**SCIENCE CHINA**
Information Sciences

**• Supplementary File •**

# Adversarial Data Splitting for Domain Generalization

Xiang GU[1], Jian SUN[1*] & Zongben XU[1]

[1] *School of Mathematics and Statistics, Xi'an Jiaotong University, Shaanxi* 710049, *China*

## Appendix A Optimization Algorithm for Finding the Hardest $S_v$

This section discusses the details of the optimization algorithm for solving the problem of Eqn. (9) in Sect. 3.2 of the paper. To solve the problem

$$\begin{aligned} &\max_{S_v, A} \sum_{(x,y)\in S_v} l\left(f_w(x), y\right) - \alpha \left\langle \nabla_w l(f_w(x), y), A \right\rangle \\ &s.t. \quad A = g_w^t, S_v \in \Gamma_\xi, \end{aligned} \tag{A1}$$

for optimizing the train/val $(S_t/S_v)$ subsets splitting to increase the domain shift, we alternately update $S_v$ and $A$ by fixing the other one as known.

**Initialization.** We first initialize $A$ with the gradient of a sample randomly selected from $S$.

After initialization, we alternately update $S_v$ and $A$ as follows.

**Updating $\boldsymbol{S_v}$.** Given $A$, $S_v$ is updated by solving

$$\begin{aligned} &\max_{S_v} \sum_{(x,y)\in S_v} l\left(f_w(x), y\right) - \alpha \left\langle \nabla_w l(f_w(x), y), A \right\rangle \\ &s.t. \quad S_v \subset S, |S_v| \,/\, |S| = \xi, \end{aligned} \tag{A2}$$

where the constraints are derived from the definition of $\Gamma_\xi$ (*i.e.*, $\Gamma_\xi = \{S_v \subset S, |S_v| \,/\, |S| = \xi\}$). Equation (A2) indicates that the optimal $S_v$ consists of $\xi\,|S|$ samples that have the largest values of $l\left(f_w(x), y\right) - \alpha \left\langle \nabla_w l(f_w(x), y), A \right\rangle$. Thus, given $A$, we compute and rank the values of $l\left(f_w(x), y\right) - \alpha \left\langle \nabla_w l(f_w(x), y), A \right\rangle$ for all $(x, y) \in S$ and select the largest $\xi|S|$ samples to constitute the $S_v$.

**Updating $\boldsymbol{A}$.** Given $S_v$ ($S_t = S - S_v$ is then given), we update $A$ to satisfy the constraint $A = g_w^t$ in Eqn. (A1). Then, $A$ is updated by

$$A = g_w^t = \frac{1}{|S_t|} \sum_{(x,y)\in S_t} \nabla_w l(f_w(x), y). \tag{A3}$$

Equation (A3) is based on the definition of $g_w^t$ that

$$\begin{aligned} g_w^t &= \nabla_\theta \mathcal{L}(\theta; S_t, w) \\ &= \frac{1}{|S_t|} \sum_{(x,y)\in S_t} \nabla_\theta l(f(x,\theta), y)|_{\theta=w} \\ &= \frac{1}{|S_t|} \sum_{(x,y)\in S_t} \nabla_w l(f(x,w), y), \end{aligned} \tag{A4}$$

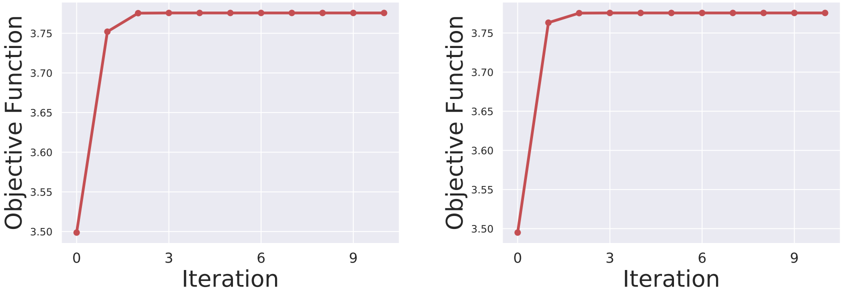
where the second equation utilizes the fact that $w$ is the initialization of $\theta$.

