A PROOF AND ADDITIONAL THEOREMS

A.1 SHORT NOTATIONS

Throughout the appendix, we will denote by $p := \mathbb{P}(Y = +1)$ and $\tilde{p} := \mathbb{P}(\tilde{Y} = +1) = p \cdot (1 - e_+) + (1 - p) \cdot e_-$. We will also shorthand $\tilde{P}, \tilde{Q}$ for $\tilde{P}_{h\times\tilde{Y}}, \tilde{Q}_{h\times\tilde{Y}}$ and P, Q for $P_{h\times Y}, Q_{h\times Y}$.

A.2 FULL TABLES OF TABLE 1

In experiments, we adopt the output activation function $g_f(v)$ instead of optimal activation functions g^* by referring to the implementation of f-GAN (Nowozin et al., 2016).

Name	$g_f(v)$	g^*	dom_{f^*}	$f^*(u)$
Total Variation (✓)	$\frac{1}{2}\tanh(v)$	$\frac{1}{2}\,\text{sign}\,\frac{p(z)}{q(z)} - 1$	$u \in [-\frac{1}{2}, \frac{1}{2}]$	u
Jenson-Shannon (✓)	$\log\frac{2}{1+e^{-v}}$	$\log\frac{2p(z)}{p(z)+q(z)}$	$u < \log 2$	$-\log(2 - e^u)$
Squared Hellinger (✗)	$1 - e^v$	$1 - \sqrt{\frac{q(z)}{p(z)}}$	$u < 1$	$\frac{u}{1-u}$
Pearson $\mathcal{X}^2$ (✓)	v	$2\left(\frac{p(z)}{q(z)} - 1\right)$	$\mathbb{R}$	$\frac{1}{4}u^2 + u$
Neyman $\mathcal{X}^2$ (✗)	$1 - e^v$	$1 - \left(\frac{q(z)}{p(z)}\right)^2$	$u < 1$	$2 - 2\sqrt{1-u}$
KL (✓)	v	$1 + \log\frac{p(z)}{q(z)}$	$\mathbb{R}$	e^{u-1}
Reverse KL (✗)	$-e^v$	$-\frac{q(z)}{p(z)}$	$\mathbb{R}_-$	$-1 - \log(-u)$
Jeffrey (✓)	v	$1 + \log\frac{p(z)}{q(z)} - \frac{q(z)}{p(z)}$	$\mathbb{R}$	$W(e^{1-u}) + \frac{1}{W(e^{1-u})} + u - 2$

Table 6: Exemplary output activation functions g_f (used for approximating g, see e.g. (Nowozin et al., 2016)), optimal activation functions, optimal conjugate functions (full table). W is the Lambert$-W$ product log function. '✓' indicates that the $f-$divergence function (in practice) is robust to label noise and '✗' means non-robust.

A.3 MAIN RESULTS: PROOF OF THEOREM 4

Proof. First note

$$
\begin{aligned}
&\mathbb{E}_{\tilde{Z}\sim\tilde{P}}\left[g(\tilde{Z})\right] \\
&= p\cdot\mathbb{E}_{\tilde{Z}\sim\tilde{P}|Y=+1}\left[g(\tilde{Z})\right] + (1-p)\cdot\mathbb{E}_{\tilde{Z}\sim\tilde{P}|Y=-1}\left[g(\tilde{Z})\right] \\
&= p\cdot\mathbb{E}_{X|Y=+1}\left[(1-e_+)\cdot g(h(X),+1) + e_+\cdot g(h(X),-1)\right] \\
&\quad + (1-p)\cdot\mathbb{E}_{X|Y=-1}\left[(1-e_-)\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)\right] \\
&= p\cdot\mathbb{E}_{X|Y=+1}\left[(1-e_+-e_-)\cdot g(h(X),+1) + e_+\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)\right] \\
&\quad + (1-p)\cdot\mathbb{E}_{X|Y=-1}\left[(1-e_+-e_-)\cdot g(h(X),-1) + e_+\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)\right] \\
&= (1-e_+-e_-)\cdot\mathbb{E}_{Z\sim P}[g(Z)] + \mathbb{E}_X[e_+\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)]
\end{aligned}
$$

The second term in the variational difference derives as:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{Z}\sim\tilde{Q}}\left[f^*(g(\tilde{Z}))\right] \\
&= \tilde{p}\cdot\mathbb{E}_X\left[f^*(g(h(X),+1))\right] + (1-\tilde{p})\cdot\mathbb{E}_X\left[f^*(g(h(X),-1))\right] \\
&= p\cdot(1-e_+-e_-)\cdot\mathbb{E}_X[f^*(g(h(X),+1))] + e_-\cdot\mathbb{E}_X[f^*(g(h(X),+1))] \\
&\quad + (1-p)\cdot(1-e_+-e_-)\cdot\mathbb{E}_X[f^*(g(h(X),-1))] + e_+\cdot\mathbb{E}_X[f^*(g(h(X),-1))] \\
&= (1-e_+-e_-)\cdot\mathbb{E}_{Z\sim Q}[f^*(g(Z))] + \mathbb{E}_X[e_-\cdot f^*(g(h(X),+1)) + e_+\cdot f^*(g(h(X),-1))]
\end{aligned}
$$

Combining $\mathbb{E}_{\tilde{Z}\sim\tilde{P}}\left[g(\tilde{Z})\right]$ and $\mathbb{E}_{\tilde{Z}\sim\tilde{Q}}\left[f^*(g(\tilde{Z}))\right]$: the leading terms combine into

$$
(1-e_+-e_-)\cdot\mathbb{E}_{Z\sim P}[g(Z)] - (1-e_+-e_-)\cdot\mathbb{E}_{Z\sim Q}[f^*(g(Z))] = (1-e_+-e_-)\cdot\mathsf{VD}_f(h,g)
$$

and the rest:

$$\begin{aligned}&\mathbb{E}_X[e_+ \cdot g(h(X), -1) + e_- \cdot g(h(X), +1)] - \mathbb{E}_X[e_- \cdot f^*(g(h(X), +1)) + e_+ \cdot f^*(g(h(X), -1))] \\ &= e_+ \cdot \Delta_f^{-1}(h, g) + e_- \cdot \Delta_f^{+1}(h, g) \\ &= \mathsf{Bias}_f(h, g)\end{aligned}$$

we proved the claim. □

A.4 Proof of Theorem 5: Multi-class extension I of Theorem 4

Proof. Denote $p_i = \mathbb{P}(Y = i)$, $\tilde{p}_i := \mathbb{P}(\tilde{Y} = i)$, $\tilde{p}_i = (1 - \sum_{j \neq i} e_j) \cdot p_i + e_i \cdot \sum_{j \neq i} p_j$. We have:

$$\begin{aligned}&\mathbb{E}_{\tilde{Z} \sim \tilde{P}} \left[g(\tilde{Z}) \right] \\ &= \sum_{i=1}^K p_i \cdot \mathbb{E}_{\tilde{Z} \sim \tilde{P} | Y = i} \left[g(\tilde{Z}) \right] = \sum_{i=1}^K p_i \cdot \mathbb{E}_{X | Y = i} \left[\sum_{j=1}^K T_{i,j} \cdot g(h(X), \tilde{Y} = j) \right] \\ &= \sum_{i=1}^K p_i \cdot \mathbb{E}_{X | Y = i} \left[(1 - \sum_{j \neq i} e_j) \cdot g(h(X), \tilde{Y} = i) + \sum_{j \neq i} e_j \cdot g(h(X), \tilde{Y} = j) \right] \\ &= \sum_{i=1}^K p_i \cdot \mathbb{E}_{X | Y = i} \left[(1 - \sum_{j=1}^K e_j) \cdot g(h(X), \tilde{Y} = i) + \sum_{j=1}^K e_j \cdot g(h(X), \tilde{Y} = j) \right] \\ &= (1 - \sum_{j=1}^K e_j) \cdot \mathbb{E}_{Z \sim P}[g(Z)] + \sum_{j=1}^K e_j \cdot \mathbb{E}_X[g(h(X), j)]\end{aligned}$$

$$\begin{aligned}&\mathbb{E}_{\tilde{Z} \sim \tilde{Q}} \left[f^*(g(\tilde{Z})) \right] \\ &= \sum_{i=1}^K \tilde{p}_i \cdot \mathbb{E}_X \left[f^*(g(h(X), \tilde{Y} = i)) \right] \\ &= \sum_{i=1}^K [(1 - \sum_{j=1}^K e_j) \cdot p_i + e_i \cdot \sum_{j=1}^K p_j] \cdot \mathbb{E}_X \left[f^*(g(h(X), \tilde{Y} = i)) \right] \\ &= (1 - \sum_{j=1}^K e_j) \cdot \mathbb{E}_{Z \sim Q}[f^*(g(Z))] + \sum_{j=1}^K e_j \cdot \mathbb{E}_X \left[f^*(g(h(X), j)) \right]\end{aligned}$$

Similar to the binary case, combining $\mathbb{E}_{\tilde{Z} \sim \tilde{P}} \left[g(\tilde{Z}) \right]$ and $\mathbb{E}_{\tilde{Z} \sim \tilde{Q}} \left[f^*(g(\tilde{Z})) \right]$ we proved the claim. □

A.5 Multi-class extension II of Theorem 4: sparse case

For sparse transition matrix, assume K is an even number, sparse noise model specifies $\frac{K}{2}$ disjoint pairs of classes (i_c, j_c) where $c \in [\frac{K}{2}]$ and $i_c < j_c$. The diagonal entry T_{i_c, i_c} becomes $1 - T_{i_c, j_c}$. Suppose $\forall c \in [\frac{K}{2}], T_{i_c, j_c} = e_{p_1}, T_{j_c, i_c} = e_{p_2}, e_{p_1} + e_{p_2} < 1$.

