Variational Representation of f-divergence and Its Applications

DA Theorey and CMI Generalization Bounds

Ziqiao Wang

School of Computer Science and Technology
Tongji University

October 30, 2024

From Entropy to Mutual Information

- Entropy: $H(X) = \mathbb{E}_{P_X}\left[\log \frac{1}{P(X)}\right]$, $H(X,Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{1}{P(X,Y)}\right]$, $H(X|Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{1}{P(X|Y)}\right]$
  - For discrete X, $H(X) \geq 0$
  - $H(X,Y) = H(X|Y) + H(Y)$
  - Conditioning reduces entropy: $H(X|Y) \leq H(X)$
  - For discrete X, $H(X) \leq \log |\mathcal{X}|$
- Relative entropy or KL divergence: $\mathrm{D_{KL}}(Q||P) = \mathbb{E}_Q\left[\log \frac{Q(X)}{P(X)}\right]$
  - $\mathrm{D_{KL}}(Q||P) \geq 0$ with equality holds iff $Q = P$.
  - Usually $\mathrm{D_{KL}}(Q||P) \neq \mathrm{D_{KL}}(P||Q)$

From Entropy to Mutual Information

- Mutual Information:
  $I(X;Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right] = \mathrm{D}_{\mathrm{KL}}\left(P_{X,Y}||P_X P_Y\right).$
  - $I(X;Y) \geq 0$ with equality holds iff $X \perp\!\!\!\perp Y$.
  - $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y).$
  - $I(X;Y) = I(Y;X)$
- Chain-rule:
  - $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$
  - $I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1)$
  - $\mathrm{D}_{\mathrm{KL}}\left(Q_{X,Y}||P_{X,Y}\right) = \mathrm{D}_{\mathrm{KL}}\left(Q_X||P_X\right) + \mathrm{D}_{\mathrm{KL}}\left(Q_{Y|X}||P_{Y|X}\right)$
- Data-processing inequality (DPI):
  If $X - Y - Z$ forms a Markov chain (i.e. $P_{X,Z|Y} = P_{X|Y}P_{Z|Y}$), then

$$I(X;Y) \geq I(X;Z)$$

From Entropy to Mutual Information

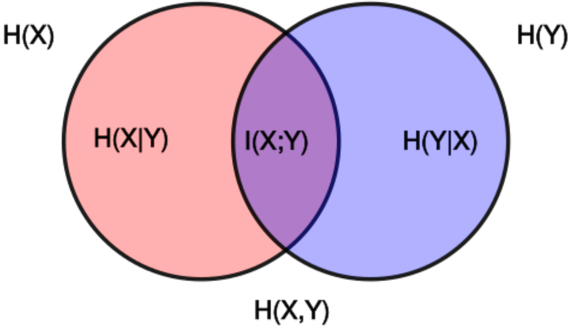


Figure 1: Venn diagram. Credit: https://en.wikipedia.org/wiki/Mutual_information

The family of f-Divergence

Definition 1 (f-divergence between two distributions)

Let P and Q be two distributions on Θ. Let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $\phi(1) = 0$. If $P \ll Q$[1]*, then f-divergence is defined as* $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, *where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative.*

[1] We say that P is absolutely continuous with respect to Q, written $P \ll Q$, if $Q(A) = 0 \implies P(A) = 0$ for all measurable sets $A \subseteq \Theta$.

The family of f-Divergence

Definition 1 (f-divergence between two distributions)

Let P and Q be two distributions on Θ. Let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $\phi(1) = 0$. If $P \ll Q$[1], then f-divergence is defined as $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative.

- ▷ Let $\phi(x) = x\log x$ (or $x\log x + c(x-1)$ for any constant c):

$$
\begin{aligned}
\mathrm{D}_\phi(P||Q) = \int \frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right)dQ = \int \log\left(\frac{dP}{dQ}\right)dP \\
=\mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] \\
=\mathrm{D}_{\mathrm{KL}}\left(P||Q\right).
\end{aligned}
$$

- ▷ Properties: Non-negativity; Data-processing inequality; Jointly Convexity

[1] We say that P is absolutely continuous with respect to Q, written $P \ll Q$, if $Q(A) = 0 \Longrightarrow P(A) = 0$ for all measurable sets $A \subseteq \Theta$.

More Examples

Divergence	Corresponding *f(t)*
χ^α-divergence, $\alpha \geq 1$	$\frac{1}{2}\lvert t-1\rvert^\alpha$
Total variation distance ($\alpha = 1$)	$\frac{1}{2}\lvert t-1\rvert$
α-divergence	$\begin{cases} \frac{t^\alpha-\alpha t-(1-\alpha)}{\alpha(\alpha-1)} & \text{if } \alpha \neq 0,\, \alpha \neq 1, \\ t\ln t - t + 1, & \text{if } \alpha = 1, \\ -\ln t + t - 1, & \text{if } \alpha = 0 \end{cases}$
KL-divergence ($\alpha = 1$)	$t\ln t$
reverse KL-divergence ($\alpha = 0$)	$-\ln t$
Jensen–Shannon divergence	$\frac{1}{2}\left(t\ln t - (t+1)\ln\left(\frac{t+1}{2}\right)\right)$
Jeffreys divergence (KL + reverse KL)	$(t-1)\ln(t)$
squared Hellinger distance ($\alpha = \frac{1}{2}$)	$\frac{1}{2}(\sqrt{t}-1)^2,\ 1-\sqrt{t}$
Pearson χ^2-divergence (rescaling of $\alpha = 2$)	$(t-1)^2,\ t^2-1,\ t^2-t$
Neyman χ^2-divergence (reverse Pearson) (rescaling of $\alpha = -1$)	$\frac{1}{t}-1,\ \frac{1}{t}-t$