We show empirically the convergence of this alternate iteration algorithm in Fig. A1, with the values of the objective function in Eqn. (A1). Figure A1 shows that the values of the objective function converge after only a few iterations. For the theoretical analysis of the convergence, we take it as our future work.

## Appendix B Proof of Theorem 1

This section proves Theorem 1 in Sect. 4.2 of the paper. We first introduce the VC-dimension-based generalization bound and the domain adaptation theory, then present two lemmas that will be used in the proof, and finally give the proof of Theorem 1.

[*] Corresponding author (email: jiansun@xjtu.edu.cn)



(a) Run 1 (b) Run 2

**Figure A1** Convergence of the alternate iteration for finding the hardest $S_v$. (a) and (b) respectively show the values of objective function in Eqn. (A1) in two different runs with different initializations.

## Appendix B.1 Preliminary

**VC-dimension-based generalization bound [1].**

**Theorem A-1.** Let $S$ be the set of training data i.i.d. sampled for distribution $\mathcal{P}$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have $\forall h$ $(h : \mathcal{X} \to \{0,1\})$ in hypothesis space $\mathcal{H}$,

$$|\epsilon_{\mathcal{P}}(h) - \hat{\epsilon}_S(h)| \leqslant \sqrt{\frac{8}{|S|}\left(VC(\mathcal{H}) \log \frac{2e\,|S|}{VC(\mathcal{H})} + \frac{4}{\delta}\right)}. \tag{B1}$$

where $\epsilon_{\mathcal{P}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathbb{I}_{\{(h(x))\neq y\}}]$ and $\hat{\epsilon}_S(h) = \frac{1}{|S|}\sum_{(x,y)\in S}\mathbb{I}_{\{(h(x))\neq y\}}$.

**Domain adaptation theory [2, 3].**

**Theorem A-2.** For any $h$ in hypothesis space $\mathcal{H}$, we have

$$\epsilon_{\mathcal{Q}}(h) \leqslant \epsilon_{\mathcal{P}}(h) + \frac{1}{2} d_{\mathcal{H}}(\mathcal{P}, \mathcal{Q}) + \lambda^*, \tag{B2}$$

where $\lambda^* \geqslant \inf_{h' \in \mathcal{H}}\{\epsilon_{\mathcal{P}}(h') + \epsilon_{\mathcal{Q}}(h')\}$ and

$$d_{\mathcal{H}}(\mathcal{P}, \mathcal{Q}) = 2 \sup_{h\in\mathcal{H}} |\mathbb{E}_{\mathcal{P}}[h = 1] - \mathbb{E}_{\mathcal{Q}}[h = 1]| \tag{B3}$$

is the $\mathcal{H}$-divergence.

## Appendix B.2 Lemmas

**Lemma A-1.** For any $S_v \in \Gamma_\xi$ and $S_t = S - S_v$, $\forall\delta \in (0,1)$, with probability at least $1-\delta$, we have $\forall f \in \mathcal{H}_{S_t}$,

$$|\epsilon^{\Psi}_{\mathcal{P}}(f) - \hat{\epsilon}^{\Psi}_{S_v}(f)| \leqslant \sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi}_{S_t}) \log \frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi}_{S_t})} + \frac{4}{\delta}\right)}, \tag{B4}$$

where $\epsilon^{\Psi}_{\mathcal{P}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathbb{I}_{\{\Psi(f(x))\neq y\}}]$ is the generalization error on distribution $\mathcal{P}$, $\hat{\epsilon}^{\Psi}_{S_v}(f) = \frac{1}{|S_v|}\sum_{(x,y)\in S_v}\mathbb{I}_{\{\Psi(f(x))\neq y\}}$ is the empirical error, $\mathcal{H}^{\Psi}_{S_t} = \{\Psi \circ f : f \in \mathcal{H}_{S_t}\}$, $\mathcal{H}_{S_t}$ is defined in Sect. 4.2 of the paper, $VC(\mathcal{H}^{\Psi}_{S_t})$ is the VC-dimension of $\mathcal{H}^{\Psi}_{S_t}$, and $\Psi(\cdot)$ is the prediction rule such as the Bayes Optimal Predictor, *i.e.*, $\Psi(f(x)) = \mathbb{I}_{\{f(x)\geqslant\frac{1}{2}\}}$.