Theorem 9. *[Multi-class extension II] In the scenario of sparse noise transition model, the variational difference between the noisy distributions $\tilde{P}$ and $\tilde{Q}$ relates to the one defined over the clean distributions in the following way:*

$$\widetilde{\mathsf{VD}}_f(h, g) = (1 - e_{p_1} - e_{p_2}) \cdot \mathsf{VD}_f(h, g) + \sum_{(i_c, j_c)} \left[e_{p_1} \cdot \Delta_f^{j_c}(g, h) + e_{p_2} \cdot \Delta_f^{i_c}(g, h) \right].$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{\tilde{Z}\sim\tilde{P}}\left[g(\tilde{Z})\right] \\
&= \sum_{i=1}^{K} p_i \cdot \mathbb{E}_{\tilde{Z}\sim\tilde{P}|Y=i}\left[g(\tilde{Z})\right] = \sum_{i=1}^{K} p_i \cdot \mathbb{E}_{X|Y=i}\left[\sum_{j=1}^{K} T_{i,j}\cdot g(h(X),\tilde{Y}=j)\right] \\
&= \sum_{i_c} p_{i_c}\cdot \mathbb{E}_{X|Y=i_c}\left[(1-e_{p_1})\cdot g(h(X),\tilde{Y}=i_c) + e_{p_1}\cdot g(h(X),\tilde{Y}=j_c)\right] \\
&\quad + \sum_{j_c} p_{j_c}\cdot \mathbb{E}_{X|Y=j_c}\left[(1-e_{p_2})\cdot g(h(X),\tilde{Y}=j_c) + e_{p_2}\cdot g(h(X),\tilde{Y}=i_c)\right] \\
&= \sum_{i=1}^{K} p_i \cdot \mathbb{E}_{X|Y=i}\Big[(1-e_{p_1}-e_{p_2})\cdot g(h(X),\tilde{Y}=i)\Big] \\
&\quad + \sum_{(i_c,j_c)} \left(e_{p_1}\cdot g(h(X),\tilde{Y}=j_c) + e_{p_2}\cdot g(h(X),\tilde{Y}=i_c)\right) \\
&= (1-e_{p_1}-e_{p_2})\cdot \mathbb{E}_{Z\sim P}[g(Z)] + \sum_{(i_c,j_c)} \mathbb{E}_X\Big[e_{p_1}\cdot g(h(X),\tilde{Y}=j_c) + e_{p_2}\cdot g(h(X),\tilde{Y}=i_c)\Big]
\end{aligned}
$$

Similarly, we have:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{Z}\sim\tilde{Q}}\left[f^*(g(\tilde{Z}))\right] \\
&= \sum_{i=1}^{K} \tilde{p}_i \cdot \mathbb{E}_X\left[f^*(g(h(X),\tilde{Y}=i))\right] \\
&= \sum_{i=1}^{K}\left[(1-\sum_{j=1}^{K} e_j)\cdot p_i + e_i\cdot\sum_{j=1}^{K} p_j\right]\cdot \mathbb{E}_X\left[f^*(g(h(X),\tilde{Y}=i))\right] \\
&= (1-e_{p_1}-e_{p_2})\cdot \mathbb{E}_{Z\sim Q}[f^*(g(Z))] \\
&\quad + \sum_{(i_c,j_c)} \mathbb{E}_X\Big[e_{p_1}\cdot f^*(g(h(X),\tilde{Y}=j_c)) + e_{p_2}\cdot f^*(g(h(X),\tilde{Y}=i_c))\Big]
\end{aligned}
$$

Combining $\mathbb{E}_{\tilde{Z}\sim\tilde{P}}\left[g(\tilde{Z})\right]$ and $\mathbb{E}_{\tilde{Z}\sim\tilde{Q}}\left[f^*(g(\tilde{Z}))\right]$ we proved the claim. □

A.6 Proof of Theorem 1: total-variation generates Bayes optimal

Proof. For total variation, we have

$$
D_f(P_{h\times Y}\|Q_{h\times Y}) = \frac{1}{2}\sum_{y,y'\in\{-1,+1\}} |\mathbb{P}(h(X)=y, Y=y') - \mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y')|
$$

We again present the main proof for the binary classification setting.

First note the following fact that

$$
\mathbb{P}(h(X)=y, Y=y) - \mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y)
$$

and

$$
\mathbb{P}(h(X)=y, Y=-y) - \mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=-y)
$$

have opposite signs. This is simply because

$$
\mathbb{P}(h(X)=y, Y=y) - \mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y) + \mathbb{P}(h(X)=y, Y=-y) - \mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=-y) = 0
$$

Because of the above, there are four possible combinations of cases:

Case 1 $\mathbb{P}(h(X)=+1,Y=+1)-\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=+1)>0, \mathbb{P}(h(X)=-1,Y=-1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=-1)>0$:

$$\begin{aligned}
&\sum_{y,y'\in\{-1,+1\}} |\mathbb{P}(h(X)=y,Y=y')-\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y')| \\
&=\mathbb{P}(h(X)=+1,Y=+1)-\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=+1) \\
&\quad-\mathbb{P}(h(X)=+1,Y=-1)+\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=-1) \\
&\quad+\mathbb{P}(h(X)=-1,Y=-1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=-1) \\
&\quad-\mathbb{P}(h(X)=-1,Y=+1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=+1) \\
&=\mathbb{P}(h(X)=Y)-\mathbb{P}(h(X)\neq Y) \\
&=2\mathbb{P}(h(X)=Y)-1
\end{aligned}$$

Therefore, maximizing D_f total variation returns the Bayes optimal classifier h^*, and the optimal value arrives at $\mathbb{P}(h^*(X)=Y)-\frac{1}{2}$.

Case 2 $\mathbb{P}(h(X)=+1,Y=+1)-\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=+1)<0, \mathbb{P}(h(X)=-1,Y=-1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=-1)>0$:

$$\begin{aligned}
&\sum_{y,y'\in\{-1,+1\}} |\mathbb{P}(h(X)=y,Y=y')-\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y')| \\
&=-\mathbb{P}(h(X)=+1,Y=+1)+\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=+1) \\
&\quad+\mathbb{P}(h(X)=+1,Y=-1)-\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=-1) \\
&\quad+\mathbb{P}(h(X)=-1,Y=-1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=-1) \\
&\quad-\mathbb{P}(h(X)=-1,Y=+1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=+1) \\
&=\mathbb{P}(Y=-1)-\mathbb{P}(Y=+1) \\
&=0
\end{aligned}$$

Case 3 $\mathbb{P}(h(X)=+1,Y=+1)-\mathbb{P}(h(X)=+1)\cdot\mathbb{P}(Y=+1)>0, \mathbb{P}(h(X)=-1,Y=-1)-\mathbb{P}(h(X)=-1)\cdot\mathbb{P}(Y=-1)<0$: This case is symmetrical to Case 2.

Case 4 This is symmetrical to Case 1:

$$D_f(P_{h\times Y}\|Q_{h\times Y})=\mathbb{P}(h(X)\neq Y)-\frac{1}{2}$$

The optimal classifier is then the opposite of h^*, but

$$\mathbb{P}(h^*(X)=Y)-\frac{1}{2}>\mathbb{P}(h^*(X)\neq Y)-\frac{1}{2}$$

so the maximizer returns a smaller value compared to Case 1.

Multi-class extension We provide arguments for the multi-class generalization. First note that

$$\begin{aligned}
&\sum_{y,y'\in\mathcal{Y}} |\mathbb{P}(h(X)=y,Y=y')-\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y')| \\
=&\sum_{y,y'\in\mathcal{Y}} |\mathbb{P}(h(X)=y|Y=y')-\mathbb{P}(h(X)=y)|\cdot\mathbb{P}(Y=y') \\
=&\frac{1}{K}\sum_{y'}\sum_{y} |\mathbb{P}(h(X)=y|Y=y')-\mathbb{P}(h(X)=y)|
\end{aligned}$$

For any classifier h and for each y, one of the following terms

$\mathbb{P}(h(X)=y,Y=1)-\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=1), ..., \mathbb{P}(h(X)=y,Y=K)-\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=K)$

must be non-negative as: $\sum_{y'}\mathbb{P}(h(X)=y,Y=y')-\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y=y')=0$.

Our following derivation focuses on confident classifiers:

Definition 3. *We call a classifier confident if for each label class y, only one class $y_k \in \mathcal{Y}$ returns positive correlation:*

$$\mathbb{P}(h(X) = y_k, Y = k) - \mathbb{P}(h(X) = y_k) \cdot \mathbb{P}(Y = k) \geq 0$$

while for all other $y' \neq y_k$, we have $\mathbb{P}(h(X) = y', Y = k) - \mathbb{P}(h(X) = y') \cdot \mathbb{P}(Y = k) \leq 0$

This above definition is saying the classifier h is "dominantly" confident in predicting one class for the each true label class.

