Appendix

Table of Contents

A Some Prerequisite Definitions and Useful Lemmas **15**

B Omitted Proofs and Additional Results in Section 4 **17**
- B.1 Proof of Theorem 4.1 17
- B.2 Proof of Corollary 4.2 17
- B.3 Proof of Theorem 4.2 17
- B.4 Proof of Theorem 4.3 18
- B.5 Proof of Theorem 4.4 18
- B.6 Proof of Corollary 4.3 18
- B.7 Proof of Theorem 4.5 18
- B.8 Additional Results: Sample Complexity Bounds 19
- B.9 Additional Discussions on the Convergence of Empirical KL Divergence 20

C Omitted Proofs and Additional Discussions in Section 5 **20**
- C.1 Additional Discussion on Theorem 5.1 20
- C.2 Proof of Theorem 5.1 20
- C.3 Proof of Corollary 5.1 21
- C.4 Proof of Theorem 5.2 23
- C.5 Proof of Corollary 5.2 23
- C.6 Proof of Theorem 5.3 23
- C.7 Derivation of Eq. (6) 25
- C.8 Additional Discussion on LIMIT 25

D Experiment Details **25**
- D.1 Objective Functions of Gradient Penalty and Controlling Label Information . . . 25
- D.2 Algorithm of Controlling Label Information and Additional Results of ERM-CL 26
- D.3 Architectures and Hyperparameters 27
- D.4 Additional Experimental Results 27
- D.5 License of the Assets 28

E Limitations **28**

A Some Prerequisite Definitions and Useful Lemmas

Definition 3 (Wasserstein Distance)**.** *Let* $d(\cdot,\cdot)$ *be a metric and let* P *and* Q *be probability measures on* $\mathcal{X}$*. Denote* $\Gamma(P,Q)$ *as the set of all couplings of* P *and* Q *(i.e. the set of all joint distributions on* $\mathcal{X}\times\mathcal{X}$ *with two marginals being* P *and* Q*), then the Wasserstein Distance of order one between* P *and* Q *is defined as* $\mathbb{W}(P,Q) \triangleq \inf_{\gamma\in\Gamma(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} d(x,x')d\gamma(x,x')$.

Definition 4 (Total Variation)**.** *The total variation between two probability measures* P *and* Q *is* $\mathrm{TV}(P,Q) \triangleq \sup_E |P(E)-Q(E)|$, *where the supremum is over all measurable set* E.

Note that the total variation equals to the Wasserstein distance under the discrete metric (or Hamming distortion) $d(x,x') = \mathbb{1}(x \neq x')$ where $\mathbb{1}$ is the indicator function.

Definition 5 (Lautum Information [43])**.** *Define the lautum information between* X *and* Y *as* $L(X;Y) \triangleq \mathrm{D}_{\mathrm{KL}}(P_X P_Y || P_{XY})$.

The key quantity in most information-theoretic generalization bounds is the mutual information between algorithm's input and output. Specifically, the core technique behind these bounds is the well-known Donsker-Varadhan representation of KL divergence [47, Theorem 3.5].

Lemma A.1 (Donsker and Varadhan's variational formula)**.** *Let Q, P be probability measures on Θ, for any bounded measurable function $f:\Theta\to\mathbb{R}$, we have* $\mathrm{D_{KL}}(Q||P)=\sup_f\mathbb{E}_{\theta\sim Q}\left[f(\theta)\right]-\log\mathbb{E}_{\theta\sim P}\left[\exp f(\theta)\right]$.

Remark A.1. *Motivated by the classic f-divergence, Acuna et al. [8] proposed a discrepancy measure called $\mathrm{D}_{\mathcal{H}\phi}$-discrepancy. Since KL divergence belongs to the family of f-divergences (e.g., choosing $x\log x$ as the Fenchel conjugate function) and both [8] and our work invoke the variational representation of the divergence, it seems our work (in Section 4) is related to theirs. However, the variational characterization of f-divergence used in [8] is based on the result of [57], and the Donsker-Varadhan representation of KL divergence (see Lemma A.1) used in this paper cannot be directly recovered from their variational characterization [58, 59]. Indeed, simply choosing $x\log x$ as the conjugate function will lead to a weaker bound than Lemma A.1. Thus, our results (in Section 4) cannot be directly recovered from the results in [8].*

Lemma A.2. *Let Q and P be probability measures on Θ. Let $\theta'\sim Q$ and $\theta\sim P$. If $g(\theta)$ is R-subgaussian, then,*

$$\left|\mathbb{E}_{\theta'\sim Q}\left[g(\theta')\right]-\mathbb{E}_{\theta\sim P}\left[g(\theta)\right]\right|\leq\sqrt{2R^2\mathrm{D_{KL}}(Q||P)}.$$

Proof. Let $f=t\cdot g$ for any $t\in\mathbb{R}$, by Lemma A.1, we have

$$\begin{aligned}\mathrm{D_{KL}}(Q||P)\geq&\sup_t\mathbb{E}_{\theta'\sim Q}\left[tg(\theta')\right]-\log\mathbb{E}_{\theta\sim P}\left[\exp t\cdot g(\theta)\right]\\=&\sup_t\mathbb{E}_{\theta'\sim Q}\left[tg(\theta')\right]-\log\mathbb{E}_{\theta\sim P}\left[\exp t(g(\theta)-\mathbb{E}_{\theta\sim P}\left[g(\theta)\right]+\mathbb{E}_{\theta\sim P}\left[g(\theta)\right])\right]\\=&\sup_t\mathbb{E}_{\theta'\sim Q}\left[tg(\theta')\right]-\mathbb{E}_{\theta\sim P}\left[tg(\theta)\right]-\log\mathbb{E}_{\theta\sim P}\left[\exp t(g(\theta)-\mathbb{E}_{\theta\sim P}\left[g(\theta)\right])\right]\\\geq&\sup_t t\left(\mathbb{E}_{\theta'\sim Q}\left[g(\theta')\right]-\mathbb{E}_{\theta\sim P}\left[g(\theta)\right]\right)-t^2R^2/2,\end{aligned}$$

where the last inequality is by the subgaussianity of $g(\theta)$.

Then consider the case of $t>0$ and $t<0$ ($t=0$ is trivial), by AM–GM inequality (i.e. the arithmetic mean is greater than or equal to the geometric mean), the following is straightforward,

$$\left|\mathbb{E}_{\theta'\sim Q}\left[g(\theta')\right]-\mathbb{E}_{\theta\sim P}\left[g(\theta)\right]\right|\leq\sqrt{2R^2\mathrm{D_{KL}}(Q||P)}.$$

This completes the proof. □

The following lemma is the Kantorovich–Rubinstein duality of Wasserstein distance [60].

