(\mathbf{x}')\rangle_c\}_{c=1}^{C}} \sum_{c=1}^{C} P_i(\mathbf{x}_i'|Y=c)\log(\langle T_i'(\mathbf{x}_i')\rangle_c) + \lambda_i(\sum_{c=1}^{C} \langle T_i'(\mathbf{x}_i')\rangle_c - 1).
\end{aligned} \tag{S. 4}$$

Setting the derivative of Eq. S. 4 w.r.t. $\langle T_i'(\mathbf{x}_i')\rangle_c$ to zero, we obtain $\langle T_i'^{*}(\mathbf{x}_i)\rangle_c = -\frac{P_i(\mathbf{x}_i'|Y=c)}{\lambda_i}$. Through substituting the value of $\langle T_i'^{*}(\mathbf{x}_i)\rangle_c$ into the constraint $\sum_{c=1}^{C}\langle T_i'(\mathbf{x}_i')\rangle_c = 1$, we can obtain $\lambda_i = -\sum_{c=1}^{C} P_i(\mathbf{x}_i'|Y=c)$, and thus get the optimal solution $\langle T_i'^{*}(\mathbf{x}_i')\rangle_c = \frac{P_i(\mathbf{x}_i'|Y=c)}{\sum_{c=1}^{C} P_i(\mathbf{x}_i'|Y=c)}$. □

S. Theorem 1. *If* $U(F)$ *is the maximum value of* $V(F,\{T_i'\}_{i=1}^{K})$*, i.e.,*

$$U(F) = \sum_{i=1}^{K}\sum_{c=1}^{C} \mathop{\mathbb{E}}_{X_i\sim P_i(X)}[\log \frac{P_i(X_i'|Y=c)}{\sum_{c=1}^{C} P_i(X_i'|Y=c)}], \tag{S. 5}$$

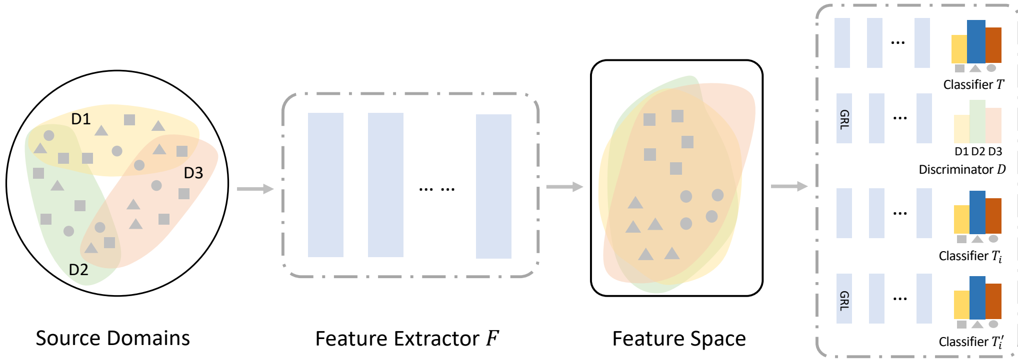
the global minimum of the minimax game is attained if and only if $P_i(X_i'|Y=1) = P_i(X_i'|Y=2) = \cdots = P_i(X_i'|Y=C)$ *for any* $i \in \{1,2,\cdots,K\}$*, where* $U(F)$ *achieves the value* $-KC\log C$*.*

Proof. Adding $KC\log C$ to $U(F)$ can obtain:

$$\begin{aligned}
U(F) + KC\log C &= \sum_{i=1}^{K}\sum_{c=1}^{C}\{\mathop{\mathbb{E}}_{X_i\sim P_i(X)}[\log \frac{P_i(X_i'|Y=c)}{\sum_{c=1}^{C} P_i(X_i'|Y=c)}] + \log C\}\\
&= \sum_{i=1}^{K}\sum_{c=1}^{C} \mathop{\mathbb{E}}_{X_i\sim P_i(X)}[\log \frac{P_i(X_i'|Y=c)}{\frac{1}{C}\sum_{c=1}^{C} P_i(X_i'|Y=c)}]\\
&= \sum_{i=1}^{K}\sum_{c=1}^{C} KL(P_i(X_i'|Y=c)||\frac{1}{C}\sum_{c=1}^{C} P_i(X_i'|Y=c)).
\end{aligned} \tag{S. 6}$$

According to the definition of the Jensen-Shannon divergence, we can obtain $U(F) = -KC\log C + \sum_{i=1}^{K} C\cdot JSD(P_i(X_i'|Y=1), P_i(X_i'|Y=2), \cdots, P_i(X_i'|Y=C))$. Since the JSD between

[1] Here, we only consider T_i' for simplicity.



S. Figure 1: Illustration of our framework. GRL represents the gradient reversal layer. All components are trained, but only F and T are preserved for test.

multiple distributions is always non-negative, and zero iff they are equal, then we have

$$\begin{aligned} &P_1(X_1'|Y=1) = P_1(X_1'|Y=2) = \cdots = P_1(X_1'|Y=C), \\ &P_2(X_2'|Y=1) = P_2(X_2'|Y=2) = \cdots = P_2(X_2'|Y=C), \\ &\cdots \\ &P_K(X_K'|Y=1) = P_K(X_K'|Y=2) = \cdots = P_K(X_K'|Y=C), \end{aligned} \tag{S. 7}$$

and the global minimum of $U(F)$ is $-KC \log C$. □

S. 3 Framework

Here, we provide an illustration of our framework in S. Figure 1 for better understanding of the proposed components. The main module consists of a feature extractor F and a classifier T. In addition, we exploit a domain discriminator D to discriminate domains, and $2K$ classifiers ($\{T_i\}_{i=1}^K$ and $\{T_i'\}_{i=1}^K$) to regularize the generated features. We insert a gradient reversal layer (GRL) [1] between F and D, and F and T_i', respectively. In the inference stage, only the main module (F and T) is required.

References

[1] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.