When Optimizing f-divergence is Robust with Label Noise

Jiaheng Wei and Yang Liu

University of California, Santa Cruz
Department of Computer Science and Engineering
{jiahengwei,yangliu}@ucsc.edu

An observation of $f-$divergence (D_f)

For an arbitrary classifier $h \in \mathcal{H}$, dataset (X, Y), define

$$\text{Joint distribution:} \quad P_{h\times Y} = \mathbb{P}(h(X) = y, Y = y'),$$

$$\text{Product distribution:} \quad Q_{h\times Y} = \mathbb{P}(h(X) = y) \cdot \mathbb{P}(Y = y').$$

Learning using D_f often returns us a good classifier h_f^*!

$$h_f^* = \underset{h\in\mathcal{H}}{\operatorname{argmax}}\, D_f\left(P_{h\times Y} || Q_{h\times Y}\right)$$

[More in Theorem 1, 3; Table 2]

Our questions

How robust is the optimization of D_f when label noise presents?

Learning with noisy labels

Only have $(X, \tilde{Y})$, noise transition matrix T: $T_{i,j} = \mathbb{P}(\tilde{Y} = j | Y = i)$.
$K-$class noise setting:

- Uniform noise: $\forall i \neq j,\ e_j = T_{i,j}$.
- Sparse noise: disjoint pairs of classes (i_c, j_c), $T_{i_c,j_c} = e_{p_1}$, $T_{j_c,i_c} = e_{p_2}$.

Why we think so?

Peer loss [Liu and Guo, ICML'20]:

Robust loss with theoretical guarantee & No need of noise rates

Motivation:

Expectation of peer loss (w.r.t. ℓ_{CE}) is similar to variational form of D_f:

$$\tilde{h}^*_{\mathsf{peer}} := \operatorname*{argmax}_{h \in \mathcal{H}} \underbrace{\mathbb{E}_{\tilde{D}}\left[-\ell_{\mathsf{CE}}(h(X), \tilde{Y})\right]}_{\Longrightarrow \mathbb{P}(h(X)=\tilde{Y})} - \underbrace{\mathbb{E}_{\tilde{D}}\left[-\ell_{\mathsf{CE}}(h(X), \tilde{Y}_{\mathsf{rand}})\right]}_{\Longrightarrow \mathbb{P}(h(X)=\tilde{Y}_{\mathsf{rand}})}$$

Fit the data! (Joint Distribution)

Be different! (Product Distribution)

Variational form of $f-$divergence

An empirical alternative

Variational form of D_f on the noisy data $X, \tilde{Y}$:

$$\tilde{h}_f^* = \underset{h\in\mathcal{H}}{\operatorname{argmax}} \sup_g \underline{\widetilde{\mathsf{VD}}_f(h,g)} := \underline{\mathbb{E}_{Z\sim\tilde{P}_{h\times\tilde{Y}}}[g(Z)] - \mathbb{E}_{Z\sim\tilde{Q}_{h\times\tilde{Y}}}[f^*(g(Z))]}$$

$g(Z) = g(h(X), \tilde{Y})$ $\qquad$ $g(Z) = g(h(X), \tilde{Y}_{\text{rand}})$

where $f^*(u) = \sup_{v\in\mathbb{R}}\{uv - f(v)\}$ is the Fenchel duality of $f(u)$.

Variational difference with noisy labels

Variational difference with noisy labels $\widetilde{\mathbf{VD}}_f(h, g)$ is an affine transformation of $\mathbf{VD}_f(h, g)$:

Theorem 4

$$\widetilde{VD}_f(h, g) = (1 - e_1 - e_2)VD_f(h, g) + Bias_f(h, g)$$

where $\text{Bias}_f(h, g) := \sum_{i \in \{1,2\}} e_i \cdot (\mathbb{E}_X[g(h(X), i)] - \mathbb{E}_X[f^*(g(h(X), i))])$.