Figure 2: Common examples of f-divergences. Credit: https://en.wikipedia.org/wiki/F-divergence

Legendre Transformation of f-divergence

Definition 2 (Convex Conjugate)

For a function $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty, +\infty\}$, its convex conjugate is

$$f^*(y) \triangleq \sup_{x \in \mathrm{dom}(f)} \langle x, y \rangle - f(x).$$

$$\begin{aligned} \mathrm{D}_\phi(P||Q) = \int \phi\left(\frac{dP}{dQ}\right) dQ &= \int \sup_g g\frac{dP}{dQ} - \phi^*(g) dQ \\ &\geq \sup_g \int g\frac{dP}{dQ} - \phi^*(g) dQ \\ &= \sup_g \mathbb{E}_P[g] - \mathbb{E}_Q[\phi^*(g)] \end{aligned}$$

▷ Variational Representation of f-divergence

$$\mathrm{D}_\phi(P||Q) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\theta \sim P}[g(\theta)] - \mathbb{E}_{\theta \sim Q}[\phi^*(g(\theta))]. \tag{1}$$

Tighter Variational Formula

- Applying "Shift Transformation" to the measurable function g:
  - Original variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\theta \sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta \sim Q}\left[\phi^*(g(\theta))\right]. \tag{2}$$

  - Reparameterization of $g \to g + \alpha$ (i.e. shifts):

$$\begin{aligned} \mathrm{D}_\phi(P||Q) &= \sup_g \sup_\alpha \mathbb{E}_{\theta \sim P}\left[g(\theta) + \alpha\right] - \mathbb{E}_{\theta \sim Q}\left[\phi^*(g(\theta) + \alpha)\right] \\ &= \sup_g \mathbb{E}_{\theta \sim P}\left[g(\theta)\right] - \inf_{\alpha \in \mathbb{R}} \left\{\mathbb{E}_{\theta \sim Q}\left[\phi^*(g(\theta) + \alpha) - \alpha\right]\right\}. \end{aligned} \tag{3}$$

Tighter Variational Formula

- Applying "Shift Transformation" to the measurable function g:
  - Original variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right]. \tag{2}$$

  - Reparameterization of $g \to g + \alpha$ (i.e. shifts):

$$\begin{aligned}\mathrm{D}_\phi(P||Q) &= \sup_{g} \sup_{\alpha} \mathbb{E}_{\theta\sim P}\left[g(\theta)+\alpha\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta)+\alpha)\right] \\ &= \sup_{g} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \inf_{\alpha\in\mathbb{R}} \left\{\mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta)+\alpha) - \alpha\right]\right\}.\end{aligned} \tag{3}$$

Eq. (3) is point-wise tighter than Eq. (2)

Example: Donsker and Varadhans (DV) representation of KL divergence

Consider KL divergence, let $\phi(x) = x \log x - x + 1$, then $\phi^*(y) = e^y - 1$. Substituting ϕ^* into Eq. (2)

$$\mathrm{D_{KL}}(P||Q) = \sup_{g \in \mathcal{G}} \mathbb{E}_P[g(\theta)] - \mathbb{E}_Q\left[e^{g(\theta)} - 1\right]. \tag{4}$$

On the other hand, Eq. (3) will give us

$$\begin{aligned}\mathrm{D_{KL}}(P||Q) &= \sup_{g \in \mathcal{G}} \mathbb{E}_P[g(\theta)] - \inf_{\alpha \in \mathbb{R}} \left\{\mathbb{E}_Q\left[e^{g(\theta)+\alpha}\right] - 1 - \alpha\right\} \\ &= \sup_{g \in \mathcal{G}} \mathbb{E}_P[g(\theta)] - \log \mathbb{E}_Q\left[e^{g(\theta)}\right], \end{aligned} \tag{5}$$

where the optimal $\alpha^* = -\log \mathbb{E}_Q\left[e^{g(\theta)}\right]$.

- Eq. (5) recovers the DV representation of KL.
- As $\log(x) \leq x - 1$ for $x > 0$, as a lower bound of KL divergence, Eq. (5) is pointwise tighter than Eq. (4).