**Proof:**

From the definition of $\mathcal{H}^{\Psi}_{S_t}$, for any $f \in \mathcal{H}_{S_t}$, there exists a $h_f \in \mathcal{H}^{\Psi}_{S_t}$ such that $h_f = \Psi \circ f$. Applying Theorem A-1, with probability at least $1-\delta$, we have $\forall f \in \mathcal{H}_{S_t}$,

$$\begin{aligned} &|\epsilon^{\Psi}_{\mathcal{P}}(f) - \hat{\epsilon}^{\Psi}_{S_v}(f)| \\ =&|\epsilon_{\mathcal{P}}(h_f) - \hat{\epsilon}_{S_v}(h_f)| \\ \leqslant&\sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi}_{S_t}) \log \frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi}_{S_t})} + \frac{4}{\delta}\right)}. \end{aligned} \tag{B5}$$

**Lemma A-2.** For any $S_v \in \Gamma_\xi$ and $S_t = S - S_v$, let $g = \arg\inf_{f\in\mathcal{H}_{S_t}} \epsilon^{\Psi}_{\mathcal{P}}(f)$ and $h = \arg\inf_{f\in\mathcal{H}_{S_t}} \hat{\epsilon}^{\Psi}_{S_v}(f)$, then $\forall\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\epsilon^{\Psi}_{\mathcal{P}}(g) \geqslant \hat{\epsilon}^{\Psi}_{S_v}(h) - \sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi}_{S_t}) \log \frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi}_{S_t})} + \frac{4}{\delta}\right)}. \tag{B6}$$

**Proof:**

From the definition of $g$ and $h$, we have $\hat{\epsilon}^{\Psi}_{S_v}(g) \geqslant \hat{\epsilon}^{\Psi}_{S_v}(h)$. $\forall \delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\epsilon^{\Psi}_{\mathcal{P}}(g) - \hat{\epsilon}^{\Psi}_{S_v}(h)\\
=&\epsilon^{\Psi}_{\mathcal{P}}(g) - \hat{\epsilon}^{\Psi}_{S_v}(g) + \hat{\epsilon}^{\Psi}_{S_v}(g) - \hat{\epsilon}^{\Psi}_{S_v}(h)\\
\geqslant&\epsilon^{\Psi}_{\mathcal{P}}(g) - \hat{\epsilon}^{\Psi}_{S_v}(g)\\
\geqslant& -\sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi}_{S_t})\log\frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi}_{S_t})}+\frac{4}{\delta}\right)}.
\end{aligned} \tag{B7}$$

In the last inequality, we utilize Lemma A-1. Thus, Eqn. (B6) holds.

## Appendix B.3 Proof of Theorem 1

**Proof:**

We denote $\mathcal{H}^{\Psi_l}$ as the hypothesis space such that $\forall h \in \mathcal{H}^{\Psi_l}$,

$$h(x) = \Psi_l(f(x)) = \begin{cases} 1 & \text{if } l(f(x),y) > \gamma,\\ 0 & \text{otherwise },\end{cases} \tag{B8}$$

for $f \in \mathcal{H}$. Then

$$\begin{aligned}
d_{\mathcal{H}^{\Psi_l}}(\mathcal{P},\mathcal{Q}) =&2\sup_{h\in\mathcal{H}^{\Psi_l}}\left|\mathbb{E}_{\mathcal{P}}[h=1]-\mathbb{E}_{\mathcal{Q}}[h=1]\right|\\
=&2\sup_{f\in\mathcal{H}}\left|\mathbb{E}_{\mathcal{P}}[\Psi_l(f(x))=1]-\mathbb{E}_{\mathcal{Q}}[\Psi_l(f(x))=1]\right|\\
=&2\sup_{f\in\mathcal{H}}\left|\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]-\mathbb{E}_{\mathcal{Q}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]\right|\\
=&2\sup_{f\in\mathcal{H}}\left\{\mathbb{E}_{\mathcal{Q}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]-\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]\right\}\\
\leqslant&2\sup_{f\in\mathcal{H}}\mathbb{E}_{\mathcal{Q}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]-2\inf_{f\in\mathcal{H}}\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}].
\end{aligned} \tag{B9}$$

