Supplementary Material for TIPI: Test Time Adaptation with Transformation Invariance

A. Tuan Nguyen[1], Thanh Nguyen-Tang[2],
Ser-Nam Lim[3*], Philip H.S. Torr[1*]
[1]University of Oxford, [2]Johns Hopkins University, [3]Meta AI
tuan@robots.ox.ac.uk, nguyent@cs.jhu.edu
sernamlim@meta.com, philip.torr@eng.ox.ac.uk

A. Derivations and Proofs

A.1. Assumption 1

First of all, the formulas for the two mutual information terms are:

$$I_T(x,y) = \int_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_T(x,y) \log \frac{p_T(x,y)}{p_T(x)p_T(y)},$$
$$I_T(x^t,y) = \int_{x^t\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_T(x^t,y) \log \frac{p_T(x^t,y)}{p_T(x^t)p_T(y)}.$$

Recall that Assumption 1 states:

$$\begin{aligned}
&I_T(x,y) = I_T(x^{tr},y)\\
\Leftrightarrow&\mathbb{E}_{p_T(x,y)}\left[\log\frac{p_T(x,y)}{p_T(x)p_T(y)}\right]\\
&\quad=\mathbb{E}_{p_T(x^{tr},y)}\left[\log\frac{p_T(x^{tr},y)}{p_T(x^{tr})p_T(y)}\right]\\
\Leftrightarrow&\mathbb{E}_{p_T(x,y)}\left[\log\frac{p_T(y|x)}{p_T(y)}\right]\\
&\quad=\mathbb{E}_{p_T(x^{tr},y)}\left[\log\frac{p_T(y|x^{tr})}{p_T(y)}\right]\\
&\quad=\mathbb{E}_{p_T(x^{tr},y)}\left[\log\frac{p_T(x^{tr},y)}{p_T(x^{tr})p_T(y)}\right]\\
\Leftrightarrow&\mathbb{E}_{p_T(x,x^{tr},y)}\left[\log\frac{p_T(y|x)}{p_T(y)}\right]\\
&\quad=\mathbb{E}_{p_T(x,x^{tr},y)}\left[\log\frac{p_T(y|x^{tr})}{p_T(y)}\right]\\
\Leftrightarrow&\mathbb{E}_{p_T(x,x^{tr},y)}\left[\log p_T(y|x)-\log p_T(y|x^{tr})\right]=0\\
\Leftrightarrow&\mathbb{E}_{p_T(x,x^{tr})}\left[\mathbb{E}_{p_T(y|x)}\left[\log p_T(y|x)-\log p_T(y|x^{tr})\right]\right]\\
&\quad=0\\
\Leftrightarrow&\mathbb{E}_{p_T(x,x^{tr})}\left[\mathrm{KL}[p_T(y|x),p_T(y|x^{tr})]\right]=0
\end{aligned}$$

*The last two authors contributed equally

A.2. Proposition 1

Note that our result also holds if l is other distance/divergence (such as the Jensen Shannon divergence), only with a modified coefficient of the regularizer term. However, in this paper, we consider l as the l_1 distance.

Proof. First of all, due to Pinsker's inequality, we have that:

$$\begin{aligned}
&l[p_T(y|x),p_T(y|x^{tr})]\le\sqrt{2\mathrm{KL}[p_T(y|x),p_T(y|x^{tr})]}\\
\Rightarrow&\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_T(y|x),p_T(y|x^{tr})]\right]\\
&\quad\le\mathbb{E}_{p_T(x,x^{tr})}\left[\sqrt{2\mathrm{KL}[p_T(y|x),p_T(y|x^{tr})]}\right]\\
&\quad\le\sqrt{\mathbb{E}_{p_T(x,x^{tr})}\left[2\mathrm{KL}[p_T(y|x),p_T(y|x^{tr})]\right]}\\
&\quad=0\\
\Rightarrow&\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_T(y|x),p_T(y|x^{tr})]\right]=0
\end{aligned}$$

We have:

$$\begin{aligned}
&\ell(\theta,p_T(x,y))-\ell(\theta,p_T(x^{tr},y))\\
=&\mathbb{E}_{p_T(x)}\left[l[p_T(y|x),p_\theta(y|x)]\right]\\
&\quad-\mathbb{E}_{p_T(x^{tr})}\left[l[p_T(y|x^{tr}),p_\theta(y|x^{tr})]\right]\\
=&\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_T(y|x),p_\theta(y|x)]\right]\\
&\quad-\mathbb{E}_{p_T(x,x^{tr}}\left[l[p_T(y|x^{tr}),p_\theta(y|x^{tr})]\right]\\
&\quad-\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_T(y|x),p_T(y|x^{tr})]\right]\\
\le&\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_\theta(y|x^{tr}),p_\theta(y|x)]\right]
\end{aligned}$$

This is because l satisfies the triangle inequality, so $l[p_T(y|x),p_\theta(y|x)] - l[p_T(y|x^{tr}),p_\theta(y|x^{tr})] - l[p_T(y|x),p_T(y|x^{tr})] \le l[p_\theta(y|x^{tr}),p_\theta(y|x)]$.