Further Improvement: Affine Transformation

- Reparameterization of $g \to tg + \alpha$ (i.e. affine transformation):

$$\mathrm{D}_{\phi}(P||Q) = \sup_{g} \sup_{t,\alpha} \mathbb{E}_{\theta\sim P}\left[tg(\theta)+\alpha\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^{*}(tg(\theta)+\alpha)\right]. \quad (6)$$

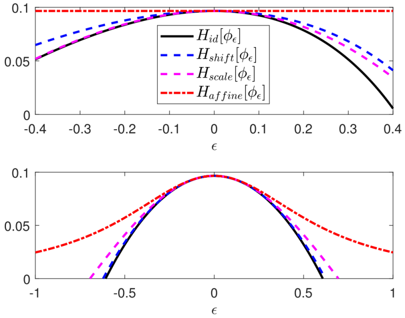


Figure 3: Visualization of Estimating f-divergences. Credit: Birrell et al. [2022]

Another Example: χ^2-divergence

For χ^2-divergence, let $\phi(x) = (x-1)^2$ for $x > 0$, then $\phi^*(y) = \frac{y^2}{4} + y$.
Plugging ϕ^* into Eq. (2):

$$\chi^2(P||Q) = \sup_g \mathbb{E}_P\left[g(\theta)\right] - \mathbb{E}_Q\left[g(\theta)\right] - \frac{\mathbb{E}_Q\left[(g(\theta))^2\right]}{4}. \tag{7}$$

Another Example: χ^2-divergence

For χ^2-divergence, let $\phi(x) = (x-1)^2$ for $x > 0$, then $\phi^*(y) = \frac{y^2}{4} + y$.
Plugging ϕ^* into Eq. (2):

$$\chi^2(P||Q) = \sup_g \mathbb{E}_P\left[g(\theta)\right] - \mathbb{E}_Q\left[g(\theta)\right] - \frac{\mathbb{E}_Q\left[(g(\theta))^2\right]}{4}. \tag{7}$$

Similarly, plugging ϕ^* into Eq. (3):

$$\chi^2(P||Q) = \sup_g \mathbb{E}_P\left[g(\theta)\right] - \mathbb{E}_Q\left[g(\theta)\right] - \frac{\mathrm{Var}_Q\left(g(\theta)\right)}{4}, \tag{8}$$

where the optimal $\alpha^* = \mathbb{E}_Q\left[g(\theta)\right]$.
By $\mathrm{Var}_Q\left(g(\theta)\right) \leq \mathbb{E}_Q\left[(g(\theta))^2\right]$, Eq. (8) is tighter than Eq. (7).

Another Example: χ^2-divergence

For χ^2-divergence, let $\phi(x) = (x-1)^2$ for $x > 0$, then $\phi^*(y) = \frac{y^2}{4} + y$.
Plugging ϕ^* into Eq. (2):

$$\chi^2(P||Q) = \sup_g \mathbb{E}_P\left[g(\theta)\right] - \mathbb{E}_Q\left[g(\theta)\right] - \frac{\mathbb{E}_Q\left[(g(\theta))^2\right]}{4}. \tag{7}$$

Similarly, plugging ϕ^* into Eq. (3):

$$\chi^2(P||Q) = \sup_g \mathbb{E}_P\left[g(\theta)\right] - \mathbb{E}_Q\left[g(\theta)\right] - \frac{\mathrm{Var}_Q\left(g(\theta)\right)}{4}, \tag{8}$$

where the optimal $\alpha^* = \mathbb{E}_Q\left[g(\theta)\right]$.
By $\mathrm{Var}_Q\left(g(\theta)\right) \leq \mathbb{E}_Q\left[(g(\theta))^2\right]$, Eq. (8) is tighter than Eq. (7).
Using Eq. (6):

$$\chi^2(P||Q) = \sup_g \frac{\left(\mathbb{E}_P\left[g(\theta)\right] - \mathbb{E}_Q\left[g(\theta)\right]\right)^2}{\mathrm{Var}_Q\left(g(\theta)\right)}, \tag{9}$$

where the optimal $t^* = \frac{2(\mathbb{E}_P[g(\theta)] - \mathbb{E}_Q[g(\theta)])}{\mathrm{Var}_Q(g(\theta))}$ and $\alpha^* = -t^*\mathbb{E}_Q\left[g(\theta)\right]$.
Eq. (9) recovers Hammersley-Chapman-Robbins lower bound.

Domain Adaptation

Problem Setup

- Given data from a source domain, i.e. $\{X_i, Y_i\} \overset{i.i.d.}{\sim} \mu$
- Obtain a model for a target domain, i.e. $\{X, Y\} \sim \nu$
- **Practical Goal**: Efficiently transfer ML models between related populations at low cost.

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;
- **Unsupervised Domain Adaptation (UDA)**:
  - Unknown distributions μ and ν
  - Labeled source-domain sample $\mathcal{S} = \{X_i, Y_i\}_{i=1}^{n} \sim \mu^{\otimes n}$
  - Unlabelled target-domain sample $\mathcal{T} = \{X_j\}_{j=1}^{m} \sim \nu_{\mathcal{X}}^{\otimes m}$
  - **Target**: find a hypothesis $h \in \mathcal{H}$ "works well" on ν.