For a given class k, if all other classes $k' \neq y_k$ are negative in $\mathbb{P}(h(X) = k', Y = k) - \mathbb{P}(h(X) = k') \cdot \mathbb{P}(Y = k)$, the total variation becomes:

$$\begin{aligned}
&\sum_{k'} |\mathbb{P}(h(X) = k', Y = k) - \mathbb{P}(h(X) = k') \cdot \mathbb{P}(Y = k)| \\
=&\mathbb{P}(h(X) = y_k, Y = k) - \mathbb{P}(h(X) = y_k) \cdot \mathbb{P}(Y = k) \\
&+ \sum_{k' \neq y_k} (\mathbb{P}(h(X) = k') \cdot \mathbb{P}(Y = k) - \mathbb{P}(h(X) = k', Y = k)) \\
=&\mathbb{P}(h(X) = y_k, Y = k) - \mathbb{P}(h(X) = y_k) \cdot \mathbb{P}(Y = k) \\
&+ \mathbb{P}(Y = k)(1 - \mathbb{P}(h(X) = y_k) - (1 - \mathbb{P}(h(X) = y_k, Y = k)) \\
=&2(\mathbb{P}(h(X) = y_k, Y = k) - \mathbb{P}(h(X) = y_k) \cdot \mathbb{P}(Y = k)).
\end{aligned}$$

Summing up, for a confident classifier, the total variation becomes (ignoring constant 2):

$$\begin{aligned}
&\sum_k \mathbb{P}(h(X) = y_k, Y = k) - \mathbb{P}(h(X) = y_k) \cdot \mathbb{P}(Y = k) \\
=&\frac{1}{K} \sum_k \mathbb{P}(h(X) = y_k | Y = k) - \mathbb{P}(h(X) = y_k) \\
=&\frac{1}{K} \sum_k \mathbb{P}(h(X) = y_k | Y = k) - 1 \\
=&\frac{1}{K} \sum_k \frac{\mathbb{P}(Y = k | h(X) = y_k)\mathbb{P}(h(X) = y_k)}{\mathbb{P}(Y = k)} - 1 \\
=&\sum_k \int_X f_X(x) \cdot \mathbb{P}(Y = k | h(x) = y_k)\mathbb{P}(h(x) = y_k)\, dx - 1 \\
=&\sum_k \int_X f_X(x) \cdot \mathbb{P}(Y = k | X = x) \cdot 1(h(x) = y_k) \cdot \mathbb{P}(h(x) = y_k)\, dx - 1 \\
=&\int_X f_X(x) \cdot \sum_k \mathbb{P}(Y = k | X = x) \cdot 1(h(x) = y_k) \cdot \mathbb{P}(h(x) = y_k)\, dx - 1 \\
\leq&\int_X f_X(x) \cdot \sum_k \mathbb{P}(Y = k | X = x) \cdot 1(h^*(X) = k) \cdot \mathbb{P}(h^*(X) = k)\, dx - 1 \\
=&\sum_k \mathbb{P}(h^*(X) = k, Y = k) - \mathbb{P}(h^*(X) = k) \cdot \mathbb{P}(Y = k).
\end{aligned}$$

where the last inequality is due to the fact that the Bayes optimal classifier selects the highest $\mathbb{P}(Y = k|X)$ for each x.

□

A.7 Proof of Theorem 3

Proof. By definition of D_f:

$$D_f(P_{h \times Y^*} \| Q_{h \times Y^*}) = \sum_{y, y'} \mathbb{P}(h(X) = y, Y^* = y') \cdot f\left(\frac{\mathbb{P}(h(X) = y, Y^* = y')}{\mathbb{P}(h(X) = y) \cdot \mathbb{P}(Y^* = y')}\right)$$

First we want to prove

$$\left|\frac{\mathbb{P}(h^*(X)=y,Y^*=y)}{\mathbb{P}(h^*(X)=y)\cdot\mathbb{P}(Y^*=y)}-1\right|>\left|\frac{\mathbb{P}(h(X)=y',Y^*=y)}{\mathbb{P}(h(X)=y')\cdot\mathbb{P}(Y^*=y)}-1\right|,\forall h\neq h^*$$

This is because:

$$\frac{\mathbb{P}(h^*(X)=y,Y^*=y)}{\mathbb{P}(h^*(X)=y)\cdot\mathbb{P}(Y^*=y)}=\frac{\mathbb{P}(h^*(X)=y|Y^*=y)}{\mathbb{P}(h^*(X)=y)}=\frac{1}{\mathbb{P}(Y^*=y)}$$

On the other hand

$$\begin{aligned}&\frac{\mathbb{P}(h(X)=y',Y^*=y)}{\mathbb{P}(h(X)=y')\cdot\mathbb{P}(Y^*=y)}\\=&\frac{\mathbb{P}(h(X)=y'|Y^*=y)}{\mathbb{P}(h(X)=y'|Y^*=y)\mathbb{P}(Y^*=y)+\mathbb{P}(h(X)=y'|Y^*=-y)\mathbb{P}(Y^*=-y)}\\=&\frac{1}{\mathbb{P}(Y^*=y)+\frac{\mathbb{P}(h(X)=y'|Y^*=-y)}{\mathbb{P}(h(X)=y'|Y^*=y)}\mathbb{P}(Y^*=-y)}\end{aligned}$$

When $\frac{\mathbb{P}(h(X)=y'|Y^*=-y)}{\mathbb{P}(h(X)=y'|Y^*=y)}<1$, $\mathbb{P}(Y^*=y)+\frac{\mathbb{P}(h(X)=y'|Y^*=-y)}{\mathbb{P}(h(X)=y'|Y^*=y)}\mathbb{P}(Y^*=-y)<\mathbb{P}(Y^*=y)+\mathbb{P}(Y^*=-y)=1$, therefore $\frac{\mathbb{P}(h(X)=y',Y^*=y)}{\mathbb{P}(h(X)=y')\cdot\mathbb{P}(Y^*=y)}>1$. Further

$$\frac{\mathbb{P}(h(X)=y',Y^*=y)}{\mathbb{P}(h(X)=y')\cdot\mathbb{P}(Y^*=y)}<\frac{1}{\mathbb{P}(Y^*=y)}=\frac{\mathbb{P}(h^*(X)=y,Y^*=y)}{\mathbb{P}(h^*(X)=y)\cdot\mathbb{P}(Y^*=y)}$$

When $\frac{\mathbb{P}(h(X)=y'|Y^*=-y)}{\mathbb{P}(h(X)=y'|Y^*=y)}>1$, denote $\alpha:=\frac{\mathbb{P}(h(X)=y'|Y^*=-y)}{\mathbb{P}(h(X)=y'|Y^*=y)}>1$. We have

$$\begin{aligned}&1-\frac{1}{\mathbb{P}(Y^*=y)+\alpha\cdot\mathbb{P}(Y^*=-y)}\\=&\frac{(\alpha-1)\mathbb{P}(Y^*=-y)}{\mathbb{P}(Y^*=y)+\alpha\cdot\mathbb{P}(Y^*=-y)}\\=&\frac{\mathbb{P}(Y^*=-y)}{\mathbb{P}(Y^*=y)}\cdot\frac{\alpha-1}{\alpha+1}\\<&\frac{\mathbb{P}(Y^*=-y)}{\mathbb{P}(Y^*=y)}\\=&\frac{1}{\mathbb{P}(Y^*=y)}-1.\end{aligned}$$

Therefore we proved

$$\left|\frac{\mathbb{P}(h^*(X)=y,Y^*=y)}{\mathbb{P}(h^*(X)=y)\cdot\mathbb{P}(Y^*=y)}-1\right|>\left|\frac{\mathbb{P}(h(X)=y',Y^*=y)}{\mathbb{P}(h(X)=y')\cdot\mathbb{P}(Y^*=y)}-1\right|,\forall h\neq h^*$$

Because $f(v)$ is monotonically increasing in $|v-1|$, we proved that

$$\begin{aligned}&D_f(P_{h\times Y^*}\|Q_{h\times Y^*})\\=&\sum_{y,y'}\mathbb{P}(h(X)=y,Y^*=y')\cdot f\left(\frac{\mathbb{P}(h(X)=y,Y^*=y')}{\mathbb{P}(h(X)=y)\cdot\mathbb{P}(Y^*=y')}\right)\\<&\sum_{y,y'}\mathbb{P}(h(X)=y,Y^*=y')f\left(\frac{\mathbb{P}(h^*(X)=y',Y^*=y')}{\mathbb{P}(h^*(X)=y')\cdot\mathbb{P}(Y^*=y')}\right)\\=&\sum_{y}\mathbb{P}(Y^*=y)f\left(\frac{\mathbb{P}(h^*(X)=y,Y^*=y)}{\mathbb{P}(h^*(X)=y)\cdot\mathbb{P}(Y^*=y)}\right)\\=&D_f(P_{h^*\times Y^*}\|Q_{h^*\times Y^*})\end{aligned}$$

The last equality is because h^* always agrees with Y^*, so $\mathbb{P}(h^*(X)=y,Y^*=y)=\mathbb{P}(Y^*=y)$ and $\mathbb{P}(h^*(X)=-y,Y^*=y)=0$.

□

A.8 PROOF OF THEOREM 6: $\mathcal{H}$-ROBUST

Proof. The proofs for the multi-class case under uniform diagonal and sparse noise setting are entirely symmetrical due to Theorem 5 and 9. We deliver the main idea for the binary case.

The proof for condition (I) is easy to see:

$$
\begin{aligned}
&\operatorname*{argmax}_{h\in\mathcal{H}} D_f(\tilde{P}_{h\times\tilde{Y}}\|\tilde{Q}_{h\times\tilde{Y}}) \\
=&\operatorname*{argmax}_{h\in\mathcal{H}} \sup_g \ \mathbb{E}_{\tilde{Z}\sim\tilde{P}}\left[g(\tilde{Z})\right] - \mathbb{E}_{\tilde{Z}\sim\tilde{Q}}\left[f^*(g(\tilde{Z}))\right] \\
=&\operatorname*{argmax}_{h\in\mathcal{H}} \sup_g \ (1-e_+-e_-)\Big[\mathbb{E}_{Z\sim P}\left[g(Z)\right] - \mathbb{E}_{Z\sim Q}\left[f^*(g(Z))\right]\Big] + \mathsf{Bias}_f(h,g) \\
=&\operatorname*{argmax}_{h\in\mathcal{H}} \sup_g \ \mathbb{E}_{Z\sim P}\left[g(Z)\right] - \mathbb{E}_{Z\sim Q}[f^*(g(Z)] \\
=&\operatorname*{argmax}_{h\in\mathcal{H}} \ D_f(P_{h\times Y}\|Q_{h\times Y}) \\
=&h_f^*.
\end{aligned}
$$

Now we prove the robustness of D_f under condition (II). Denote by h' the classifier that maximizes $D_f(\tilde{P}_{h'\times\tilde{Y}}\|\tilde{Q}_{h'\times\tilde{Y}})$ and

$$
D_f(\tilde{P}_{h'\times\tilde{Y}}\|\tilde{Q}_{h'\times\tilde{Y}}) > D_f(\tilde{P}_{h_f^*\times\tilde{Y}}\|\tilde{Q}_{h_f^*\times\tilde{Y}})
$$