Lemma A.3 (KR duality)**.** *For any two distributions P and Q, we have*

$$\mathbb{W}(P,Q)=\sup_{f\in 1-\mathrm{Lip}(\rho)}\int_{\mathcal{X}}fdP-\int_{\mathcal{X}}fdQ,$$

where the supremum is taken over all 1-Lipschitz functions in the metric d, i.e. $|f(x)-f(x')|\leq d(x,x')$ for any $x,x'\in\mathcal{X}$.

To connect total variation with KL divergence , we will use Pinsker's inequality [47, Theorem 6.5] and Bretagnolle-Huber inequality [48, Lemma 2.1] in this paper, for more discussion about these two inequalities, we refer readers to [61].

Lemma A.4 (Pinsker's inequality)**.** $\mathrm{TV}(P,Q)\leq\sqrt{\frac{1}{2}\mathrm{D_{KL}}(P||Q)}$.

Lemma A.5 (Bretagnolle-Huber inequality)**.** $\mathrm{TV}(P,Q)\leq\sqrt{1-e^{-\mathrm{D_{KL}}(P||Q)}}$.

Below is the variational formula, or golden formula of mutual information.

Lemma A.6 (Polyanskiy and Wu [47, Corollary 3.1.])**.** *For two random variables X and Y, we have*

$$I(X;Y)=\inf_P\mathbb{E}_X\left[\mathrm{D_{KL}}(Q_{Y|X}||P)\right],$$

where the infimum is achieved at $P=Q_Y$.

$$\begin{tikzcd}
 & S'_{X'} \arrow[d] \\
S \arrow[r] \arrow[d] & W \arrow[dl] \\
F &
\end{tikzcd}$$

Figure 1: The relationship between random variables in UDA, where $F = R_{\mu'}(W) - R_S(W)$.

B Omitted Proofs and Additional Results in Section 4

B.1 Proof of Theorem 4.1

Proof. Let $Q = \mu'$, $P = \mu$ and $g = \ell$, then Theorem 4.1 comes directly from Lemma A.2. $\square$

B.2 Proof of Corollary 4.2

Proof. When the loss is bounded in $[0, M]$, similar to the proof of Theorem 4.1, let $Q = \mu$, $P = \mu'$ and $g = \ell$, then the following bound also holds by Lemma A.2,

$$\left|\widetilde{\mathrm{Err}}(w)\right| \leq \sqrt{\frac{M^2}{2}\mathrm{D}_{\mathrm{KL}}(\mu||\mu')}.$$

Then, recall Theorem 4.1 and by $\min\{A, B\} \leq \frac{1}{2}(A + B)$, the remaining part is straightforward,

$$\left|\widetilde{\mathrm{Err}}(w)\right| \leq \frac{M}{\sqrt{2}}\sqrt{\min\{\mathrm{D}_{\mathrm{KL}}(\mu||\mu'), \mathrm{D}_{\mathrm{KL}}(\mu'||\mu)\}} \leq \frac{M}{2}\sqrt{\mathrm{D}_{\mathrm{KL}}(\mu||\mu') + \mathrm{D}_{\mathrm{KL}}(\mu'||\mu)}.$$

This completes the proof. $\square$

B.3 Proof of Theorem 4.2

Proof. Let $w^* = \arg\min_{w\in\mathcal{W}} \mathbb{E}_{Z'}\left[\ell(f_w(X'), Y')\right] + \mathbb{E}_Z\left[\ell(f_w(X), Y)\right]$. By Lemma A.1,

$$\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X) \geq \sup_{t\in\mathbb{R}, w\in\mathcal{W}} \mathbb{E}_{X'}\left[t\ell(f_w(X'), f_{w^*}(X'))\right] - \log \mathbb{E}_X\left[e^{t\ell(f_w(X), f_{w^*}(X))}\right].$$

Recall that $\ell(f_{w'}(X), f_w(X))$ is R-subgaussian, by using Lemma A.2 (let $Q = P_{X'}$, $P = P_X$ and $g(\cdot) = \ell(f_{w'}(\cdot), f_w(\cdot))$), we have

$$\left|\mathbb{E}_{X'}\left[\ell(f_w(X'), f_{w^*}(X'))\right] - \mathbb{E}_X\left[\ell(f_w(X), f_{w^*}(X))\right]\right| \leq \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)}. \tag{7}$$

For any $f_w \in \mathcal{F}$, by the triangle property of the loss, we have

$$\begin{aligned}
&\mathbb{E}_{Z'}\left[\ell(f_w(X'), Y')\right]\\
\leq&\mathbb{E}_{X'}\left[\ell(f_w(X'), f_{w^*}(X'))\right] + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right]\\
\leq&\mathbb{E}_X\left[\ell(f_w(X), f_{w^*}(X))\right] + \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)} + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right] \qquad (8)\\
=&\int_x \ell(f_w(x), f_{w^*}(x))dP_X(x) + \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)} + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right]\\
=&\int_x\int_y \ell(f_w(x), f_{w^*}(x))dP_{Y|X=x}(y)dP_X(x) + \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)} + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right]\\
\leq&\int_x\int_y \ell(f_w(x), y) + \ell(y, f_{w^*}(x))dP_{Y|X=x}(y)dP_X(x) + \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)} + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right]\\
&\qquad (9)\\
=&\mathbb{E}_Z\left[\ell(f_w(X), Y)\right] + \mathbb{E}_Z\left[\ell(Y, f_{w^*}(X))\right] + \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)} + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right],
\end{aligned}$$

where Eq. (8) is by Eq. (7) and Eq. (9) is again by the triangle property of the loss function.

Thus, $\widetilde{\mathrm{Err}}(w) \leq \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(\mathrm{P}_{\mathrm{X}'}||\mathrm{P}_{\mathrm{X}})} + \lambda^*$, which completes the proof. $\square$

B.4 Proof of Theorem 4.3

Proof. By Lemma A.1,

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X) \geq & \sup_{t\in\mathbb{R}, W, W'\in\mathcal{W}^2} \mathbb{E}_{X'}\left[t\ell(f_W(X'), f_{W'}(X'))\right] - \log \mathbb{E}_X\left[e^{t\ell(f_W(X), f_{W'}(X))}\right] \\
& \geq \sup_{t\in\mathbb{R}} \mathbb{E}_{W,W'}\left[\mathbb{E}_{X'}\left[t\ell(f_W(X'), f_{W'}(X'))\right] - \log \mathbb{E}_X\left[e^{t\ell(f_W(X), f_{W'}(X))}\right]\right] \\
& \geq \sup_{t\in\mathbb{R}} \mathbb{E}_{W,W',X'}\left[t\ell(f_W(X'), f_{W'}(X'))\right] - \log \mathbb{E}_{W,W',X}\left[e^{t\ell(f_W(X), f_{W'}(X))}\right],
\end{aligned}
$$

where the last inequality is by applying Jensen's inequality to the concavity of logarithm function.