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;
- **Unsupervised Domain Adaptation (UDA)**:
  - Unknown distributions μ and ν
  - Labeled source-domain sample $S = \{X_i, Y_i\}_{i=1}^n \sim \mu^{\otimes n}$
  - Unlabelled target-domain sample $T = \{X_j\}_{j=1}^m \sim \nu_{\mathcal{X}}^{\otimes m}$
  - **Target**: find a hypothesis $h \in \mathcal{H}$ "works well" on ν.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$.
- Target error: $R_\nu(h) \triangleq \mathbb{E}_{(X,Y)\sim\nu}\left[\ell(h(X), Y)\right]$,
  Source error: $R_\mu(h) \triangleq \mathbb{E}_{(X,Y)\sim\mu}\left[\ell(h(X), Y)\right]$.

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;
- **Unsupervised Domain Adaptation (UDA)**:
  - Unknown distributions μ and ν
  - Labeled source-domain sample $S = \{X_i, Y_i\}_{i=1}^{n} \sim \mu^{\otimes n}$
  - Unlabelled target-domain sample $T = \{X_j\}_{j=1}^{m} \sim \nu_{\mathcal{X}}^{\otimes m}$
  - **Target**: find a hypothesis $h \in \mathcal{H}$ "works well" on ν.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$.
- Target error: $R_\nu(h) \triangleq \mathbb{E}_{(X,Y)\sim\nu}\left[\ell(h(X), Y)\right]$,
  Source error: $R_\mu(h) \triangleq \mathbb{E}_{(X,Y)\sim\mu}\left[\ell(h(X), Y)\right]$.
- We use $\ell(h, h')$ to denote $\ell(h(x), h'(x))$,
  i.e. the disagreement of h and h' on x.

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_{\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\nu}\left[\ell(h,h')\right]\right|.$$

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_{\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\nu}\left[\ell(h,h')\right]\right|.$$

- Assumptions:
  - Triangle property: $\ell(y_1,y_2) \le \ell(y_1,y_3) + \ell(y_3,y_2)$ for any $y_1, y_2, y_3 \in \mathcal{Y}$.
  - Bounded loss: e.g., $\ell \in [0,1]$

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\ell(h,h')\right]\right|.$$

- Assumptions:
  - Triangle property: $\ell(y_1,y_2) \leq \ell(y_1,y_3) + \ell(y_3,y_2)$ for any $y_1,y_2,y_3 \in \mathcal{Y}$.
  - Bounded loss: e.g., $\ell \in [0,1]$

Theorem 1 ($\mathcal{H}\Delta\mathcal{H}$-divergence Bound)

Then, for any $h \in \mathcal{H}$,

$$R_\nu(h) \leq R_\mu(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) + \lambda^*,$$

where $\lambda^* = \min_{h^*\in\mathcal{H}} R_\nu(h^*) + R_\mu(h^*)$.

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\ell(h,h')\right]\right|.$$

- Assumptions:
  - Triangle property: $\ell(y_1,y_2) \le \ell(y_1,y_3) + \ell(y_3,y_2)$ for any $y_1,y_2,y_3 \in \mathcal{Y}$.
  - Bounded loss: e.g., $\ell \in [0,1]$

Theorem 1 ($\mathcal{H}\Delta\mathcal{H}$-divergence Bound)

Then, for any $h \in \mathcal{H}$,

$$R_\nu(h) \le R_\mu(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) + \lambda^*,$$

where $\lambda^* = \min_{h^*\in\mathcal{H}} R_\nu(h^*) + R_\mu(h^*)$.

Can we extend $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence?

From $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence

- f-divergence: $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where ϕ is convex and $\phi(1) = 0$.

From $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence

- f-divergence: $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where ϕ is convex and $\phi(1)=0$.
- Its variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right].$$

From $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence

- f-divergence: $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where ϕ is convex and $\phi(1)=0$.
- Its variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right].$$

- Acuna et al. [2021] defines:

$$\widetilde{\mathrm{D}}_\phi^{h,\mathcal{H}}(\mu||\nu) \triangleq \sup_{h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\phi^*(\ell(h,h'))\right]\right|.$$

$\Longrightarrow$ Additional absolute value function added.

Gap between Theory and Algorithm in Acuna et al. [2021]

$$\widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu) \triangleq \sup_{h'\in\mathcal{H}} \left|\mathbb{E}_{\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\nu}\left[\phi^*(\ell(h,h'))\right]\right|.$$

▷ Theory (Target Error Bound):

$$R_{\nu}(h) \leq R_{\mu}(h) + \widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu) + \lambda^*,$$

$\Longrightarrow$ Absolute value function is necessary for establishing this bound

Gap between Theory and Algorithm in Acuna et al. [2021]

$$\widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu) \triangleq \sup_{h'\in\mathcal{H}} |\mathbb{E}_{\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\nu}\left[\phi^*(\ell(h,h'))\right]|.$$

- Theory (Target Error Bound):

$$R_{\nu}(h) \leq R_{\mu}(h) + \widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu) + \lambda^*,$$

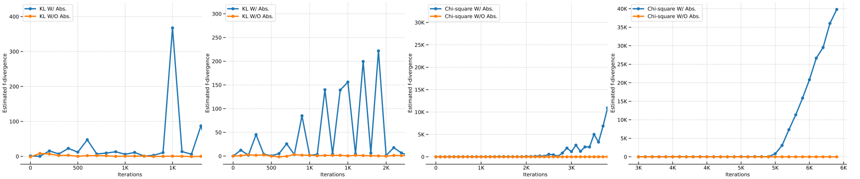
$\Longrightarrow$ Absolute value function is necessary for establishing this bound