But

$$
\begin{aligned}
&D_f(\tilde{P}_{h'\times\tilde{Y}}\|\tilde{Q}_{h'\times\tilde{Y}}) \\
=&(1-e_+-e_-)\Big[\mathbb{E}_{Z\sim P}\left[\tilde{g}^*([h'(X),Y])\right] - \mathbb{E}_{Z\sim Q}\left[f^*(\tilde{g}^*([h'(X),Y])\right]\Big] + \mathsf{Bias}_f(h',\tilde{g}^*) \\
\le&\max_{h\in\mathcal{H}}(1-e_+-e_-)\cdot\sup_g\Big[\mathbb{E}_{Z\sim P}\left[g([h(X),Y])\right] - \mathbb{E}_{Z\sim Q}\left[f^*(g([h(X),Y])\right]\Big] + \mathsf{Bias}_f(h_f^*,g^*) \\
& \qquad\qquad (\mathsf{Bias}_f(h,\tilde{g}^*)\le\mathsf{Bias}_f(h_f^*,g^*)) \\
=&\max_{h\in\mathcal{H}}(1-e_+-e_-)\cdot D_f(P_{h\times Y}\|Q_{h\times Y}) + \mathsf{Bias}_f(h_f^*,g^*) \qquad \text{(variational form of } D_f) \\
=&(1-e_+-e_-)\cdot D_f(P_{h_f^*\times Y}\|Q_{h_f^*\times Y}) + \mathsf{Bias}_f(h_f^*,g^*) \\
\le&\sup_g\Big[\mathbb{E}_{Z=[h_f^*(X),Y]\sim P}\left[g(Z)\right] - \mathbb{E}_{Z=[h_f^*(X),Y]\sim Q}\left[f^*(g(Z))\right]\Big] + \mathsf{Bias}_f(h_f^*,g) \\
=&D_f(\tilde{P}_{h_f^*\times\tilde{Y}}\|\tilde{Q}_{h_f^*\times\tilde{Y}}),
\end{aligned}
$$

which is a contradiction. □

A.9 PROOF OF LEMMA 1: IMPACT OF BIAS TERM FOR DIFFERENT f-DIVERGENCES

Denote $p_y = \mathbb{P}(Y=y), \tilde{p}_y := \mathbb{P}(\tilde{Y}=y)$, $\dfrac{\tilde{P}(y,y')}{\tilde{Q}(y,y')} := \frac{\mathbb{P}(h(x)=y,\tilde{Y}=y')}{\mathbb{P}(h(x)=y)\cdot\mathbb{P}(\tilde{Y}=y')}$. We first prove that $\left|\dfrac{\tilde{P}(y,y')}{\tilde{Q}(y,y')}-1\right|$ approaches to 0 as a function of $1-e_+-e_-$:

Proposition 10. *When* $e_+, e_- < 0.5$, $\left|\dfrac{\tilde{P}(y,y')}{\tilde{Q}(y,y')}-1\right| = O\left((1-e_+-e_-)\right)$.

Proof.

$$\left|\frac{\tilde{P}(y,y')}{\tilde{Q}(y,y')}-1\right| = \left|\frac{\mathbb{P}(h(x)=y|\tilde{Y}=y')}{\mathbb{P}(h(x)=y)}-1\right|$$
$$= \left|\frac{\mathbb{P}(h(x)=y|\tilde{Y}=y') - \tilde{p}_{y'}\mathbb{P}(h(x)=y|\tilde{Y}=y') - (1-\tilde{p}_{y'})\mathbb{P}(h(x)=y|\tilde{Y}=-y')}{\mathbb{P}(h(x)=y)}\right|$$
$$= \frac{1-\tilde{p}_{y'}}{\mathbb{P}(h(x)=y)}\cdot|\mathbb{P}(h(x)=y|\tilde{Y}=y') - \mathbb{P}(h(x)=y|\tilde{Y}=-y')|$$

$|\mathbb{P}(h(x)=y|\tilde{Y}=y') - \mathbb{P}(h(x)=y|\tilde{Y}=-y')|$ derives as

$$\begin{aligned}
&|\mathbb{P}(h(x)=y|\tilde{Y}=y') - \mathbb{P}(h(x)=y|\tilde{Y}=-y')| \\
=&\Big|p_{y'}\mathbb{P}(h(x)=y|Y=y')\cdot\mathbb{P}(Y=y'|\tilde{Y}=y') + (1-p_{y'})\mathbb{P}(h(x)=y|Y=-y')\cdot\mathbb{P}(Y=-y'|\tilde{Y}=y') \\
&- p_{y'}\mathbb{P}(h(x)=y|Y=y)\cdot\mathbb{P}(Y=y'|\tilde{Y}=-y') \\
&- (1-p_{y'})\mathbb{P}(h(x)=y|Y=-y')\cdot\mathbb{P}(Y=-y'|\tilde{Y}=-y')\Big| \\
=&\Big|p_{y'}\mathbb{P}(h(x)=y|Y=y')\cdot(\mathbb{P}(Y=y'|\tilde{Y}=y') - \mathbb{P}(Y=y'|\tilde{Y}=-y')) \\
&+ (1-p_{y'})\mathbb{P}(h(x)=y|Y=-y')\cdot(\mathbb{P}(Y=-y'|\tilde{Y}=y') - \mathbb{P}(Y=-y'|\tilde{Y}=-y'))\Big| \\
=&\Big|p_{y'}\mathbb{P}(h(x)=y|Y=y') - (1-p_{y'})\mathbb{P}(h(x)=y|Y=-y')|\cdot|\mathbb{P}(Y=y'|\tilde{Y}=y') - \mathbb{P}(Y=y'|\tilde{Y}=-y')\Big|
\end{aligned}$$

The last equation is satisfied because $\forall y$:

$$\begin{aligned}
&\mathbb{P}(Y=y|\tilde{Y}=y) + \mathbb{P}(Y=-y|\tilde{Y}=y) = \mathbb{P}(Y=-y|\tilde{Y}=-y) + \mathbb{P}(Y=y|\tilde{Y}=-y) \\
\Longleftrightarrow&\mathbb{P}(Y=y|\tilde{Y}=y) - \mathbb{P}(Y=y|\tilde{Y}=-y) = \mathbb{P}(Y=-y|\tilde{Y}=-y) - \mathbb{P}(Y=-y|\tilde{Y}=y)
\end{aligned}$$

Now focus on $|\mathbb{P}(Y=y'|\tilde{Y}=y') - \mathbb{P}(Y=y'|\tilde{Y}=-y')|$:

$$\begin{aligned}
&|\mathbb{P}(Y=y'|\tilde{Y}=y') - \mathbb{P}(Y=y'|\tilde{Y}=-y')| \\
=&\left|\frac{\mathbb{P}(\tilde{Y}=y'|Y=y')\cdot p_{y'}}{\mathbb{P}(\tilde{Y}=y')} - \frac{\mathbb{P}(\tilde{Y}=-y'|Y=y')\cdot p_{y'}}{\mathbb{P}(\tilde{Y}=-y')}\right| \\
=&\frac{p_{y'}}{\tilde{p}_{y'}\cdot(1-\tilde{p}_{y'})}|(1-e_{y'})\cdot(1-\tilde{p}_{y'}) - e_{y'}\cdot\tilde{p}_{y'}| \\
=&\frac{p_{y'}}{\tilde{p}_{y'}\cdot(1-\tilde{p}_{y'})}\,|(1-e_{y'})\cdot(1-p_{y'}(1-e_{y'}) - (1-p_{y'})e_{-y'}) - e_{y'}\cdot(p_{y'}(1-e_{y'}) + (1-p_{y'})e_{-y'})| \\
=&\frac{p_{y'}}{\tilde{p}_{y'}\cdot(1-\tilde{p}_{y'})}|(1-p_{y'})(1-e_+-e_-)| \\
=&\frac{p_{y'}\cdot(1-p_{y'})}{\tilde{p}_{y'}\cdot(1-\tilde{p}_{y'})}|1-e_+-e_-|
\end{aligned}$$

Putting everything up together:

$$\left|\frac{\tilde{P}(y,y')}{Q(y,y')}-1\right| = p_{y'}(1-p_{y'})\left(p_{y'}\mathbb{P}(h(X)=y|Y=y') - (1-p_{y'})\mathbb{P}(h(X)=y|Y=-y')\right)\frac{|1-e_+-e_-|}{\tilde{p}_{y'}}$$

When $e_+, e_- < 0.5$, we have

$$\tilde{p}_{y'} = p_{y'}(1-e_{y'}) + p_{-y'}e_{-y'} \geq 0.5\min\{p, 1-p\}.$$

Therefore

$$\begin{aligned}
&p_{y'}(1-p_{y'})\left(p_{y'}\mathbb{P}(h(X)=y|Y=y') - (1-p_{y'})\mathbb{P}(h(X)=y|Y=-y')\right)\frac{|1-e_+-e_-|}{\tilde{p}_{y'}} \\
\leq&\, 2\max\{p,1-p\}(p_{y'}\mathbb{P}(h(X)=y|Y=y') - (1-p_{y'})\mathbb{P}(h(X)=y|Y=-y'))\cdot|1-e_+-e_-|
\end{aligned}$$

□

Next, we prove Lemma 1:

Proof. Shorthand $\tilde{P}(y,y') := \mathbb{P}(h(x)=y,\tilde{Y}=y'), \tilde{Q}(y,y') := \mathbb{P}(h(x)=y)\cdot\mathbb{P}(\tilde{Y}=y')$. Denote $x := \frac{\mathbb{P}(h(x)=y,\tilde{Y}=y')}{\mathbb{P}(h(x)=y)\cdot\mathbb{P}(\tilde{Y}=y')} - 1 = \frac{\tilde{P}(y,y')}{\tilde{Q}(y,y')} - 1$. Next we prove $\mathsf{Bias}_f(h,\tilde{g}^*) = O(x^2)$ for different f-divergences.

Because $\mathsf{Bias}_f(h,\tilde{g}^*) := \sum_{j=1}^K e_j \cdot \Delta_f^j(h,\tilde{g}^*) = \sum_{j=1}^K e_j \cdot \mathbb{E}_X[\tilde{g}^*(h(X),y) - f^*(\tilde{g}^*(h(X),y))]$, we analyze each of the term in expectation: $\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y'))$.