By Lemma A.2,

$$
\left|\mathbb{E}_{W,W',X'}\left[\ell(f_W(X'), f_{W'}(X'))\right] - \mathbb{E}_{W,W',X}\left[\ell(f_W(X), f_{W'}(X))\right]\right| \leq \sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)}.
$$

This concludes the proof. ☐

B.5 Proof of Theorem 4.4

Proof. From the definition, we have

$$
\begin{aligned}
\left|\widetilde{\mathrm{Err}}(w)\right| &= \left|\mathbb{E}_{Z'}\left[\ell(f_w(X'), Y')\right] - \mathbb{E}_Z\left[\ell(f_w(X), Y)\right]\right| \\
&\leq \beta\mathbb{W}(\mu, \mu').
\end{aligned}
$$

where the last inequality is by the KR duality of Wasserstein distance (see Lemma A.3). ☐

B.6 Proof of Corollary 4.3

Proof. When d is the discrete metric, Wasserstein distance is equal to the total variation, then by Theorem 4.4,

$$
\left|\widetilde{\mathrm{Err}}(w)\right| \leq \beta\mathrm{TV}(\mu', \mu),
$$

The remaining part is by using Lemma A.4 and Lemma A.5:

$$
\beta\mathrm{TV}(\mu', \mu) \leq \beta\sqrt{\min\left\{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mu'||\mu), 1 - e^{-\mathrm{D}_{\mathrm{KL}}(\mu'||\mu)}\right\}}.
$$

Then, if ℓ is bounded by M, we replace β by M above, which completes the proof. ☐

B.7 Proof of Theorem 4.5

Proof. Let $w^* = \arg\min_{w\in\mathcal{W}} \mathbb{E}_{Z'}\left[\ell(f_w(X'), Y')\right] + \mathbb{E}_Z\left[\ell(f_w(X), Y)\right]$.

If $\ell(f_w(X), f_{w'}(X))$ is L-Lipschitz in $\mathcal{X}$ for any $w, w' \in \mathcal{W}$, then similar to Theorem 4.4, it's easy to show that

$$
\mathbb{E}_{X'}\left[\ell(f_w(X'), f^*(X'))\right] - \mathbb{E}_X\left[\ell(f_w(X), f^*(X))\right] \leq \beta\mathbb{W}(P'_X, P_X) \tag{10}
$$

For any $f_w \in \mathcal{F}$, by the triangle property of the loss, we have

$$
\begin{aligned}
& \mathbb{E}_{Z'}\left[\ell(f_w(X'), Y')\right] \\
\leq & \mathbb{E}_{X'}\left[\ell(f_w(X'), f_{w^*}(X'))\right] + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right] \\
\leq & \mathbb{E}_X\left[\ell(f_w(X), f_{w^*}(X))\right] + \beta\mathbb{W}(P'_X, P_X) + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right] \qquad (11) \\
\leq & \mathbb{E}_Z\left[\ell(f_w(X), Y)\right] + \mathbb{E}_Z\left[\ell(Y, f_{w^*}(X))\right] + \beta\mathbb{W}(P'_X, P_X) + \mathbb{E}_{Z'}\left[\ell(f_{w^*}(X'), Y')\right],
\end{aligned}
$$

where Eq. (11) is by Eq. (10) and the last inequality is again by the triangle property of the loss function. This completes the proof. ☐

B.8 Additional Results: Sample Complexity Bounds

Theorem B.1. *Let $\hat{\mu}$ and $\hat{\mu}'$ be the empirical distributions consist of n source data and m target data sampled i.i.d. from μ and μ', respectively. Let $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R} \text{ s.t. } \mathbb{E}_{\mu}\left[e^{g(Z)}\right] < \infty\}$ with finite VC-dimension d_1, and let VC-dimension of $\{\exp \circ g | g \in \mathcal{G}\}$ be d_2. W.L.O.G. assume that $\alpha \leq \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right] \leq \mathbb{E}_{\mu}\left[e^{g(Z)}\right]$ for some constant $\alpha > 0$ and any $g \in \mathcal{G}$. Then for $\forall \delta \in (0,1)$ the following holds with probability at least $1-\delta$,*

$$\mathrm{D}_{\mathrm{KL}}(\mu' || \mu) - \mathrm{D}_{\mathrm{KL}}(\hat{\mu}' || \hat{\mu}) \leq \sqrt{\frac{4}{n}\left(d_1 \log \frac{2en}{d_1} + \log \frac{4}{\delta}\right)} + \frac{1}{\alpha}\sqrt{\frac{4}{m}\left(d_2 \log \frac{2em}{d_2} + \log \frac{4}{\delta}\right)}.$$

Proof. Recall Lemma A.1, we have

$$\mathrm{D}_{\mathrm{KL}}(\mu' || \mu) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'}\left[g(Z')\right] - \log \mathbb{E}_{\mu}\left[e^{g(Z)}\right],$$

and

$$\mathrm{D}_{\mathrm{KL}}(\hat{\mu}' || \hat{\mu}) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\hat{\mu}'}\left[g(Z')\right] - \log \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right].$$

Then, with the probability at least $1-\delta$,

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\mu' || \mu) - \mathrm{D}_{\mathrm{KL}}(\hat{\mu}' || \hat{\mu}) \\
=&\sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'}\left[g(Z')\right] - \log \mathbb{E}_{\mu'}\left[e^{g(Z)}\right] - \left(\sup_{g \in \mathcal{G}} \mathbb{E}_{\hat{\mu}'}\left[g(Z')\right] - \log \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]\right) \\
\leq&\sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'}\left[g(Z')\right] - \log \mathbb{E}_{\mu}\left[e^{g(Z)}\right] - \left(\mathbb{E}_{\hat{\mu}'}\left[g(Z')\right] - \log \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]\right) \\
=&\sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'}\left[g(Z')\right] - \mathbb{E}_{\hat{\mu}'}\left[g(Z')\right] + \log \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right] - \log \mathbb{E}_{\mu}\left[e^{g(Z)}\right] \\
\leq&\sup_{g \in \mathcal{G}} \left|\mathbb{E}_{\mu'}\left[g(Z')\right] - \mathbb{E}_{\hat{\mu}'}\left[g(Z')\right]\right| + \sup_{g \in \mathcal{G}} \left|\log \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right] - \log \mathbb{E}_{\mu}\left[e^{g(Z)}\right]\right| \\
\leq&\sup_{g \in \mathcal{G}} \left|\mathbb{E}_{\mu'}\left[g(Z')\right] - \mathbb{E}_{\hat{\mu}'}\left[g(Z')\right]\right| + \sup_{g \in \mathcal{G}} \frac{1}{\alpha}\left|\mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right] - \mathbb{E}_{\mu}\left[e^{g(Z)}\right]\right| \qquad (12) \\
\leq&\sqrt{\frac{4}{n}\left(d_1 \log \frac{2en}{d_1} + \log \frac{4}{\delta}\right)} + \frac{1}{\alpha}\sqrt{\frac{4}{m}\left(d_2 \log \frac{2em}{d_2} + \log \frac{4}{\delta}\right)}, \qquad (13)
\end{aligned}$$