- f-Domain Adversarial Learning (f-DAL) Algorithm:

$$\min_{h} R_{\hat{\mu}}(h) + \underbrace{\max_{h'} \mathbb{E}_{\hat{\mu}}\left[\ell(h,h')\right] - \mathbb{E}_{\hat{\nu}}\left[\phi^*(\ell(h,h'))\right]}_{d(\hat{\mu},\hat{\nu};h)}.$$

$\Longrightarrow$ $d(\hat{\mu},\hat{\nu};h)$ **drops the absolute value function** compared with $\widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu)$

Overestimation by Absolute Value Function



(a) KL (Office-31) (b) KL (Office-Home) (c) χ^2 (Office-31) (d) χ^2 (Office-Home)

Figure 4: The y-axis is the estimated corresponding f-divergence and the x-axis is the number of iterations.

- f-DAL algorithm fails if the absolute value function is added.

Our work: New f-Domain Discrepancy (f-DD)

Ziqiao Wang and Yongyi Mao. "On f-Divergence Principled Domain Adaptation: An Improved Framework." To appear at NeurIPS 2024.

▷ Our f-DD:

$$\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) \triangleq \sup_{t\in\mathbb{R},h'} \mathbb{E}_{\nu}\left[t\ell(h,h')\right] - \inf_{\alpha\in\mathbb{R}} \mathbb{E}_{\mu}\left[\phi^*(t\ell(h,h')+\alpha)-\alpha\right].$$

Our work: New f-Domain Discrepancy (f-DD)

Ziqiao Wang and Yongyi Mao. "On f-Divergence Principled Domain Adaptation: An Improved Framework." To appear at NeurIPS 2024.

- Our f-DD:

$$\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) \triangleq \sup_{t\in\mathbb{R},h'} \mathbb{E}_{\nu}\left[t\ell(h,h')\right] - \inf_{\alpha\in\mathbb{R}} \mathbb{E}_{\mu}\left[\phi^*(t\ell(h,h')+\alpha)-\alpha\right].$$

- Target Error Bound: For any $h \in \mathcal{H}$,

$$R_{\nu}(h) \leq R_{\mu}(h) + \inf_{t\geq 0} \frac{\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) + K_{\mu}(t)}{t} + \lambda^*, \tag{10}$$

where $K_{\mu}(t)$ is the upper bound for the "general cumulant generating function (CGF)" for μ.

Our work: New f-Domain Discrepancy (f-DD)

Ziqiao Wang and Yongyi Mao. "On f-Divergence Principled Domain Adaptation: An Improved Framework." To appear at NeurIPS 2024.

- Our f-DD:

$$\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) \triangleq \sup_{t\in\mathbb{R},h'} \mathbb{E}_{\nu}\left[t\ell(h,h')\right] - \inf_{\alpha\in\mathbb{R}} \mathbb{E}_{\mu}\left[\phi^*(t\ell(h,h')+\alpha)-\alpha\right].$$

- Target Error Bound: For any $h \in \mathcal{H}$,

$$R_{\nu}(h) \leq R_{\mu}(h) + \inf_{t\geq 0} \frac{\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) + K_{\mu}(t)}{t} + \lambda^*, \tag{10}$$

where $K_{\mu}(t)$ is the upper bound for the "general cumulant generating function (CGF)" for μ.

- If ϕ is twice differentiable and ϕ'' is monotone, then

$$R_{\nu}(h) \leq R_{\mu}(h) + \sqrt{\frac{2}{\phi''(1)}\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu)} + \lambda^*. \tag{11}$$

e.g., $\phi''(1) = 1$ for KL recovers [Wang and Mao, 2023a, Theorem 4.2].

Shaper Bound: Localization Technique

- Restricted Hypothesis Space (Rashomon set): $\mathcal{H}_r \triangleq \{h \in \mathcal{H} | R_\mu(h) \leq r\}$

- Localized f-DD: For a given $h \in \mathcal{H}_{r_1}$

$$\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) \triangleq \sup_{h' \in \mathcal{H}_r, t \geq 0} \mathbb{E}_\nu \left[t\ell(h,h')\right] - \inf_{\alpha \in \mathbb{R}} \mathbb{E}_\mu \left[\phi^*(t\ell(h,h') + \alpha) - \alpha\right].$$

Shaper Bound: Localization Technique

- Restricted Hypothesis Space (Rashomon set):
  $\mathcal{H}_r \triangleq \{h \in \mathcal{H} | R_\mu(h) \leq r\}$

- Localized f-DD: For a given $h \in \mathcal{H}_{r_1}$

$$\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) \triangleq \sup_{h' \in \mathcal{H}_r, t \geq 0} \mathbb{E}_\nu \left[t\ell(h,h')\right] - \inf_{\alpha \in \mathbb{R}} \mathbb{E}_\mu \left[\phi^*(t\ell(h,h') + \alpha) - \alpha\right].$$

- Target Error Bound:
  For any h, h' and $C_1, C_2 > 0$ satisfying
  $\inf_\alpha \mathbb{E}_\mu \left[\phi^*(C_1\ell(h,h') + \alpha) - \alpha\right] \leq C_1(1 + C_2)\mathbb{E}_\mu \left[\ell(h,h')\right]$, then:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1}\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*,$$

  where $\lambda_r^* = \min_{h^* \in \mathcal{H}_r} R_\mu(h^*) + R_\nu(h^*)$ and
  $R_\mu^r(h) = \sup_{h' \in \mathcal{H}_r} \mathbb{E}_\mu \left[\ell(h,h')\right]$.