Jenson-Shannon For Jenson-Shannon divergence, we have:

$$\begin{aligned}
&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') + \log(2 - e^{\tilde{g}^*(y,y')}) \\
=&\log\frac{2\tilde{P}(y,y')}{\tilde{P}(y,y')+\tilde{Q}(y,y')} + \log\left(2 - e^{\log\frac{2\tilde{P}(y,y')}{\tilde{P}(y,y')+\tilde{Q}(y,y')}}\right) \\
=&\log\frac{2\tilde{P}(y,y')}{\tilde{P}(y,y')+\tilde{Q}(y,y')} + \log\frac{2\tilde{Q}(y,y')}{\tilde{P}(y,y')+\tilde{Q}(y,y')} \\
=&\log\frac{2}{1+\frac{\tilde{Q}(y,y')}{\tilde{P}(y,y')}} + \log\frac{2}{1+\frac{\tilde{P}(y,y')}{\tilde{Q}(y,y')}} \\
=&\log\frac{2}{1+\frac{1}{x+1}} + \log\frac{2}{1+x+1} \\
=&\log\frac{4}{2+x+1+\frac{1}{x+1}}
\end{aligned}$$

Using Taylor expansion we know

$$\log\frac{4}{2+x+1+\frac{1}{x+1}} = 0 + x\cdot\left(-\frac{4\cdot(1-\frac{1}{(x+1)^2})}{2+x+1+\frac{1}{x+1}}\right)\Bigg|_{x=0} + O(x^2) = O(x^2).$$

Squared Hellinger

$$\begin{aligned}
&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') - \frac{\tilde{g}^*(y,y')}{1-\tilde{g}^*(y,y')} = \frac{\tilde{g}^{*2}(y',y)}{1-\tilde{g}^*(y,y')} \\
=&\sqrt{\frac{\tilde{P}(h(x)=y,\tilde{Y}=y')}{\tilde{Q}(h(x)=y,\tilde{Y}=y')}} + \sqrt{\frac{\tilde{Q}(h(x)=y,\tilde{Y}=y')}{\tilde{P}(h(x)=y,\tilde{Y}=y')}} - 2 \\
=&\sqrt{1+x} + \sqrt{\frac{1}{1+x}} - 2.
\end{aligned}$$

$$\sqrt{1+x} + \sqrt{\frac{1}{1+x}} - 2 = \left[1 + \frac{1}{2\cdot\sqrt{1+x}} + 1 - \frac{1}{2}(1+x)^{-1.5}\right]\Bigg|_{x=0} - 2 + O(x^2) = O(x^2).$$

Pearson $\mathcal{X}^2$

$$\begin{aligned}
&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') - \tilde{g}^*(y,y') - \frac{1}{4}(\tilde{g}^*(y,y'))^2 \\
=&-\left(\frac{\tilde{P}(y,y')}{\tilde{Q}(y,y')} - 1\right)^2 = -x^2 = O(x^2).
\end{aligned}$$

Neyman $\mathcal{X}^2$

$$\begin{aligned}
&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') - 2 - 2\sqrt{1-\tilde{g}^*(y,y')} \\
=&-\left(\frac{\tilde{Q}(y,y')}{\tilde{P}(y,y')} - 1\right)^2 = -\left(\frac{1}{1+x} - 1\right)^2 = O(x^2).
\end{aligned}$$

KL

$$\begin{aligned}&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') - e^{\tilde{g}^*(y,y')-1}\\ =&1 + \log\frac{\tilde{P}(h(x)=y,\tilde{Y}=y')}{\tilde{Q}(h(x)=y,\tilde{Y}=y')} - \frac{\tilde{P}(h(x)=y,\tilde{Y}=y')}{\tilde{Q}(h(x)=y,\tilde{Y}=y')}\end{aligned}$$

$$1 + \log(1+x) - (1+x) = (\frac{1}{1+x} - 1)\Big|_{x=0} + O(x^2) = O(x^2).$$

Reverse KL

$$\begin{aligned}&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') + 1 + \log(-\tilde{g}^*(y,y'))\\ =&1 - \frac{\tilde{Q}(h(x)=y,\tilde{Y}=y')}{\tilde{P}(h(x)=y,\tilde{Y}=y')} + \log\frac{\tilde{Q}(h(x)=y,\tilde{Y}=y')}{\tilde{P}(h(x)=y,\tilde{Y}=y')}\end{aligned}$$

$$1 + \log(1+x)^{-1} - (1+x)^{-1} = 1 + \Big[-\frac{1}{1+x} - (1 - \frac{1}{(1+x)^2})\Big]\Big|_{x=0} + O(x^2) = O(x^2)$$

Jeffrey

$$1 - \tilde{g}^*(y,y') = \frac{\tilde{Q}(h(x)=y,\tilde{Y}=y')}{\tilde{P}(h(x)=y,\tilde{Y}=y')} - \log\frac{\tilde{P}(h(x)=y,\tilde{Y}=y')}{\tilde{Q}(h(x)=y,\tilde{Y}=y')} = \frac{1}{1+x} - \log(1+x)$$

And

$$\tilde{g}^*(y,y') = \frac{x}{1+x} + \log(1+x)$$

$$\begin{aligned}&\tilde{g}^*(y,y') - f^*(\tilde{g}^*(y,y')) = \tilde{g}^*(y,y') - W(e^{1-\tilde{g}^*(y,y')}) - \frac{1}{W(e^{1-\tilde{g}^*(y,y')})} - \tilde{g}^*(y,y') + 2\\ =&2 - W(e^{1-\tilde{g}^*(y,y')}) - \frac{1}{W(e^{1-\tilde{g}^*(y,y')})}\\ =&2 - W(e^{1-\tilde{g}^*(y,y')})\Big|_{x=0} - W'(e^{1-\tilde{g}^*(y,y')})\Big|_{x=0}\cdot x - \frac{1}{W(e^{1-\tilde{g}^*(y,y')})}\Big|_{x=0}\\ &- \Big[\frac{1}{W(e^{1-\tilde{g}^*(y,y')})}\Big]'\Big|_{x=0}\cdot x + O(x^2)\\ =&- W'(e^{1-\tilde{g}^*(y,y')})\Big|_{x=0}\cdot x - \Big[\frac{1}{W(e^{1-\tilde{g}^*(y,y')})}\Big]'\Big|_{x=0}\cdot x + O(x^2)\\ =&O(x^2)\end{aligned}$$

□

A.10 Proof of Theorem 7: $\mathcal{H}$-robustness of total-variation

Proof. We present the binary derivation but it extends easily to the multi-class case.

For TV, since $f(v) = \frac{1}{2}|v-1|$, $f^*(u) = u$, we immediately have $\forall y'\ g^*(h=y',y) - f^*(g(h=y',y)) = 0$ and therefore

$$\Delta_f^y(h,g) = \mathbb{E}_X[g(h(X),y)] - \mathbb{E}_X\left[f^*\left(g(h(X),y)\right)\right] \equiv 0, \forall y$$

and further for the binary case

$$\mathsf{Bias}_f(h,g) := e_+ \cdot \Delta_f^{-1}(h,g) + e_- \cdot \Delta_f^{+1}(h,g) = 0$$

and for the multi-class case

$$\mathsf{Bias}_f(h,g) = \sum_j e_j \Delta_f^j(h,g) = 0.$$

Therefore $\mathsf{Bias}_f(h, g) \equiv 0$. We then know TV is $\mathcal{H}$-robust for an arbitrary $\mathcal{H}$ using Theorem 6.

TV's robustness can also be derived straightforwardly for binary classification:

$$
\begin{aligned}
&\sup_g \mathbb{E}_{\tilde{Z}\sim\tilde{P}}\left[g(\tilde{Z})\right] - \mathbb{E}_{\tilde{Z}\sim\tilde{Q}}\left[f^*(g(\tilde{Z}))\right] \\
=& \sup_{|g|\le 1/2}\Big\{(1-e_+-e_-)\left[\mathbb{E}_{Z\sim P}\left[g(Z)\right] - \mathbb{E}_{Z\sim Q}\left[f^*(g(Z))\right]\right] \\
&+ \mathbb{E}\left[e_+\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)\right] \\
&- \mathbb{E}\left[e_+\cdot f^*\left(g(h(X),-1)\right) + e_-\cdot f^*\left(g(h(X),+1)\right)\right]\Big\} \\
=& \sup_{|g|\le 1/2}\Big\{(1-e_+-e_-)\left[\mathbb{E}_{Z\sim P}\left[g(Z)\right] - \mathbb{E}_{Z\sim Q}\left[f^*(g(Z))\right]\right] \\
&+ \mathbb{E}\left[e_+\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)\right] \\
&- \mathbb{E}\left[e_+\cdot g(h(X),-1) + e_-\cdot g(h(X),+1)\right]\Big\} \\
=& \sup_{|g|\le 1/2}(1-e_+-e_-)\left[\mathbb{E}_{Z\sim P}\left[g(Z)\right] - \mathbb{E}_{Z\sim Q}\left[f^*(g(Z))\right]\right] \\
=& (1-e_+-e_-)D_f(P_{h\times Y}\|Q_{h\times Y})
\end{aligned}
$$

That is for total variation, minimizing the f-divergence between $h(X)$ and $\tilde{Y}$ is the same as minimizing the f-divergence between $h(X)$ and the clean distribution Y. The above proof generalizes to multi-class easily. For example for the uniform diagonal noise, we have

$$
\Delta_f^j(h, g) = \sum_{j=1}^{K} e_j \cdot \left[\mathbb{E}_X[g(h(X), \tilde{Y} = j)] - \mathbb{E}_X\left[f^*(g(h(X), \tilde{Y} = j))\right]\right] = 0.
$$

Similar argument holds for sparse noise too.

□

A.11 When f-divergence measure is $\mathcal{H}^*$-robust?