In the fourth equation, we utilize the assumption that $\mathbb{E}_{\mathcal{Q}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}] \geqslant \mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]$. Given any $S_v \in \Gamma_\xi$ and $S_t = S - S_v$, we replace $\mathcal{H}$ by $\mathcal{H}_{S_t}$, then

$$\begin{aligned}
d_{\mathcal{H}^{\Psi_l}_{S_t}}(\mathcal{P},\mathcal{Q}) \leqslant&2\sup_{f\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{Q}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]-2\inf_{f\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]\\
=&2C_1(\mathcal{Q},S_t)-2\inf_{f\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]
\end{aligned} \tag{B10}$$

where $C_1(\mathcal{Q},S_t)=\sup_{f\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{Q}}[\mathbb{I}_{\{l(f(x),y)>\gamma\}}]$. Applying Theorem A-2, for any $f \in \mathcal{H}_{S_t}$, we have

$$\epsilon^{\Psi_l}_{\mathcal{Q}}(f) \leqslant \epsilon^{\Psi_l}_{\mathcal{P}}(f) + C_1(\mathcal{Q},S_t) - \inf_{f'\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f'(x),y)>\gamma\}}] + \lambda^*(S_t), \tag{B11}$$

where $\lambda^*(S_t) \geqslant \inf_{f'\in\mathcal{H}_{S_t}}\{\epsilon^{\Psi_l}_{\mathcal{P}}(f') + \epsilon^{\Psi_l}_{\mathcal{Q}}(f')\}$. We let $C^*(\mathcal{Q},S_t) = C_1(\mathcal{Q},S_t) + \lambda^*(S_t)$, then

$$\epsilon^{\Psi_l}_{\mathcal{Q}}(f) \leqslant \epsilon^{\Psi_l}_{\mathcal{P}}(f) - \inf_{f'\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f'(x),y)>\gamma\}}] + C^*(\mathcal{Q},S_t). \tag{B12}$$

Applying Lemma A-1 to the first term of the right side in Eqn. (B12), $\forall\delta \in (0,1)$, with probability at least $1-\delta$, we have $\forall f \in \mathcal{H}_{S_t}$,

$$\epsilon^{\Psi_l}_{\mathcal{P}}(f) \leqslant \hat{\epsilon}^{\Psi_l}_{S_v}(f) + \sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi_l}_{S_t})\log\frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi_l}_{S_t})}+\frac{4}{\delta}\right)}. \tag{B13}$$

Applying Lemma A-2 to the third term of the right side in Eqn. (B12), $\forall\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\inf_{f'\in\mathcal{H}_{S_t}}\mathbb{E}_{\mathcal{P}}[\mathbb{I}_{\{l(f'(x),y)>\gamma\}}] \geqslant \inf_{f'\in\mathcal{H}_{S_t}}\frac{1}{|S_v|}\sum_{(x,y)\in S_v}\mathbb{I}_{\{l(f'(x),y)>\gamma\}} - \sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi_l}_{S_t})\log\frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi_l}_{S_t})}+\frac{4}{\delta}\right)}. \tag{B14}$$

Combining Eqns. (B12), (B13), and (B14) and using the union bound, for any $\delta \in (0,1)$, with probability at least $1-2\delta$, we have $\forall f \in \mathcal{H}_{S_t}$,

$$\epsilon^{\Psi_l}_{\mathcal{Q}}(f) \leqslant \hat{\epsilon}^{\Psi_l}_{S_v}(f) + B(S_v) + 2\sqrt{\frac{8}{|S_v|}\left(VC(\mathcal{H}^{\Psi_l}_{S_t})\log\frac{2e\,|S_v|}{VC(\mathcal{H}^{\Psi_l}_{S_t})}+\frac{4}{\delta}\right)} + C^*(\mathcal{Q},S_t), \tag{B15}$$

where

$$B(S_v) = -\inf_{f' \in \mathcal{H}_{S_t}} \frac{1}{|S_v|} \sum_{(x,y)\in S_v} \mathbb{I}_{\{l(f'(x),y)>\gamma\}}. \tag{B16}$$