Localization Technique

- Target Error Bound:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1} \mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*.$$

  - $R_\mu^r(h) \leq r + r_1 \Longrightarrow$ Small r, r_1
  - If $r < \lambda^*$, then it's possible that $\lambda_r^* > \lambda^* \Longrightarrow$ Large r

Localization Technique

- Target Error Bound:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1}\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*.$$

  - $R_\mu^r(h) \leq r + r_1 \Longrightarrow$ Small r, r_1
  - If $r < \lambda^*$, then it's possible that $\lambda_r^* > \lambda^* \Longrightarrow$ Large r

- Localized KL-DD:
$\inf_\alpha \mathbb{E}_\mu\left[\phi^*(C_1\ell(h,h') + \alpha) - \alpha\right] \leq C_1(1+C_2)\mathbb{E}_\mu\left[\ell(h,h')\right]$

Localization Technique

- Target Error Bound:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1} \mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*.$$

  - $R_\mu^r(h) \leq r + r_1 \implies$ Small r, r_1
  - If $r < \lambda^*$, then it's possible that $\lambda_r^* > \lambda^* \implies$ Large r

- Localized KL-DD:
$\inf_\alpha \mathbb{E}_\mu \left[\phi^*(C_1 \ell(h,h') + \alpha) - \alpha\right] \leq C_1(1 + C_2)\mathbb{E}_\mu \left[\ell(h,h')\right]$

$$\Longleftarrow \begin{cases} C_1 > 0 \\ C_2 \in (0,1) \\ \left(e^{C_1} - C_1 - 1\right)\left[1 + (C_2^2 - 1)\min\{r_1 + r, 1\}\right] \leq C_1 C_2 \end{cases}$$

Generalization Bound via Localized f-DD

Theorem (informal)

For any $h \in \mathcal{H}_{r_1}$, w.p. at least $1-\delta$, we have

$$R_\nu(h) \leq R_{\hat{\mu}}(h) + \frac{\mathrm{D}_{\mathrm{KL}}^{h,\mathcal{H}_r}(\hat{\nu}||\hat{\mu})}{C_1} + C_2 R_\mu^r(h) + \mathcal{O}\left(\frac{\log(1/\delta)}{n} + \frac{\log(1/\delta)}{m}\right)$$
$$+ \mathcal{O}\left(\sqrt{\frac{(r_1+r)\log(1/\delta)}{n}} + \sqrt{\frac{r\log(1/\delta)}{m}}\right) + \text{Complexity.} + \lambda_r^*.$$

Small $r, r_1 \implies$ fast decaying rate (i.e. $\mathcal{O}\left(\frac{1}{n} + \frac{1}{m}\right)$).

Algorithm

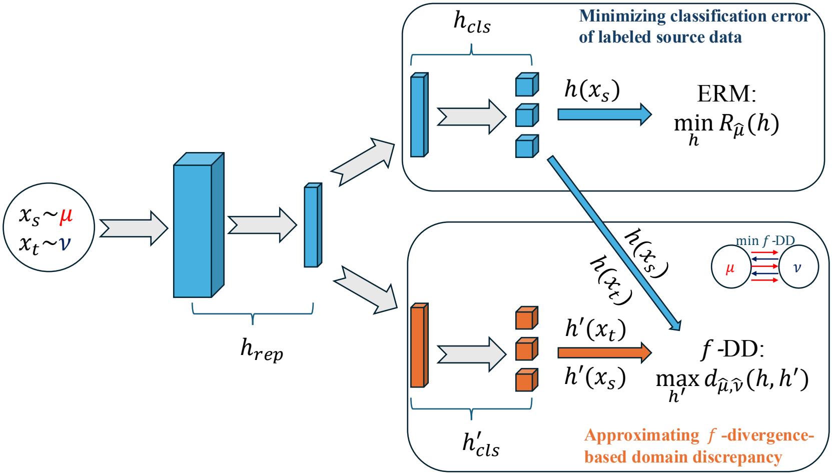


Figure 5: Overview of f-DD.