Theorem 11. *For binary classification, suppose $\Delta_f^y(h, g)$ has the following form:*

$$
\Delta_f^y(h, \tilde{g}^*) = w_h \cdot t(\mathsf{FIT}(h = y, \tilde{Y} = y)) + (1 - w_h) \cdot t(\mathsf{FIT}(h = -y, \tilde{Y} = y)) \tag{8}
$$

$$
\Delta_f^y(h_f^*, g^*) = w_{h_f^*} \cdot t(\mathsf{FIT}(h_f^* = y, Y = y)) + (1 - w_{h_f^*}) \cdot t(\mathsf{FIT}(h_f^* = -y, Y = y)) \tag{9}
$$

where $w_h, w_{h_f^*} \in [0, 1]$. When $t(x)$ is monotonically decreasing as a function of $|x - 1|$ on both sides of $[1, \infty]$ and $(-\infty, 1)$, then $\mathsf{Bias}_f(h_f^*, g^*) \ge \max_{h\in\mathcal{H}^*} \mathsf{Bias}_f(h, \tilde{g}^*)$. Further, according to Theorem 6, the corresponding f-divergence measure is $\mathcal{H}^*$-robust.

Proof. For binary case, when $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1)$ and $e_+ = e_-$, we have $\mathbb{P}(\tilde{Y} = +1) = \mathbb{P}(\tilde{Y} = -1)$.

Denote $\mathcal{H}^*_- := \{h \in \mathcal{H} : \max_{y'} \mathsf{FIT}(h(X) = -y', \tilde{Y} = y') \leq \min_y \mathsf{FIT}(h^*_f = -y, Y = y)\}$, we first prove $\mathcal{H}^*_- \subseteq \mathcal{H}^*$. It is equivalent to prove, $\forall h \in \mathcal{H}^*_-$, $h \in \mathcal{H}^*$:

$$
\begin{aligned}
\mathcal{H}^*_- =& \{h \in \mathcal{H} : \max_{y'} \mathsf{FIT}(h(X) = -y', \tilde{Y} = y') \leq \min_y \mathsf{FIT}(h^*_f = -y, Y = y)\} \\
\subseteq& \Big\{h \in \mathcal{H} : \max_{y'} \frac{\mathbb{P}(h(X) = -y'|\tilde{Y} = y')}{\mathbb{P}(h(X) = -y')} \leq \min_y \frac{\mathbb{P}(h^*_f(X) = -y|Y = y)}{\mathbb{P}(h^*_f(X) = -y)}\Big\} \cup \{h^*_f\} \\
\subseteq& \Big\{h \in \mathcal{H} : \max_{y'} \frac{\mathbb{P}(h(X) = y'|\tilde{Y} = -y')}{\mathbb{P}(h(X) = y')} \leq \min_y \frac{\mathbb{P}(h^*_f(X) = y|Y = -y)}{\mathbb{P}(h^*_f(X) = y)}\Big\} \cup \{h^*_f\} \\
\subseteq& \Big\{h \in \mathcal{H} : \max_{y'} \frac{\mathbb{P}(h(X) = y'|\tilde{Y} = -y') - \mathbb{P}(h(X) = y'|\tilde{Y} = y')}{\mathbb{P}(h(X) = y')} \\
& \leq \min_y \frac{\mathbb{P}(h^*_f(X) = y|Y = -y) - \mathbb{P}(h^*_f(X) = y|Y = y)}{\mathbb{P}(h^*_f(X) = y)}\Big\} \cup \{h^*_f\} \\
\subseteq& \Big\{h \in \mathcal{H} : 1 - \min_{y'} \frac{\mathbb{P}(h(X) = y'|\tilde{Y} = y')}{\mathbb{P}(h(X) = y')} \leq 1 - \max_y \frac{\mathbb{P}(h^*_f(X) = y|Y = y)}{\mathbb{P}(h^*_f(X) = y)}\Big\} \cup \{h^*_f\} \\
\subseteq& \Big\{h \in \mathcal{H} : \min_{y'} \frac{\mathbb{P}(h(X) = y'|\tilde{Y} = y')}{\mathbb{P}(h(X) = y')} \geq \max_y \frac{\mathbb{P}(h^*_f(X) = y|Y = y)}{\mathbb{P}(h^*_f(X) = y)}\Big\} \cup \{h^*_f\} \\
\subseteq& \mathcal{H}^* = \Big\{h \in \mathcal{H} : \min_{y'} \mathsf{FIT}(h(X) = y', \tilde{Y} = y') \geq \max_y \mathsf{FIT}(h^*_f = y, Y = y)\Big\} \cup \{h^*_f\}
\end{aligned}
$$

Since $t(x)$ is monotonically decreasing as a function of $|x - 1|$, for $h \in \mathcal{H}^*$,

$$
\begin{aligned}
\mathsf{Bias}_f(h^*_f, g^*) =& e_+ \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot t(\mathsf{FIT}(h^*_f(X) = y, Y = +1)) \\
&+ e_- \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot t(\mathsf{FIT}(h^*_f(X) = y, Y = -1)) \\
=& e_+ \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot t(\mathsf{FIT}(h^*_f(X) = y, Y = y)) \cdot \mathbb{P}(Y = +1) \\
&+ e_+ \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot t(\mathsf{FIT}(h^*_f(X) = y, Y = -y)) \cdot \mathbb{P}(Y = -1) \\
&+ e_- \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot t(\mathsf{FIT}(h^*_f(X) = y, Y = y)) \cdot \mathbb{P}(Y = -1) \\
&+ e_- \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot t(\mathsf{FIT}(h^*_f(X) = y, Y = -y)) \cdot \mathbb{P}(Y = +1) \\
=& \frac{e_+ + e_-}{2} \cdot \sum_y \mathbb{P}(h^*_f(X) = y) \cdot \Big[t(\mathsf{FIT}(h^*_f(X) = y, Y = y)) + t(\mathsf{FIT}(h^*_f(X) = y, Y = -y))\Big] \\
\geq& \max_{h \in \mathcal{H}^*} \frac{e_+ + e_-}{2} \cdot \Big[t(\max_y \mathsf{FIT}(h^*_f(X) = y, Y = y)) + t(\min_y \mathsf{FIT}(h^*_f(X) = y, Y = -y))\Big]
\end{aligned}
$$

$$
\begin{aligned}
\max_{h \in \mathcal{H}^*} \mathsf{Bias}_f(h, \tilde{g}^*) =& \max_{h \in \mathcal{H}^*} e_+ \cdot \sum_y \mathbb{P}(h(X) = y) \cdot t(\mathsf{FIT}(h(X) = y, \tilde{Y} = +1)) \\
&+ e_- \cdot \sum_y \mathbb{P}(h(X) = y) \cdot t(\mathsf{FIT}(h(X) = y, \tilde{Y} = -1)) \\
=& \max_{h \in \mathcal{H}^*} [e_+ \cdot \mathbb{P}(\tilde{Y} = +1) + e_- \cdot \mathbb{P}(\tilde{Y} = -1)] \cdot \sum_{\tilde{y}} \mathbb{P}(h(X) = \tilde{y}) \cdot t(\mathsf{FIT}(h(X) = \tilde{y}, \tilde{Y} = \tilde{y})) \\
&+ [e_+ \cdot \mathbb{P}(\tilde{Y} = -1) + e_- \cdot \mathbb{P}(\tilde{Y} = +1)] \cdot \sum_{\tilde{y}} \mathbb{P}(h(X) = \tilde{y}) \cdot t(\mathsf{FIT}(h(X) = \tilde{y}, \tilde{Y} = -\tilde{y})) \\
\leq& \max_{h \in \mathcal{H}^*} \frac{e_+ + e_-}{2} \cdot \Big[t(\max_y \mathsf{FIT}(h(X) = y, \tilde{Y} = y)) + t(\min_y \mathsf{FIT}(h(X) = y, \tilde{Y} = -y))\Big]
\end{aligned}
$$

Thus, $\mathsf{Bias}_f(h_f^*, g^*) \geq \max_{h \in \mathcal{H}^*} \mathsf{Bias}_f(h, \tilde{g}^*)$. According to Theorem 6, the corresponding f-divergence measure is $\mathcal{H}^*$-robust.

□

A.12 Proof of Theorem 8: Robustness of D_f

Proof. Earlier we proved Theorem 11, next we show presented conditions in Eqn. (8) and (9) and $t(x)$ can be satisfied by the listed divergences:

The proof for $\Delta_f^y(h_f^*, g^*)$ can be viewed as a special case of $\Delta_f^y(h, \tilde{g}^*)$. The following derivations will therefore focus on $\Delta_f^y(h, \tilde{g}^*)$ and will not repeat for $\Delta_f^y(h_f^*, g^*)$.

Jenson-Shannon For Jenson-Shannon, we have $f^*(u) = -\log(2 - e^u)$, and

$$\begin{aligned}\tilde{g}^*(y, y') &= \log \frac{2 \cdot \mathbb{P}(h(X) = y, \tilde{Y} = y')}{\mathbb{P}(h(X) = y, \tilde{Y} = y') + \mathbb{Q}(h(X) = y, \tilde{Y} = y')} \\ &= \log \frac{2 \cdot \mathbb{P}(h(X) = y | \tilde{Y} = y')}{\mathbb{P}(h(X) = y | \tilde{Y} = y') + \mathbb{P}(h(X) = y)}\end{aligned}$$

Therefore,

$$\begin{aligned}\Delta_f^y(h, g) =& \mathbb{P}(h(X) = -y) \cdot \log \frac{4 \cdot \mathsf{FIT}(h(X) = -y, \tilde{Y} = y)}{(1 + \mathsf{FIT}(h(X) = -y, \tilde{Y} = y))^2} \\ &+ \mathbb{P}(h(X) = y) \cdot \log \frac{4 \cdot \mathsf{FIT}(h(X) = y, \tilde{Y} = y)}{(1 + \mathsf{FIT}(h(X) = y, \tilde{Y} = y))^2}\end{aligned}$$

$t(x) = \log \frac{4x}{(1+x)^2}$ satisfies the requirement specified in Theorem 11.