Using the fact that $|S_v| = \xi |S|$ and let $C_H = \sup_{S'_v \in \Gamma_\xi} VC(\mathcal{H}^{\Psi_l}_{S-S'_v}) \log \frac{2e\xi|S|}{VC(\mathcal{H}^{\Psi_l}_{S-S'_v})}$, we have

$$\epsilon^{\Psi_l}_{\mathcal{Q}}(f) \leqslant \hat{\epsilon}^{\Psi_l}_{S_v}(f) + B(S_v) + 2\sqrt{\frac{8}{\xi\,|S|}\left(C_H + \frac{4}{\delta}\right)} + C^*(\mathcal{Q}, S_t). \tag{B17}$$

## Appendix C Analysis of $L_2$-normalization Mitigating Gradient Explosion

Meta-learning approaches for DG [4,5] often suffer from gradient explosion, *i.e.*, the norm of gradient of loss *w.r.t.* the parameters of model is infinite. We find experimentally that the gradient explosion can be mitigated in our approach by introducing the $L_2$-normalization, as in Sect. 5.2 of the paper. We next theoretically analyze the reasons for this finding.

For the sake of simplicity, we analyze the gradient norm of loss *w.r.t.* parameters of the classifier in the meta-learning process for DG, with feature extractor as a fixed function. Without loss of generality, we consider the case that $K = 2$ (*i.e.*, binary classification), $s = 1$ and $m = 0$. Then we have the following proposition.

**Proposition A-1.** Under the above setting, if the input feature of the classifier is $L_2$-normalized, the gradient norm of the generalization loss *w.r.t.* parameters of the classifier in the meta-learning process of DG is bounded.

**Proof:**

Given feature $z$, the loss of binary classification is

$$\mathcal{L}(w; z) = -y \log(\sigma(w^T z)) - (1-y)\log(1 - \sigma(w^T z)), \tag{C1}$$

where $\sigma$ is the sigmoid function. Let $w' = w - \alpha \nabla_w \mathcal{L}(w; z)$, then

$$\nabla_w \mathcal{L}(w'; z) = (I - \alpha H) \nabla_{w'} \mathcal{L}(w'; z), \tag{C2}$$

where $H$ is the Hessian matrix. The gradient norm

$$\left\|\nabla_w \mathcal{L}(w'; z)\right\| \leqslant \|I - \alpha H\| \left\|\nabla_{w'} \mathcal{L}(w'; z)\right\| \leqslant (1 + |\alpha| \, \|H\|) \left\|\nabla_{w'} \mathcal{L}(w'; z)\right\|. \tag{C3}$$

Since $\nabla_{w'} \mathcal{L}(w'; z) = (p - y)z$ and $H = p(1-p)zz^T$ where $p = \sigma(w^T z)$,

$$\|H\| = \sup_{u:\|u\|=1} \|Hu\| \leqslant \sup_{u:\|u\|=1} \left\|zz^T u\right\| \leqslant \|z\|^2 \tag{C4}$$

and

$$\left\|\nabla_{w'} \mathcal{L}(w'; z)\right\| \leqslant \|z\|. \tag{C5}$$

If $\|z\| = 1$, combining Eqns. (C3), (C4) and (C5), we have

$$\left\|\nabla_w \mathcal{L}(w'; z)\right\| \leqslant 1 + |\alpha|. \tag{C6}$$

Hence the norm of gradient is bounded.

According to Proposition A-1, $L_2$-normalization can mitigate gradient explosion under the above setting. The analysis of gradient norm of loss *w.r.t.* parameters of both classifier and feature extractor in the meta-learning process is much more complex, left for our future work.

## Appendix D Applying ADS to Large-Scale Datasets

In this section, we discuss how to apply our method of ADS to large-scale datasets. The main bottleneck of ADS to scale up is that the computing of gradients on all data for learning the splitting is time-consuming for large-scale datasets. To implement our ADS on large-scale datasets, we can randomly sample a subset of training data to learn the splitting and train the model on it, in each iteration of the alternate training algorithm in Sect. 3.2 of the paper.

### References

1 Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

2 Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *ML*, 79(1-2):151–175, 2010.

3 Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, 2007.

4 Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019.

5 Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.