Experiments

- Three specific discrepancy measures:
  - KL-DD, χ^2-DD,
    the weighted Jeffereys-DD: $\gamma_1 \mathrm{D}_{\mathrm{KL}}(\hat{\mu}||\hat{\nu}) + \gamma_2 \mathrm{D}_{\mathrm{KL}}(\hat{\nu}||\hat{\mu})$
- Objective Function:
  Bounded $\ell \to$ Unbounded $\hat{\ell}$ (Optimizing over t may not be necessary)

$$\min_h R_{\hat{\mu}}(h) + \max_{h'} \left\{ \mathbb{E}_{\hat{\mu}} \left[\hat{\ell}(h, h') \right] - \inf_{\alpha} \mathbb{E}_{\hat{\nu}} \left[\phi^*(\hat{\ell}(h, h') + \alpha) - \alpha \right] \right\}.$$

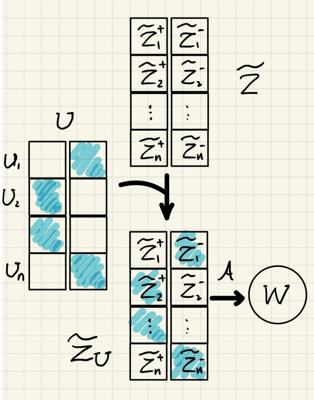
Table 1: Accuracy (%) on UDA Classification Tasks

Method	Office-31	Office-Home	Digits
Acuna et al. [2021]	89.5	68.5	96.3
Our weighted Jeffereys-DD	**90.1**	**70.2**	**97.1**

Generalization and CMI Setting

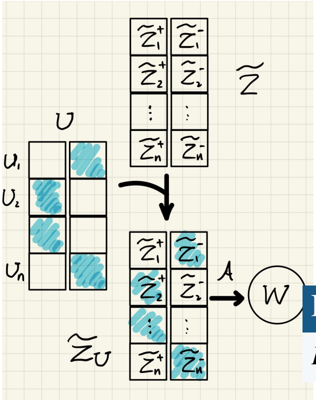
- Training dataset: $S = \{Z_i\}_{i=1}^n \in \mathcal{Z}$, drawn i.i.d. from μ
- Hypothesis space: $\mathcal{W} \subseteq \mathbb{R}^d$
- Learning algorithm: $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ by $P_{W|S}$
- Loss: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- We're interested in
  - Population risk: $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$
  - Empirical risk: $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$
  - Expected generalization error: $\mathcal{E} \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$

Generalization and CMI Setting



- Let $\widetilde{Z}$ drawn i.i.d. from μ
- Let $(U_1, U_2, \ldots, U_n)^T \sim \text{Unif}(\{0,1\}^n)$.
- Learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
- $\mathcal{E} =$
  $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(-1)^{U_i}\underbrace{(\ell(W,\widetilde{Z}_i^-) - \ell(W,\widetilde{Z}_i^+))}_{\Delta L_i}]$

Generalization and CMI Setting



- Let $\widetilde{Z}$ drawn i.i.d. from μ
- Let $(U_1, U_2, \ldots, U_n)^T \sim \mathrm{Unif}(\{0,1\}^n)$.
- Learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
- $\mathcal{E} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(-1)^{U_i} \underbrace{(\ell(W, \widetilde{Z}_i^-) - \ell(W, \widetilde{Z}_i^+))}_{\Delta L_i}]$

Lemma 1 (Wang and Mao [2023b])

For $\ell \in [0,1]$, $|\mathcal{E}| \le \frac{1}{n}\sum_{i=1}^n \sqrt{2I(\Delta L_i; U_i)}$.

Variational Representation of f-information

Let $I_\phi(X;Y) \triangleq \mathrm{D}_\phi(P_{X,Y}||P_X P_Y)$ be the f-information.
Let $G_i = (-1)^{U_i}\Delta L_i$, $P = P_{\Delta L_i, U_i}$ and $Q = P_{\Delta L_i} P_{U_i'}$.

$$\mathbb{E}_P\left[G_i\right] \leq \inf_{t\in\mathbb{R}} \frac{1}{t}\left(I_\phi(\Delta L_i;U_i) + \inf_{\alpha\in\mathbb{R}}\left\{\mathbb{E}_Q\left[\phi^*(tG_i+\alpha)\right]-\alpha\right\}\right) \tag{12}$$

$$\leq \inf_{t\in\mathbb{R}} \frac{1}{t}\left(I_\phi(\Delta L_i;U_i) + \mathbb{E}_Q\left[\phi^*(tG_i)\right]\right). \tag{13}$$

All the previous information-theoretic analysis focuses on upper bounding $\mathbb{E}_Q\left[\phi^*(tG_i)\right]$.

Our Work: New f-information Bounds

Ziqiao Wang and Yongyi Mao. "Generalization Bounds via Conditional f-Information." To appear at NeurIPS 2024.

Recall variational representation:

$$I_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{P_{X,Y}}\left[g(X,Y)\right] - \mathbb{E}_{P_X P_{Y'}}\left[\phi^*(g(X,Y'))\right].$$

Lemma 2 (Informal)

Let $g = \phi^{*-1} \circ (tf)$ *and let* Y' *be an independent copy of* Y. *If* $\mathbb{E}_{X,Y'}\left[f(X,Y')\right] = 0$, *then*

$$\sup_t \mathbb{E}_{X,Y}\left[\phi^{*-1}(tf(X,Y))\right] \leq I_\phi(X;Y).$$

Clearly,

$$\sup_t \mathbb{E}_{\Delta L_i, U_i}\left[\phi^{*-1}\left(tG_i\right)\right] \leq I_\phi(\Delta L_i; U_i).$$

Example: Mutual Information-based Bounds

Let $\phi(x) = x\log x + x - 1$ with $\phi^*(y) = e^y - 1$ and $\phi^{*-1}(z) = \log(1+z)$.