Squared-Hellinger For Squared-Hellinger, we have $f^*(u) = \frac{u}{1-u}$, and

$$\tilde{g}^*(y, y') = 1 - \sqrt{\frac{\mathbb{P}(h(X) = y)}{\mathbb{P}(h(X) = y | \tilde{Y} = y)}}$$

Therefore

$$\begin{aligned}\Delta_f^y(h, g) =& \mathbb{P}(h(X) = -y) \cdot \left[2 - \sqrt{\mathsf{FIT}(h(X) = -y, \tilde{Y} = y)} - \frac{1}{\sqrt{\mathsf{FIT}(h(X) = -y, \tilde{Y} = y)}}\right] \\ &+ \mathbb{P}(h(X) = y) \cdot \left[2 - \sqrt{\mathsf{FIT}(h(X) = y, \tilde{Y} = y)} - \frac{1}{\sqrt{\mathsf{FIT}(h(X) = y, \tilde{Y} = y)}}\right]\end{aligned}$$

Clearly $t(x) = 2 - \sqrt{x} - \frac{1}{\sqrt{x}}$ satisfies the requirement specified in Theorem 11.

Pearson $\mathcal{X}^2$

$$\begin{aligned}\Delta_f^y(h, \tilde{g}^*) &= \mathbb{E}[\tilde{g}^*(h(X), y) - f^*(\tilde{g}^*(h(X), y))] \\ &= \mathbb{E}[\tilde{g}^*(h(X), y) - \tilde{g}^*(h(X), y) - \frac{1}{4}(\tilde{g}^*(h(X), y))^2] \\ &= -\frac{1}{4}\mathbb{E}[(\tilde{g}^*(h(X), y))^2]\end{aligned}$$

$$\begin{aligned}\Delta_f^y(h,\tilde{g}^*) =& -\mathbb{P}(h(X)=y)\cdot\left(\frac{\mathbb{P}(h(X)=y,\tilde{Y}=y)}{\mathbb{P}(h(X)=y)\cdot\mathbb{P}(\tilde{Y}=y)}-1\right)^2\\ &-\mathbb{P}(h(X)=-y)\cdot\left(\frac{\mathbb{P}(h(X)=-y,\tilde{Y}=y)}{\mathbb{P}(h(X)=-y)\cdot\mathbb{P}(\tilde{Y}=y)}-1\right)^2\\ =& -\mathbb{P}(h(X)=y)\cdot(\mathsf{FIT}(h(X)=y,\tilde{Y}=y)-1)^2\\ &-\mathbb{P}(h(X)=-y)\cdot(\mathsf{FIT}(h(X)=-y,\tilde{Y}=y)-1)^2\end{aligned}$$

Correspondingly $t(x)=-(x-1)^2$, which satisfies the requirement specified in Theorem 11.

Neyman $\mathcal{X}^2$ For Neyman $\mathcal{X}^2$ we have $f^*(u)=2-2\sqrt{1-u}$, and

$$\begin{aligned}\Delta_f^y(h,\tilde{g}^*) &= \mathbb{E}[\tilde{g}^*(h(X),y)-f^*(\tilde{g}^*(h(X),y))]\\ &= \mathbb{E}[\tilde{g}^*(h(X),y)+2\sqrt{1-\tilde{g}^*(h(X),y)}-2]\end{aligned}$$

Since

$$\tilde{g}^*(y,y')=1-\left(\frac{Q(h(X)=y,\tilde{Y}=y')}{P(h(X)=y,\tilde{Y}=y')}\right)^2=1-\left(\frac{\mathbb{P}(h(X)=y)}{\mathbb{P}(h(X)=y|\tilde{Y}=y')}\right)^2$$

Therefore

$$\tilde{g}^*(h(X)=y,y')+2\sqrt{1-\tilde{g}^*(h(X)=y,y')}-2=-\left(\frac{\mathbb{P}(h(X)=y)}{\mathbb{P}(h(X)=y|\tilde{Y}=y')}-1\right)^2$$

And further we have

$$\begin{aligned}&\Delta_f^y(h,\tilde{g}^*)\\ =& -\mathbb{P}(h(X)=y)\cdot\left(\frac{\mathbb{P}(h(X)=y)}{\mathbb{P}(h(X)=y|\tilde{Y}=y)}-1\right)^2\\ &-\mathbb{P}(h(X)=-y)\cdot\left(\frac{\mathbb{P}(h(X)=-y)}{\mathbb{P}(h(X)=-y|\tilde{Y}=y)}-1\right)^2\\ =& -\mathbb{P}(h(X)=y)\cdot(\mathsf{FIT}(h(X)=y,\tilde{Y}=y)^{-1}-1)^2\\ &-\mathbb{P}(h(X)=-y)\cdot(\mathsf{FIT}(h(X)=-y,\tilde{Y}=y)^{-1}-1)^2\end{aligned}$$

$t(x)=-(x^{-1}-1)^2$ satisfies the requirement specified in Theorem 11.

KL For KL, we have $f^*(u)=e^{u-1}$, and

$$\tilde{g}^*(y,y')=1+\log\frac{\mathbb{P}(h(X)=y|\tilde{Y}=y')}{\mathbb{P}(h(X)=y')}$$

Therefore

$$\begin{aligned}\Delta_f^y(h,g) =& \mathbb{P}(h(X)=-y)\cdot[1+\log\mathsf{FIT}(h(X)=-y,\tilde{Y}=y)-\mathsf{FIT}(h(X)=-y,\tilde{Y}=y)]\\ &+\mathbb{P}(h(X)=y)\cdot[1+\log\mathsf{FIT}(h(X)=y,\tilde{Y}=y)-\mathsf{FIT}(h(X)=y,\tilde{Y}=y)]\end{aligned}$$

Clearly $t(x)=1+\log x-x$ satisfies the requirement specified in Theorem 11.

Reverse-KL For Reverse-KL, we have $f^*(u) = -1 - \log(-u)$, and

$$\tilde{g}^*(y, y') = -\log \frac{\mathbb{P}(h(X) = y)}{\mathbb{P}(h(X) = y|\tilde{Y} = y)}$$

Therefore

$$\begin{aligned}\Delta_f^y(h, g) =& \mathbb{P}(h(X) = -y) \cdot \left[1 - \log \mathsf{FIT}(h(X) = -y, \tilde{Y} = y) - \frac{1}{\mathsf{FIT}(h(X) = -y, \tilde{Y} = y)}\right] \\ &+ \mathbb{P}(h(X) = y) \cdot \left[1 - \log \mathsf{FIT}(h(X) = y, \tilde{Y} = y) - \frac{1}{\mathsf{FIT}(h(X) = y, \tilde{Y} = y)}\right]\end{aligned}$$

Clearly $t(x) = 1 - \log x - \frac{1}{x}$ satisfies the requirement specified in Theorem 11.

□

B SUPPLEMENTARY EXPERIMENT RESULTS

In our experiment settings, D_f measures fail to work well on almost all sparse high noise setting. This is largely due to the super unbalanced noisy labels, e.g., for each pair, the ratio of samples between the two classes is in the range of $[\frac{1}{3}, \frac{1}{2}]$.

B.1 SUPPLEMENTARY TABLE OF TABLE 3: METHODS COMPARISON WITHOUT BIAS CORRECTION

In Table 7, the performance of Pearson $\mathcal{X}^2$, Jeffrey divergence on MNIST, Fashion MNIST, CIFAR-10 and CIFAR-100 are included in Table 7.

Dataset	Noise	CE	BLC	FLC	DMI	PL	**Pearson**	**Jeffrey**
MNIST	Sparse, Low	97.21	95.23	97.37	97.76	98.59	**99.24(99.07±0.16)**	**99.24(99.11±0.08)**
	Sparse, High	48.55	55.86	49.67	49.61	**60.27**	58.63(58.58±0.05)	49.21(49.17±0.04)
	Uniform, Low	97.14	94.27	95.51	97.72	99.06	99.13(99.03±0.09)	**99.14(99.06±0.05)**
	Uniform, High	93.25	85.92	87.75	95.50	97.77	97.89(97.76±0.10)	**97.94(97.80±0.12)**
	Random (0.2)	98.26	97.46	97.61	98.82	99.25	**99.28(99.27±0.02)**	99.29(99.22±0.11)
	Random (0.7)	97.00	93.52	87.74	95.47	98.52	**98.70(98.54±0.10)**	**98.67(98.53±0.15)**
Fashion MNIST	Sparse, Low	84.36	86.02	88.15	85.65	88.32	**88.93(88.81±0.11)**	**88.96(88.68±0.21)**
	Sparse, High	43.33	46.97	47.63	47.16	**51.92**	44.62(44.36±0.21)	45.57(45.26± 0.28)
	Uniform, Low	82.98	84.48	86.58	83.69	**89.31**	87.27(87.15±0.09)	88.13(87.85±0.18)
	Uniform, High	79.52	78.10	82.41	77.94	84.69	**85.30(85.26±0.06)**	84.92(84.63±0.26)
	Random (0.2)	85.47	83.40	77.61	86.21	**89.78**	89.65(89.44±0.21)	89.74(89.33+0.29)
	Random (0.7)	82.05	78.41	73.42	80.89	**87.22**	86.72(86.29±0.32)	87.21(87.19±0.04)
CIFAR-10	Sparse, Low	87.20	72.96	76.17	**92.32**	91.35	91.40(91.24±0.27)	91.55(91.24±0.15)
	Sparse, High	61.81	56.30	66.12	27.94	**69.70**	46.36(46.27±0.07)	46.21(45.78±0.27)
	Uniform, Low	85.68	72.73	77.12	90.39	91.70	**92.37(92.27±0.07)**	**92.17(92.02±0.08)**
	Uniform, High	71.38	54.41	64.22	82.68	83.42	83.61(83.09±0.38)	**83.80(83.73±0.05)**
	Random (0.5)	78.40	59.31	68.97	85.06	86.47	86.03(85.56±0.32)	86.04(85.75±0.19)
	Random (0.7)	68.26	38.59	54.39	77.91	57.81	76.92 (76.82±0.07)	**79.46(79.08±0.25)**
CIFAR-100	Uniform	63.87	51.40	60.04	64.39	67.94	**68.42(68.16±0.16)**	**68.86(68.63±0.18)**
	Sparse	40.45	36.57	43.39	40.53	**44.25**	37.54(37.50±0.07)	37.43(37.06±0.25)
	Random (0.2)	65.84	61.21	61.52	66.23	62.92	**69.90(69.72±0.15)**	**69.56(69.42±0.14)**
	Random (0.5)	56.92	22.21	55.88	56.06	49.62	**60.81(60.36±0.26)**	**60.95(60.73±0.15)**

Table 7: Experiment results comparison (w/o bias correction): The best performance in each setting (row) is highlighted in **blue**. All f-divergences will be highlighted if they are better than the baselines we compare to. We report the maximum accuracy of each D_f measures along with (mean $\pm$ standard deviation).