where Eq. (12) is by

$$\begin{aligned}
\left|\log \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right] - \log \mathbb{E}_{\mu}\left[e^{g(Z)}\right]\right| = \left|\log \frac{\mathbb{E}_{\mu}\left[e^{g(Z)}\right]}{\mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]}\right| &= \left|\log\left(1 + \frac{\mathbb{E}_{\mu}\left[e^{g(Z)}\right]}{\mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]} - 1\right)\right| \\
&\leq \left|\frac{\mathbb{E}_{\mu}\left[e^{g(Z)}\right]}{\mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]} - 1\right| \\
&= \left|\frac{1}{\mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]}\left(\mathbb{E}_{\mu}\left[e^{g(Z)}\right] - \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]\right)\right| \\
&\leq \frac{1}{\alpha}\left|\mathbb{E}_{\mu}\left[e^{g(Z)}\right] - \mathbb{E}_{\hat{\mu}}\left[e^{g(Z)}\right]\right|,
\end{aligned}$$

and Eq. (13) is by the classic VC-dimension generalization bound [62]. This concludes the proof. ☐

With Theorem B.1 and Theorem 4.1, we immediately have the following corollary.

Corollary B.1. *Let the conditions in Theorem B.1 and Theorem 4.1 hold, then for any $w \in \mathcal{W}$,*

$$\left|\widetilde{\mathrm{Err}}(w)\right| \leq R\sqrt{2\mathrm{D}_{\mathrm{KL}}(\hat{\mu}' || \hat{\mu}) + 2\sqrt{\frac{4}{n}\left(d_1 \log \frac{2en}{d_1} + \log \frac{4}{\delta}\right)} + \frac{2}{\alpha}\sqrt{\frac{4}{m}\left(d_2 \log \frac{2em}{d_2} + \log \frac{4}{\delta}\right)}}.$$

B.9 Additional Discussions on the Convergence of Empirical KL Divergence

Although characterizing the convergence of the empirical KL divergence to the real KL is not easy without relying several additional assumptions (as in Theorem B.1), the result of convergence rate of empirical distribution to the real distribution in the KL sense is already known in the discrete space. The following theorem comes directly from the classic result in [63, Theorem 11.2.1],

Theorem B.2. *Let $\hat{\mu}$ and $\hat{\mu}'$ be defined as in Theorem B.1. Assume the space of $\mathcal{Z}$ is finite (i.e. $|\mathcal{Z}| \leq \infty$), then for $\forall \delta \in (0,1)$, with the probability at least $1-\delta$,*

$$\mathrm{D_{KL}}(\hat{\mu}||\mu) \leq \frac{|\mathcal{Z}|}{n}\log{(n+1)} + \frac{1}{n\log\delta}, \qquad \mathrm{D_{KL}}(\hat{\mu}'||\mu') \leq \frac{|\mathcal{Z}|}{m}\log{(m+1)} + \frac{1}{m\log\delta}.$$

Thus, it suffices to ensure that the empirical KL converge to the real KL with the similar rate, although we do not know if there might exist more optimal convergence rate.

C Omitted Proofs and Additional Discussions in Section 5

C.1 Additional Discussion on Theorem 5.1

To derive the bound in Theorem 5.1, we need to make use of the second equality in Eq. (1). Indeed, by the definition of Err (the first equality in Eq. (1)), the unlabelled sample $S'_{X'_j}$ does not explicitly appear, so one can easily apply the similar information-theoretic analysis starting from the first equality in Eq. (1), and obtain an upper bound that consists of $I(W;Z_i)$ and $\mathrm{D_{KL}}(\mu||\mu')$. Precisely, the following bound holds,

Theorem C.1. *Assume $\ell(f_w(X'),Y')$ is R-subgaussian for any $w \in \mathcal{W}$. Then*

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sqrt{2R^2 I(W;Z_i)} + \sqrt{2R^2 \mathrm{D_{KL}}(\mu||\mu')}.$$

The proof of Theorem C.1 is nearly the same to the proof of [38, Corollary 2] and [40, Corollary 1].

It's important to note that although

$$I(W;Z_i) \leq I(W;Z_i|X'_j) = \mathbb{E}_{X'_j}\left[I^{X'_j}(W;Z_i)\right],$$

the bound in Theorem 5.1 is incomparable to the bound based on $I(W;Z_i)$. This is mainly due to the fact that we use the disintegrated version of mutual information, $I^{X'_j}(W;Z_i)$, and the expectation over X'_j is outside of the square root, which is a convex function. Using $I^{X'_j}(W;Z_i)$ instead of $I(W;Z_i)$ allows us to figure out more details about the role of unlabelled target data in the algorithm. Additionally, one can also prove a bound based on $I(W;Z_i|X'_j)$ (e.g., simply applying Jensen's inequality to Theorem 5.1), which is close to an individual and UDA version of [41, Theorem 3].

In essence, the first term in Theorem 5.1 characterize the expected generalization gap on the source domain (i.e. $\mathbb{E}_{W,S}\left[R_\mu(W) - R_S(W)\right]$), then the bound suggests us that it's possible to invoke the unlabelled target data to further improve the performance on source domain, and the simplest case is the semi-supervised learning (when $\mu = \mu'$).

C.2 Proof of Theorem 5.1

Proof. By Lemma A.1,

$$\begin{aligned}
&\mathrm{D_{KL}}\left(P_{W,Z_i|X'_j=x'_j}||P_{W,Z'|X'_j=x'_j}\right)\\
=&\mathrm{D_{KL}}\left(P_{W,Z_i|X'_j=x'_j}||P_{W|X'_j=x'_j}P_{Z'}\right) && (14)\\
\geq&\sup_t \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[t\ell(f_W(X_i),Y_i)\right] - \log\mathbb{E}_{P_{W|X'_j=x'_j}P_{Z'}}\left[\exp t\ell(f_W(X'),Y')\right]\\
\geq&\sup_t \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[t\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{P_{W|X'_j=x'_j}}\left[tR_{\mu'}(W)\right] - \log\mathbb{E}_{P_{W|X'_j=x'_j}P_{Z'}}\left[e^{t(\ell(f_W(X'),Y') - \mathbb{E}_{Z'}\left[\ell(f_W(X'),Y')\right])}\right] && (15)\\
\geq&\sup_t \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[t\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{P_{W|X'_j=x'_j}}\left[tR_{\mu'}(W)\right] - R^2t^2/2,
\end{aligned}$$

where Eq. (14) is by the independence between algorithm output W and unseen target domain data Z', Eq. (15) is by Jensen's inequality for the exponential function and the last inequality is by the subgaussian assumption.