More theoretical results

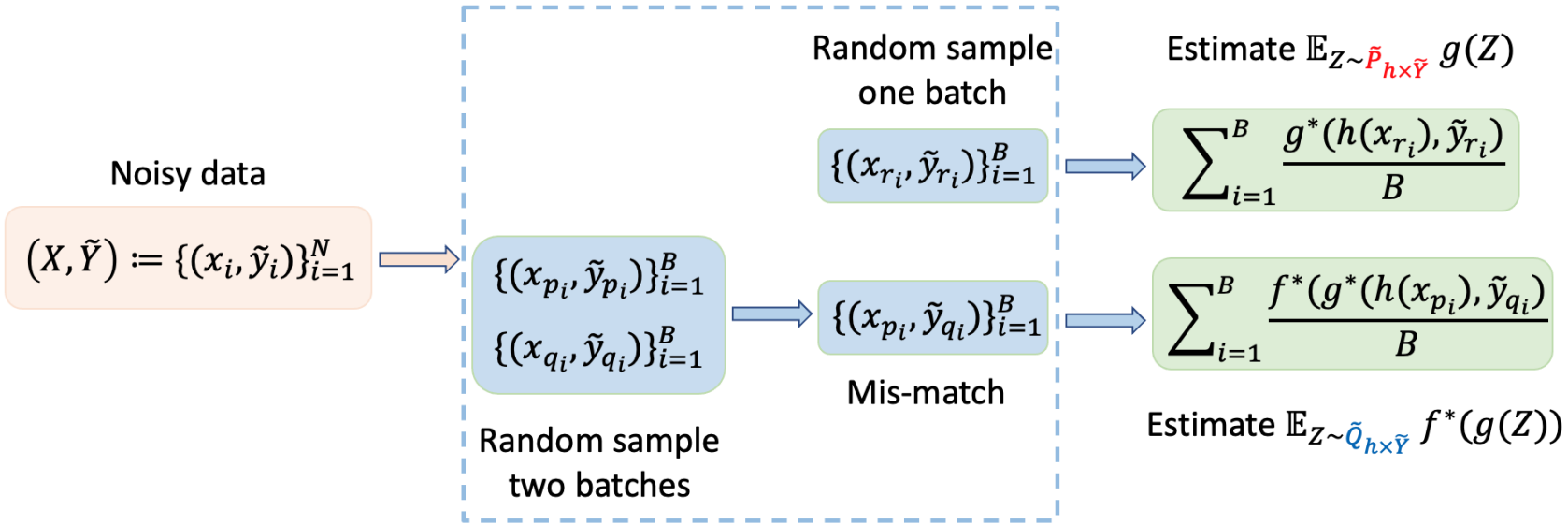
Our theoretical results also include:

- **Impact of Bias$_f$ is diminishing when noise rates are high** [Lemma 1]
- **When optimizing D_fs are robust** [Theorem 6, 7, 8]
- **Multi-class extension of Theorem 4** [Theorem 5, 9]

Implementation

Practical workflow

We adopt fixed g^* without updating from **f-GAN** [Nowozin et.al]

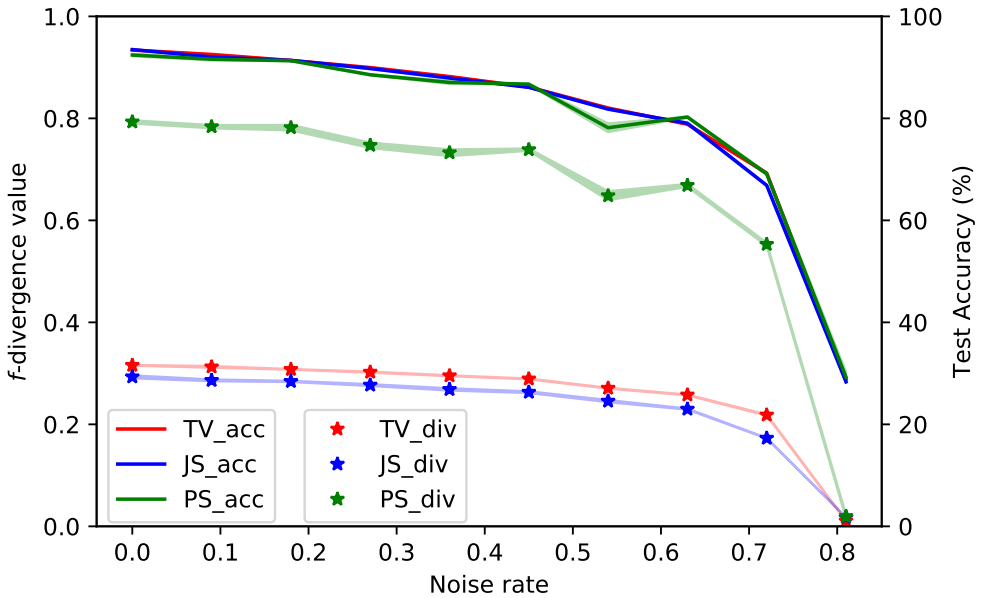


Robustness of typical $f-$divergence functions

What $f-$divergence functions are robust in practice

D_f	T-V	J-S	KL	Pearson	Jeffrey	Reverse-KL	S-H
Robustness	✓	✓	✓	✓	✓	✗	✗

How robust are D_f's on CIFAR-10



More works and acknowledgement

Other Relevant Works

- A more challenging noise setting, ICLR'21
  Learning with Instance-Dependent Label Noise: A Sample Sieve Approach
- High-order statistics: CVPR'21 (oral)
  A Second-Order Approach to Learning with Instance-Dependent Label Noise

Acknowledgement

This work is partially supported by the National Science Foundation (NSF) under grant IIS-2007951 and the Office of Naval Research under grant N00014-20-1-22.

Q&A

Meet us in Poster Session 3: May 3 at 17:00-19:00 PDT!
Our code is available at
`https://github.com/UCSC-REAL/Robust-f-divergence-measures`