Table 1. **ImageNet-C, Batchsize 64**: Results for 15 types of corruption with the highest level of severity

Method	Classification Error (lower is better)															
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Average
Source	97.0	96.3	97.4	82.1	90.3	85.3	77.5	83.4	76.9	76	40.9	94.6	83.5	79.1	67.4	81.8
ETA	64.9	62.1	63.4	66.1	67.1	52.2	47.4	48.1	54.2	39.9	32.1	55.0	42.1	39.1	45.1	51.9
ETA+TIPI	**63.0**	**60.9**	**62.1**	**65.2**	**65.4**	**51.6**	**46.7**	**47.6**	**53.7**	**39.9**	**31.6**	**54.1**	**41.2**	**38.2**	**43.6**	**51.0**

Table 2. **ImageNet-C, Batchsize 2**: Results for 15 types of corruption with the highest level of severity

Method	Classification Error (lower is better)															
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Average
Source	97.0	96.3	97.4	**82.1**	90.3	85.3	77.5	83.4	76.9	76	40.9	94.6	83.5	79.1	67.4	81.8
ETA	97.7	97.4	97.5	97.8	97.9	95.1	93.3	92.2	92.3	89.2	78.4	96.4	90.2	90.1	92.3	93.2
ETA+TIPI	**87.8**	**88.8**	**87.4**	88.3	**89.3**	**82.4**	**70.5**	**73.2**	**75.8**	**60.7**	**40.8**	**93.9**	**64.0**	**58.8**	**65.4**	**75.4**

Using Pinsker's inequality again, we have that:

$$\begin{aligned}
&\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_\theta(y|x^{tr}), p_\theta(y|x)]\right] \\
\leq &\mathbb{E}_{p_T(x,x^{tr})}\left[\sqrt{2\mathrm{KL}[p_\theta(y|x^{tr}), p_\theta(y|x)]}\right] \\
\leq &\mathbb{E}_{p_T(x)}\left[\sqrt{2 \max_{x^{tr}\sim T(x^{tr}|x)} \mathrm{KL}[p_\theta(y|x^{tr}), p_\theta(y|x)]}\right]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{p_T(x,x^{tr})}\left[l[p_\theta(y|x^{tr}), p_\theta(y|x)]\right] \\
\leq &\mathbb{E}_{p_T(x,x^{tr})}\left[\sqrt{2\mathrm{KL}[p_\theta(y|x), p_\theta(y|x^{tr})]}\right] \\
\leq &\mathbb{E}_{p_T(x)}\left[\sqrt{2 \max_{x^{tr}\sim T(x^{tr}|x)} \mathrm{KL}[p_\theta(y|x), p_\theta(y|x^{tr})]}\right]
\end{aligned}$$

This concludes our proof. □

B. Additional Experiments

B.1. Incorporating TIPI into EATA [1]

EATA [1] proposes a datapoints selection strategy for TENT, and achieves state-of-the-art performance on the test time adaptation task. As discussed before, this line of research (datapoint selection for surrogate optimization) is complementary to ours. Indeed, in this subsection, we show that TIPI can be incorporated into EATA, thereby improving EATA's robustness (EATA only uses TENT and is not robust against small batchsizes). We refer the readers to [1] for a detailed discussion on their datapoint selection strategy.

We conduct the experiment with ETA, the variation without weight regularization (ETA outperforms EATA for out-of-distribution data). Specifically, we compare ETA to ETA+TIPI, a variant of ETA with the same datapoint selection strategy, but with the TIPI surrogate objective. Following [1], we conduct the experiment on ImageNet-C with the highest level of severity. The 15 types of corruption are: gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg compression.

Table 1 and Table 2 show the results of this comparison for batch sizes 64 and 2. For a large batch size, ETA+TIPI further improves over ETA, pushing a new SOTA result. For a small batch size, incorporating the TIPI objective help to stabilize the optimization, thus avoiding network collapse.

Furthermore, note that we keep the datapoint selection strategy in EATA as is in this experiment. This strategy is developed specifically for the TENT objective and we did not make any modifications for TIPI. Yet, using it for TIPI already improves the model's performance. This shows that indeed the two lines of research are complementary. It also suggests that investigating a datapoint selection strategy tailored for TIPI is a promising direction, and could potentially improve TIPI's performance even further.

References

[1] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. *arXiv preprint arXiv:2204.02610*, 2022. 2