Thus,

$$\left|\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{P_{W|X'_j=x'_j}}\left[R_{\mu'}(W)\right]\right|\leq\sqrt{2R^2\mathrm{D}_{\mathrm{KL}}\left(P_{W,Z_i|X'_j=x'_j}||P_{W|X'_j=x'_j}P_{Z'}\right)}. \tag{16}$$

Exploiting the fact that

$$\begin{aligned}
|\mathrm{Err}|=&\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,Z_i}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W,Z'}\left[\ell(f_W(X'),Y')\right]\right|\\
=&\left|\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{X'_j}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W,Z'|X'_j=x'_j}\left[\ell(f_W(X'),Y')\right]\right]\right|\\
\leq&\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{X'_j}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W,Z'|X'_j=x'_j}\left[\ell(f_W(X'),Y')\right]\right|\\
\leq&\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu'}(W)\right]\right|,
\end{aligned}$$

where the last two inequalities are by the Jensen's inequality for the absolute function.

Since

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(P_{W,Z_i|X'_j=x'_j}||P_{W|X'_j=x'_j}P_{Z'}\right)=&\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[\log\frac{P_{W,Z_i|X'_j=x'_j}}{P_{W|X'_j=x'_j}P_{Z'}}\right]\\
=&\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[\log\frac{P_{W|Z_i,X'_j=x'_j}P_{Z_i}}{P_{W|X'_j=x'_j}P_{Z'}}\right]\\
=&\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[\log\frac{P_{W|Z_i,X'_j=x'_j}}{P_{W|X'_j=x'_j}}\right]+\mathbb{E}_{P_{Z_i}}\left[\log\frac{P_{Z_i}}{P_{Z'}}\right]\\
=&I(W;Z_i|X'_j=x'_j)+\mathrm{D}_{\mathrm{KL}}(\mu||\mu').
\end{aligned}$$

Recall Eq. (16), we have

$$\begin{aligned}
|\mathrm{Err}|\leq&\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu'}(W)\right]\right|\\
\leq&\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2\mathrm{D}_{\mathrm{KL}}\left(P_{W,Z_i|X'_j=x'_j}||P_{W|X'_j=x'_j}P_{Z'}\right)}\\
=&\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2(I(W;Z_i|X'_j=x'_j)+\mathrm{D}_{\mathrm{KL}}(\mu||\mu'))}\\
\leq&\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2I^{X'_j}(W;Z_i)}+\sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(\mu||\mu')}.
\end{aligned}$$

This completes the proof. □

C.3 Proof of Corollary 5.1

Proof. We now modify the proof in Theorem 5.1.

Recall that

$$|\mathrm{Err}| \leq \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu'}(W)\right]\right|.$$

We first decompose the right hand side,

$$\begin{aligned}
&\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu'}(W)\right]\right| \\
=&\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu}(W)\right] + \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu}(W)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu'}(W)\right]\right| \\
\leq&\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu}(W)\right]\right| + \left|\mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu}(W) - R_{\mu'}(W)\right]\right| \\
\leq&\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu}(W)\right]\right| + \frac{M}{\sqrt{2}}\sqrt{\min\{\mathrm{D_{KL}}(\mu||\mu'), \mathrm{D_{KL}}(\mu'||\mu)\}},
\end{aligned}$$

where the last inequality is by Corollary 4.2.

Then for the first term in RHS, notice that

$$\begin{aligned}
&\mathrm{D_{KL}}\left(P_{W,Z|X'_j=x'_j}||P_{W,Z_i|X'_j=x'_j}\right) \\
=&\mathrm{D_{KL}}\left(P_{W|X'_j=x'_j}P_Z||P_{W,Z_i|X'_j=x'_j}\right) \\
\geq&\sup_t \mathbb{E}_{P_{W|X'_j=x'_j}P_Z}\left[t\ell(f_W(X),Y)\right] - \log\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[\exp t\ell(f_W(X_i),Y_i)\right] \\
\geq&\sup_t \mathbb{E}_{P_{W|X'_j=x'_j}P_Z}\left[t\ell(f_W(X),Y)\right] - \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[t\ell(f_W(X_i),Y_i)\right] \\
&- \log\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[e^{t(\ell(f_W(X_i),Y_i)-\mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[\ell(f_W(X_i),Y_i)\right])}\right] \\
\geq&\sup_t \mathbb{E}_{P_{W|X'_j=x'_j}}\left[tR_\mu(W)\right] - \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}}\left[t\ell(f_W(X_i),Y_i)\right] - M^2t^2/8,
\end{aligned}$$

where the last inequality is due to the fact that ℓ is bounded by M and $\ell(f_W(X_i),Y_i)$ is $M/2$-subgaussian.

Thus,

$$\begin{aligned}
\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right] - \mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu}(W)\right]\right| \leq&\sqrt{\frac{M^2}{2}\mathrm{D_{KL}}\left(P_{W|X'_j=x'_j}P_Z||P_{W,Z_i|X'_j=x'_j}\right)} \\
=&\sqrt{\frac{M^2}{2}L\left(W,Z_i|X'_j=x'_j\right)}.
\end{aligned}$$

Plugging this inequality with the decomposition into the inequality at the beginning of the proof, we have

$$|\mathrm{Err}| \leq \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{\frac{M^2}{2}L^{X'_j}\left(W,Z_i\right)} + \frac{M}{\sqrt{2}}\sqrt{\min\{\mathrm{D_{KL}}(\mu||\mu'), \mathrm{D_{KL}}(\mu'||\mu)\}}.$$

Similar development also holds for $\mathrm{D_{KL}}\left(P_{W,Z_i|X'_j=x'_j}||P_{W|X'_j=x'_j}P_Z\right)$, thus

$$|\mathrm{Err}| \leq \frac{M}{\sqrt{2}nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{\min\left\{I^{X'_j}(W;Z_i), L^{X'_j}(W;Z_i)\right\}} + \frac{M}{\sqrt{2}}\sqrt{\min\left\{\mathrm{D_{KL}}(\mu||\mu'), \mathrm{D_{KL}}(\mu'||\mu)\right\}}.$$

This completes the proof. □

C.4 Proof of Theorem 5.2

Proof. Similar to the proof of Theorem 5.1, we exploit the fact that

$$
\begin{aligned}
&|\mathrm{Err}| \\
\leq & \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W|X'_j=x'_j}\left[R_\mu(W)\right]+\mathbb{E}_{W|X'_j=x'_j}\left[R_\mu(W)\right]-\mathbb{E}_{W|X'_j=x'_j}\left[R_{\mu'}(W)\right]\right| \\
\leq & \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\left|\mathbb{E}_{W,Z_i|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W|X'_j=x'_j}\left[R_\mu(W)\right]\right|+\beta\mathbb{W}(\mu,\mu') \\
\leq & \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j,Z_i}\left|\mathbb{E}_{W|Z_i=z_i,X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]-\mathbb{E}_{W|X'_j=x'_j}\left[\ell(f_W(X_i),Y_i)\right]\right|+\beta\mathbb{W}(\mu,\mu') \\
\leq & \frac{\beta'}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j,Z_i}\mathbb{W}(P_{W|X'_j,Z_i},P_{W|X'_j})+\beta\mathbb{W}(\mu,\mu'),
\end{aligned}
$$

which concludes the proof. □

C.5 Proof of Corollary 5.2

Proof. Similar to the proof of Corollary 4.3, replacing Wasserstein distance by the total variation and replacing β and β' by M, will give us the first inequality,

$$
\left|\widetilde{\mathrm{Err}}\right| \leq \frac{M}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j,Z_i}\left[\mathrm{TV}(P_{W|Z_i,X'_j},P_{W|X'_j})\right]+M\mathrm{TV}(\mu,\mu').
$$