- ▷ Lemma 2 gives us $I(\Delta L_i; U_i) \geq \sup_t \mathbb{E}\left[\log\left(1+t(-1)^{U_i}\Delta L_i\right)\right]$.

Example: Mutual Information-based Bounds

Let $\phi(x) = x\log x + x - 1$ with $\phi^*(y) = e^y - 1$ and $\phi^{*-1}(z) = \log(1+z)$.

- Lemma 2 gives us $I(\Delta L_i; U_i) \geq \sup_t \mathbb{E}\left[\log\left(1+t(-1)^{U_i}\Delta L_i\right)\right]$.
- Let $f(x) = \log(1+x) - x + ax^2$ and set $a = \frac{|\mathbb{E}[G_i]|}{2\mathbb{E}\left[G_i^2\right]} + \frac{1}{2}$.

  Additional lemma: $f(x) \geq 0$ holds when $a \geq \frac{1}{2}$ and $|x| \leq 1 - \frac{1}{2a}$

Example: Mutual Information-based Bounds

Let $\phi(x) = x \log x + x - 1$ with $\phi^*(y) = e^y - 1$ and $\phi^{*-1}(z) = \log(1+z)$.

- Lemma 2 gives us $I(\Delta L_i; U_i) \geq \sup_t \mathbb{E}\left[\log\left(1+t(-1)^{U_i}\Delta L_i\right)\right]$.
- Let $f(x) = \log(1+x) - x + ax^2$ and set $a = \frac{|\mathbb{E}[G_i]|}{2\mathbb{E}\left[G_i^2\right]} + \frac{1}{2}$.
  Additional lemma: $f(x) \geq 0$ holds when $a \geq \frac{1}{2}$ and $|x| \leq 1 - \frac{1}{2a}$
- Hence,
  $\sup_{t>-1} \mathbb{E}\left[\log\left(1+tG_i\right)\right] \geq \sup_{t\in[\frac{1}{2a}-1, 1-\frac{1}{2a}]} \mathbb{E}\left[tG_i - at^2G_i^2\right]$. The supremum is attained when $t^* = \frac{\mathbb{E}[G_i]}{2a\mathbb{E}\left[G_i^2\right]}$, which is achievable.
- We have
  $I(\Delta L_i; U_i) \geq \sup_{t>-1} \mathbb{E}_{\Delta L_i, U_i}\left[\log\left(1+t(-1)^{U_i}\Delta L_i\right)\right] \geq \frac{\mathbb{E}^2[G_i]}{4a\mathbb{E}\left[G_i^2\right]}$,
  which simplifies to

$$|\mathbb{E}[G_i]| \leq \sqrt{2\left(|\mathbb{E}\left[G_i\right]| + \mathbb{E}\left[G_i^2\right]\right) I(\Delta L_i; U_i)}. \tag{14}$$

CMI Bounds

Theorem 2

Assume the loss difference $\ell(w, z_1) - \ell(w, z_2)$ is bounded in $[-1, 1]$ for any $w \in \mathcal{W}$ and $z_1, z_2 \in \mathcal{Z}$, we have

$$|\mathcal{E}| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2 \left(\mathbb{E}\left[\Delta L_i^2\right] + \left|\mathbb{E}\left[G_i\right]\right| \right) I(\Delta L_i; U_i)}.$$

Notably, using solely $I(\Delta L_i; U_i)$ (and other variants of CMI measures) to characterize generalization is loose.

Corollary 3

Under the conditions of Theorem 2, we have

$$|\mathcal{E}| \leq \frac{1}{n} \sum_{i=1}^{n} \left(2I(\Delta L_i; U_i) + 2\sqrt{2\mathrm{Var}\left(L_i^+\right) I(\Delta L_i; U_i)} \right).$$

Further Comments

- Similar bounds can be obtained for other f-information (f-divergence) such as χ^2-divergence, squared Hellinger distance, Jensen-Shannon divergence, ...
- We also extend results to the unbounded loss function case by using the truncation trick.
- For more work on f-divergence, check Nguyen et al. [2010], Jiao et al. [2017], Birrell et al. [2022], Agrawal and Horel [2020, 2021], Polyanskiy and Wu [2022].

Jeremiah Birrell, Markos A Katsoulakis, and Yannis Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *IEEE Transactions on Information Theory*, 68(7):4553–4572, 2022.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

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

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *The 22nd Conference on Learning Theory*, 2009.

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pages 66–75. PMLR, 2021.

Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation. In *International Conference on Learning Representations*, 2023a.

Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In *International Conference on Machine Learning*. PMLR, 2023b.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861, 2010.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479. IEEE, 2017.

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. In *International Conference on Machine Learning*, pages 115–124. PMLR, 2020.

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *The Journal of Machine Learning Research*, 22(1):5662–5720, 2021.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge university press, 2022.

Thanks!