The second inequality is by Lemma A.4,

$$
\left|\widetilde{\mathrm{Err}}\right| \leq \frac{M}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j,Z_i}\sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(P_{W|Z_i,X'_j}||P_{W|X'_j})}+\sqrt{\frac{M^2}{2}\mathrm{D}_{\mathrm{KL}}(\mu||\mu')}.
$$

Again, one can also apply Lemma A.5 here. This concludes the proof. □

C.6 Proof of Theorem 5.3

Proof. Recall Theorem 5.1 and by Jensen's inequality we have

$$
\begin{aligned}
|\mathrm{Err}| \leq & \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2 I^{X'_j}(W;Z_i)}+\sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(\mu||\mu')} \\
\leq & \sqrt{\frac{2R^2}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}I(W;Z_i|X'_j)}+\sqrt{2R^2\mathrm{D}_{\mathrm{KL}}(\mu||\mu')}.
\end{aligned}
$$

First notice that,

$$
\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}I(W;Z_i|X'_j) \leq \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}I(W;Z_i|S'_{X'}) = \frac{1}{n}\sum_{i=1}^{n}I(W;Z_i|S'_{X'}).
$$

Then, since $S \perp\!\!\!\perp S'_{X'}$ and $Z_i \perp\!\!\!\perp Z_{1:i-1}$ for any $i \in [n]$, by chain rule of the mutual information, we have

$$\begin{aligned}
I(W;S|S'_{X'}) = \sum_{i=1}^{n} I(W;Z_i|S'_{X'},Z_{1:i-1}) &= \sum_{i=1}^{n} I(W;Z_i|S'_{X'},Z_{1:i-1}) + I(Z_i;Z_{1:i-1}) \\
&= \sum_{i=1}^{n} I(W,Z_{1:i-1};Z_i|S'_{X'}) \\
&= \sum_{i=1}^{n} I(W;Z_i|S'_{X'}) + I(Z_i;Z_{1:i-1}|S'_{X'},W) \\
&\geq \sum_{i=1}^{n} I(W;Z_i|S'_{X'}).
\end{aligned}$$

Thus, the generalization error bound becomes,

$$|\mathrm{Err}| \leq \sqrt{\frac{2R^2}{n} I(W;S|S'_{X'})} + \sqrt{2R^2 \mathrm{D_{KL}}(\mu||\mu')}.$$

Recall the updating rule of W and notice that W_0 is independent of S and $S'_{X'}$, the following process is by using chain rule of mutual information and data processing inequality recurrently,

$$\begin{aligned}
I(W_T;S|S'_{X'}) =& I(W_{T-1} - \eta_T g(W_{T-1},Z_{B_T},X'_{B_T}) + N_T;S|S'_{X'}) \\
\leq& I(W_{T-1}, -\eta_T g(W_{T-1},Z_{B_T},X'_{B_T}) + N_T;S|S'_{X'}) \\
=& I(W_{T-1};S|S'_{X'}) + I(\eta_T g(W_{T-1},Z_{B_T},X'_{B_T}) + N_T;S|S'_{X'},W_{T-1}) \\
&\vdots \\
=& \sum_{t=1}^{T} I(\eta_t g(W_{t-1},Z_{B_t},X'_{B_t}) + N_t;S|S'_{X'},W_{t-1}).
\end{aligned}$$

For each $t \in [T]$, denote $g(W_{t-1},Z_{B_t},X'_{B_t})$ as G_t, then

$$\begin{aligned}
I(\eta_t g(W_{t-1},Z_{B_t},X'_{B_t}) + N_t;S|S'_{X'},W_{t-1}) =& \mathbb{E}_{S'_{X'},W_{t-1},S}\left[\mathrm{D_{KL}}(P_{G_t+\frac{N_t}{\eta_t}|S,S'_{X'},W_{t-1}}||P_{G_t+\frac{N_t}{\eta_t}|S'_{X'},W_{t-1}})\right] \\
\leq& \mathbb{E}_{S'_{X'},W_{t-1},S}\left[\mathrm{D_{KL}}(P_{G_t+\frac{N_t}{\eta_t}|S,S'_{X'},W_{t-1}}||P_{N_t/\eta_t})\right] \\
=& \frac{\eta_t^2}{2\sigma_t^2}\mathbb{E}_{S'_{X'},W_{t-1},S}\left[||G_t||^2\right],
\end{aligned}$$

where the inequality is by Lemma A.6 and the last equality is by the KL divergence between two Gaussian distributions.

Finally, putting everything together,

$$|\mathrm{Err}| \leq \sqrt{\frac{R^2}{n}\sum_{t=1}^{T}\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{S'_{X'},W_{t-1},S}\left[||g(W_{t-1},Z_{B_t},X'_{B_t})||^2\right]} + \sqrt{2R^2\mathrm{D_{KL}}(\mu||\mu')},$$

which concludes the proof. □

C.7 Derivation of Eq. (6)

Recall the expected cross-entropy loss, we have

$$\begin{aligned}
\mathbb{E}_{W,Z_i}\left[\ell(f_W(T_i),Y_i)\right] &= \mathbb{E}_{Z_i,W}\left[-\log Q_{Y_i|T_i,W}\right] \\
&= \mathbb{E}_{Z_i,W}\left[\log \frac{P_{Y_i|T_i,W}}{Q_{Y_i|T_i,W}P_{Y_i|T_i,W}}\right] \\
&= H(Y_i|T_i,W) + \mathbb{E}_{X_i,W}\left[\mathrm{D}_{\mathrm{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right] \\
&= \mathbb{E}_{Z_i,W}\left[\log \frac{P_{Y_i|T_i}P_{W|T_i}}{P_{Y_i|T_i,W}P_{Y_i|T_i}P_{W|T_i}}\right] + \mathbb{E}_{T_i,W}\left[\mathrm{D}_{\mathrm{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right] \\
&= \mathbb{E}_{Z_i,W}\left[\log \frac{P_{Y_i|T_i}P_{W|T_i}}{P_{Y_i,W|T_i}P_{Y_i|T_i}}\right] + \mathbb{E}_{T_i,W}\left[\mathrm{D}_{\mathrm{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right] \\
&= H(Y_i|T_i) - I(W;Y_i|T_i) + \mathbb{E}_{T_i,W}\left[\mathrm{D}_{\mathrm{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right]
\end{aligned}$$

C.8 Additional Discussion on LIMIT

As mentioned in Section 5, [51] proposed an approach called LIMIT, refers to limiting label information memorization in training, to control label information. Roughly speaking, to update the parameters of the classifier, they construct an auxiliary network to predict the gradient instead of using the real gradient, in which case the true label is not directly used for training the classifier. To provide accurate gradients, they also need to train the auxiliary network by using the true labels. We find that the training of LIMIT is unstable and hard to tune the hyperparameters under the UDA setting. Thus, we choose to use the pseudo label strategy proposed in Section 5 instead of pseudo gradient strategy.

D Experiment Details

The implementation in this paper is on PyTorch [64], and all the experiments are carried out on NVIDIA Tesla V100 GPUs (32 GB).

D.1 Objective Functions of Gradient Penalty and Controlling Label Information

For every iteration, the objective function after adding the gradient penalty becomes

$$\min_{W} \hat{L}(W, Z_{B_t}, X'_{B_t}) + \lambda_1 ||g(W, Z_{B_t}, X'_{B_t})||^2,$$

where $\hat{L}(W, Z_{B_t}, X'_{B_t})$ is some loss function for the source and target domain data in the current mini-batch and λ_1 is the trade-off coefficient. For example, if we combine ERM with gradient penalty then $\hat{L}(W, Z_{B_t}, X'_{B_t}) = \frac{1}{|B_t|}\sum_{k\in B_t}\ell(f_W(X_k), Y_k)$ and ℓ could be the cross-entropy loss. Moreover, if we combine KL guided marginal alignment algorithm [9] with gradient penalty then the objective function is

$$\min_{W,\theta} \frac{1}{|B_t|}\sum_{k\in B_t}\ell(f_W(T_k), Y_k) + \beta_1 \mathrm{D}_{\mathrm{KL}}(P_{T'}||P_T) + \beta_2 \mathrm{D}_{\mathrm{KL}}(P_T||P_{T'}) + \lambda_1 ||g(W, Z_{B_t}, X'_{B_t})||^2,$$

where θ is the parameters of the representation network and the gradient is

$$g(W, Z_{B_t}, X'_{B_t}) = \frac{1}{|B_t|}\sum_{k\in B_t}\nabla_{W,\theta}\ell(f_W(T_k), Y_k) + \beta_1 \nabla_\theta \mathrm{D}_{\mathrm{KL}}(P_{T'}||P_T) + \beta_2 \nabla_\theta \mathrm{D}_{\mathrm{KL}}(P_T||P_{T'}).$$

In [9], the representation distribution is modelled as an Gaussian distribution, i.e., $T \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2 \mathrm{I}_d|X)$ and $T' \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2 \mathrm{I}_d|X')$. Additionally, let the batch size be $b = |B_t|$, the empirical KL divergence is estimated by the mini-batch data, as given in [9],

$$
\begin{aligned}
&\beta_1 \mathrm{D}_{\mathrm{KL}}(P_{T'}||P_T) + \beta_2 \mathrm{D}_{\mathrm{KL}}(P_T||P_{T'}) \\
\approx& \beta_1 \frac{1}{b} \sum_{k\in B_t} [\log P_{T'_k} - \log P_{T_k}] + \beta_2 \frac{1}{b} \sum_{k\in B_t} [\log P_{T_k} - \log P_{T'_k}] \\
\approx& \beta_1 \frac{1}{b} \sum_{k\in B_t} \left[\log \frac{1}{b} \sum_{k\in B_t} P_{T'_k|X'_k} - \log \frac{1}{b} \sum_{k\in B_t} P_{T_k|X_k}\right] + \beta_2 \frac{1}{b} \sum_{k\in B_t} \left[\log \frac{1}{b} \sum_{k\in B_t} P_{T_k|X_k} - \log \frac{1}{b} \sum_{k\in B_t} P_{T'_k|X'_k}\right],
\end{aligned}
$$

where $P_{T_k|X_k} = \mathcal{N}(\mu_\theta, \sigma_\theta^2 \mathrm{I}_d|X_k)$ and $P_{T'_k|X'_k} = \mathcal{N}(\mu_\theta, \sigma_\theta^2 \mathrm{I}_d|X'_k)$. To be more precise, μ_θ and σ_θ are the outputs of the representation network. Since the forward pass requires the sampling of T and T', we need to use the reparameterization trick [65] for the backward pass.

When we train the model with controlling label information, the objective function becomes

$$\min_{W} \hat{L}(W, Z_{B_t}, X'_{B_t}) + \lambda_2 ||W - \widetilde{W}||^2,$$

where $\widetilde{W}$ is the auxiliary classifier and λ_2 is the trade-off hyperparameter.

Similarly, when we combine KL guided marginal alignment algorithm with controlling label information, then the objective function in every iteration is

$$\min_{W,\theta} \frac{1}{|B_t|} \sum_{k\in B_t} \ell(f_W(T_k), Y_k) + \beta_1 \mathrm{D}_{\mathrm{KL}}(P_{T'}||P_T) + \beta_2 \mathrm{D}_{\mathrm{KL}}(P_T||P_{T'}) + \lambda_2 ||W - \widetilde{W}||^2.$$

In addition, the training objective for the auxiliary classifier is

$$\min_{\widetilde{W}} \frac{1}{|B_t|} \sum_{k\in B_t} \ell(f_{\widetilde{W}}(T'_k), f_W(T'_k)) + \frac{1}{|B_t|} \sum_{k\in B_t} \ell(f_{\widetilde{W}}(T_k), f_W(T_k)). \tag{17}$$

In practice, removing the second term would not affect the performance. Note that we need to disenable the automatic differentiation of T, T' and W when executing the backward pass for the auxiliary classifier. The detailed algorithm of controlling label information is given in the next section.

D.2 Algorithm of Controlling Label Information and Additional Results of ERM-CL

Algorithm 1 Controlling Label Information

```
Require: Source domain labelled dataset $S$, Target domain unlabelled dataset $S'_{X'}$, Batch size $b$,
    Classification loss function $\ell_c$, Marginal alignment loss function $\ell_r$, Initial classifier parameter
    $\boldsymbol{w}_0 = \widetilde{\boldsymbol{w}}_0$, Initial representation network parameter $\boldsymbol{\theta}_0$, Learning rate $\eta$, Lagrange multiplier $\lambda_2$
    while $\boldsymbol{w}_t$ not converged do
        Update iteration: $t \leftarrow t + 1$
        Sample $\mathcal{Z}_{\mathcal{B}} = \{\boldsymbol{z}_i\}_{i=1}^b$ from source domain training set $S$
        Sample $\mathcal{X}'_{\mathcal{B}} = \{\boldsymbol{x}'_i\}_{i=1}^b$ from target domain training set $S'_{X'}$
        Compute distance from the auxiliary classifier $\boldsymbol{dis} \leftarrow ||\boldsymbol{w}_t - \widetilde{\boldsymbol{w}}_t||^2$
        Compute marginal alignment loss $L_r \leftarrow \frac{1}{b}\sum_{i=1}^b \ell_r(\theta_t, \boldsymbol{z}_i, \boldsymbol{x}'_i)$
        Compute classification loss $L_c \leftarrow \frac{1}{b}\sum_{i=1}^b \ell_c(\boldsymbol{w}_t, \theta_t, \boldsymbol{z}_i, \boldsymbol{x}'_i)$
        Compute gradient:
        $g_{\mathcal{B}} \leftarrow \nabla(L_c + L_r + \lambda_2 \boldsymbol{dis})$
        Update parameter: $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \cdot g_{\mathcal{B}}$, $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \cdot g_{\mathcal{B}}$
        Obtain the pseudo labels $\mathcal{Y}'_{\mathcal{B}} \leftarrow f_{\boldsymbol{w}_t}(g_{\boldsymbol{\theta}_t}(\mathcal{X}'_{\mathcal{B}}))$
        Compute auxiliary classification loss $L_a \leftarrow \frac{1}{b}\sum_{i=1}^b \ell_c(\widetilde{\boldsymbol{w}}_t, \theta_t, \boldsymbol{x}'_i, \boldsymbol{y}'_i)$
        Compute auxiliary classifier gradient:
        $\widetilde{g}_{\mathcal{B}} \leftarrow \nabla L_a$
        Update auxiliary classifier parameter: $\widetilde{\boldsymbol{w}}_{t+1} \leftarrow \widetilde{\boldsymbol{w}}_t - \eta \cdot \widetilde{g}_{\mathcal{B}}$
    end while
```

Table 2: RotatedMNIST and Digits Experiments of **ERM-CL**. Results of ERM are reported from [9].

	RotatedMNIST ($\mathbf{0}^\circ$ as source domain)						Digits			
Method	$\mathbf{15}^\circ$	$\mathbf{30}^\circ$	$\mathbf{45}^\circ$	$\mathbf{60}^\circ$	$\mathbf{75}^\circ$	**Ave**	**M → U**	**U → M**	**S → M**	**Ave**
ERM	97.5±0.2	84.1±0.8	53.9±0.7	34.2±0.4	22.3±0.5	58.4	73.1±4.2	54.8±6.2	65.9±1.4	64.6
ERM-GP	**97.5±0.1**	**86.2±0.5**	**62.0±1.9**	**34.8±2.1**	**26.1±1.2**	**61.2**	**91.3±1.6**	**72.7±4.2**	68.4±0.2	77.5
ERM-CL	97.3±0.1	84.1±0.1	56.9±2.5	34.2±1.9	25.5±1.6	59.6	88.9±0.4	71.2±3.6	**73.5±1.4**	**77.9**

If we only provide the pseudo labels for the target domain data to the auxiliary classifier, i.e. removing the second term in Eq (17), the Algorithm 1 is the algorithm for combining any marginal alignment algorithm with controlling label information.

Even without incorporating with the marginal alignment algorithm, e.g., ERM, in which case L_r is removed, Algorithm 1 still boosts the performance in practice.

Table 2 shows that **ERM-CL** can overall outperform the basic **ERM** and is close to the performance of **ERM-GP**.

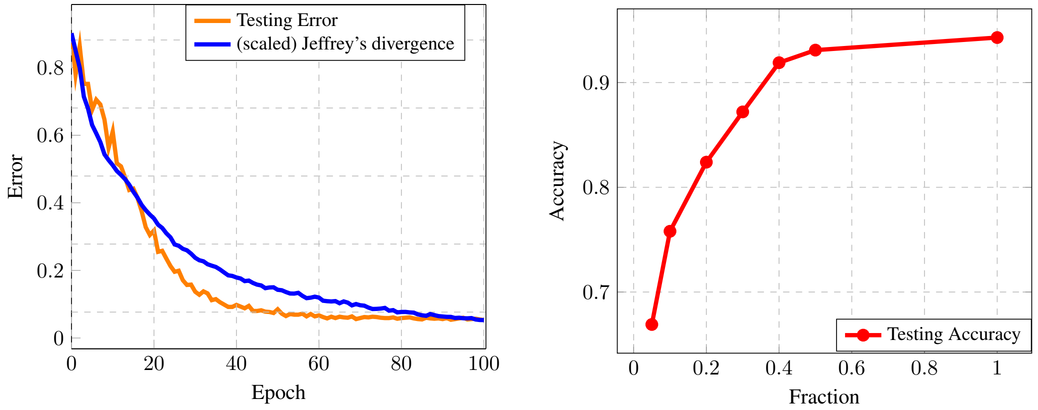
D.3 Architectures and Hyperparameters

The network architecture in this work is the same as in [56] and [9], where a simple CNN is used.

Other settings are also the same as [56] and [9], for example, each algorithm is trained for 100 epochs. To select the hyperparameters (λ_1 and λ_2) for **ERM-GP**, **ERM-KL**, **KL-GP** and **KL-CL**, we perform random search. Specifically, λ_1 is searched between $[0.1, 0.9]$ and λ_2 is searched between $[10^{-6}, 0.8]$. Other hyperparameters searching range could be found in the source code.

D.4 Additional Experimental Results

The representation version of Corollary 4.2 hints that small Jeffrey's divergence will make the testing error small. In Figure 2a, we show that the dynamic of Jeffrey's divergence (computed in representation space) can well characterize the evolution of the testing error during the training phase. In Figure 2b, we show that the number of target data has some impact on the testing performance on the target data. When we use less than half of the available unlabelled target data, the performance increases as the number of data increases. When we use more than half of the unlabelled target data, the improvement on the performance is very small.



(a) Training Dynamic

(b) Testing Performance vs. # of Target Data

Figure 2: **KL** on **S**→**M**. The left figure is the comparison of the Jeffrey's divergence in the representation space and the testing error. The right figure is the evolution of testing accuracy with respect to the different fraction of unlabelled target data used for training.

D.5 License of the Assets

MNIST is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license. SVHN is licensed under the GNU General Public License v3.0. The source code from the DomainBed suite is released under the MIT license.

E Limitations

A central notion in our bounds is KL divergence (which includes mutual information as a special case). Although generic and universally applicable, KL divergence has a fundamental limitation in capturing the natural metric in the underlying space, which may cause the bounds incapable of extracting certain structural properties in some settings.

In the mutual information-based bounds, the key random variable is weight W. For over-parametrized models, this variable may not be sufficiently indicative as the algorithm's output. Replacing W by a random variable on the space $\mathcal{F}$ of classifiers may lead to tighter bounds.