B.2 D_f MEASURES WITH BIAS CORRECTION ON MNIST

In Table 8, we test the impact of bias correction on MNIST with 4 noise settings. Except for the spare high noise setting which is a huge challenge for all implemented methods, experiment results of other 3 noise settings further demonstrate the negligible effect of bias term in the optimization of D_f measures.

Noise	J-S	Gap	PS	Gap	KL	Gap	Jeffrey	Gap
Sparse, Low	**98.88(98.82±0.06)**	-0.27	**99.05(98.98±0.05)**	-0.19	**99.29(99.19±0.09)**	**+0.08**	**99.13(99.06±0.06)**	-0.11
Sparse, High	21.39(21.36±0.03)	-37.54	49.22(49.17±0.05)	-9.41	49.07(49.05±0.02)	-0.07	49.14(49.06±0.09)	-0.07
Uniform, Low	**99.18(99.10±0.05)**	**+0.05**	99.13(99.01±0.10)	**+0.00**	**99.30(99.24±0.08)**	**+0.20**	**99.20(99.12±0.09)**	**+0.06**
Uniform, High	97.76(97.68±0.07)	-0.10	97.72(97.65±0.06)	-0.17	97.91(97.74±0.14)	-0.23	**98.16(97.98±0.14)**	**+0.22**

Table 8: D_f measures with bias correction on MNIST: Digits highlighted in **blue** means better than baseline methods, in **red** means better than without bias correction. PS: Pearson.

B.3 D_f MEASURES WITH BIAS CORRECTION ON FASHION MNIST

In Table 9, we test the impact of bias correction on Fashion MNIST with 4 noise settings. We can reach the same conclusion on bias correction as MNIST.

Noise	J-S	Gap	PS	Gap	KL	Gap	Jeffrey	Gap
Sparse, Low	**89.37(88.83±0.34)**	**+0.57**	87.99(87.90±0.13)	-0.94	82.29(82.03±0.22)	-7.48	82.04(81.65±0.26)	-6.92
Sparse, High	✗	✗	39.02(38.85±0.10)	-5.60	46.98(46.23±0.62)	**+8.02**	38.94(38.75±0.15)	-6.63
Uniform, Low	88.98(88.61±0.26)	**+0.40**	87.72(87.66±0.06)	**+0.45**	89.04(88.78±0.18)	**+0.72**	89.05(88.87±0.15)	**+0.92**
Uniform, High	**85.56(85.33± 0.19)**	-0.06	**85.57(85.03±0.37)**	**+0.27**	**85.15(84.94±0.15)**	-0.54	84.76(84.48±0.31)	-0.16

Table 9: D_f measures with bias correction on Fashion MNIST: Digits highlighted in **blue** means better than baseline methods, in **red** means better than without bias correction. $\times$: experiment failed to stablize. PS: Pearson.

C EXPERIMENT DETAILS

C.1 NOISE TRANSITION MATRIX FOR SECTION 5.2: ROBUSTNESS OF D_f MEASURES

We use CIFAR-10 dataset together with the uniform noise transition matrix to flip the noisy labels. In the following noise transition matrix, e is in $[0.00, 0.01, 0.02, ..., 0.09]$ and the noise rate of each set of noisy labels is $9 * e$.

$$\begin{bmatrix}
1-9*e & e & e & e & e & e & e & e & e & e \\
e & 1-9*e & e & e & e & e & e & e & e & e \\
e & e & 1-9*e & e & e & e & e & e & e & e \\
e & e & e & 1-9*e & e & e & e & e & e & e \\
e & e & e & e & 1-9*e & e & e & e & e & e \\
e & e & e & e & e & 1-9*e & e & e & e & e \\
e & e & e & e & e & e & 1-9*e & e & e & e \\
e & e & e & e & e & e & e & 1-9*e & e & e \\
e & e & e & e & e & e & e & e & 1-9*e & e \\
e & e & e & e & e & e & e & e & e & 1-9*e
\end{bmatrix}$$

C.2 NOISE TRANSITION MATRIX FOR MNIST AND FASHION MNIST DATASET

Sparse-low noise matrix:

$$\begin{bmatrix}
0.7 & 0.3 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0.2 & 0.8 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.7 & 0.3 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.2 & 0.8 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.7 & 0.3 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.2 & 0.8 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.7 & 0.3 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.2 & 0.8 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.7 & 0.3 \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.2 & 0.8
\end{bmatrix}$$

Sparse-high noise matrix:

$$\begin{bmatrix}
0.3 & 0.7 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0.2 & 0.8 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.3 & 0.7 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.2 & 0.8 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.3 & 0.7 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.2 & 0.8 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.3 & 0.7 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.2 & 0.8 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.3 & 0.7 \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.2 & 0.8
\end{bmatrix}$$

Uniform-low noise matrix:

$$\begin{bmatrix}
0.258 & 0.075 & 0.09 & 0.085 & 0.07 & 0.082 & 0.077 & 0.091 & 0.092 & 0.08 \\
0.08 & 0.253 & 0.09 & 0.085 & 0.07 & 0.082 & 0.077 & 0.091 & 0.092 & 0.08 \\
0.08 & 0.075 & 0.268 & 0.085 & 0.07 & 0.082 & 0.077 & 0.091 & 0.092 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.263 & 0.07 & 0.082 & 0.077 & 0.091 & 0.092 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.085 & 0.248 & 0.082 & 0.77 & 0.91 & 0.92 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.085 & 0.07 & 0.26 & 0.077 & 0.091 & 0.092 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.085 & 0.07 & 0.082 & 0.255 & 0.091 & 0.092 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.085 & 0.07 & 0.082 & 0.077 & 0.269 & 0.092 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.085 & 0.07 & 0.082 & 0.077 & 0.091 & 0.27 & 0.08 \\
0.08 & 0.075 & 0.09 & 0.085 & 0.07 & 0.082 & 0.077 & 0.091 & 0.092 & 0.258
\end{bmatrix}$$

Uniform-high noise matrix:

$$\begin{bmatrix}
0.58 & 0.045 & 0.047 & 0.055 & 0.053 & 0.022 & 0.068 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.0575 & 0.047 & 0.055 & 0.053 & 0.022 & 0.068 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.577 & 0.055 & 0.053 & 0.022 & 0.068 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.585 & 0.053 & 0.022 & 0.068 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.055 & 0.583 & 0.022 & 0.068 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.055 & 0.053 & 0.552 & 0.068 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.055 & 0.053 & 0.022 & 0.598 & 0.054 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.055 & 0.053 & 0.022 & 0.068 & 0.584 & 0.056 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.055 & 0.053 & 0.022 & 0.068 & 0.054 & 0.586 & 0.02 \\
0.05 & 0.045 & 0.047 & 0.055 & 0.053 & 0.022 & 0.068 & 0.054 & 0.056 & 0.55
\end{bmatrix}$$

Random 0.2 noise matrix:

$$\begin{bmatrix}
0.82 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.82 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.82 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.82 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.02 & 0.82 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.82 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.82 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.82 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.82 & 0.02 \\
0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.82
\end{bmatrix}$$

Random 0.7 noise matrix:

$$\begin{bmatrix}
0.36 & 0.07 & 0.08 & 0.07 & 0.08 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.06 & 0.39 & 0.07 & 0.06 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.07 & 0.38 & 0.08 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.08 & 0.07 & 0.07 & 0.36 & 0.07 & 0.08 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.07 & 0.07 & 0.07 & 0.37 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.07 & 0.07 & 0.08 & 0.06 & 0.37 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.38 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.06 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.38 & 0.07 & 0.07 \\
0.06 & 0.07 & 0.07 & 0.07 & 0.08 & 0.07 & 0.07 & 0.07 & 0.37 & 0.07 \\
0.07 & 0.07 & 0.06 & 0.07 & 0.07 & 0.07 & 0.08 & 0.07 & 0.07 & 0.37
\end{bmatrix}$$

C.3 NOISE TRANSITION MATRIX FOR CIFAR-10 DATASET

Sparse-low noise matrix:

$$\begin{bmatrix}
0.7 & 0.3 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0.1 & 0.9 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.7 & 0.3 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.1 & 0.9 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.7 & 0.3 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.1 & 0.9 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.7 & 0.3 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.1 & 0.9 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.7 & 0.3 \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.1 & 0.9
\end{bmatrix}$$

Sparse-high noise matrix:

$$\begin{bmatrix}
0.4 & 0.6 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0.2 & 0.8 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.4 & 0.6 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0.2 & 0.8 & 0. & 0. & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.4 & 0.6 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0.2 & 0.8 & 0. & 0. & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.4 & 0.6 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0.2 & 0.8 & 0. & 0. \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.4 & 0.6 \\
0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.2 & 0.8
\end{bmatrix}$$

Uniform-low noise matrix:

$$\begin{bmatrix}
0.82 & 0.03 & 0.01 & 0.023 & 0.017 & 0.022 & 0.021 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.83 & 0.01 & 0.023 & 0.017 & 0.022 & 0.021 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.81 & 0.023 & 0.017 & 0.022 & 0.021 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.823 & 0.017 & 0.022 & 0.021 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.023 & 0.817 & 0.022 & 0.021 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.023 & 0.017 & 0.822 & 0.021 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.023 & 0.017 & 0.022 & 0.821 & 0.018 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.023 & 0.017 & 0.022 & 0.021 & 0.818 & 0.019 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.023 & 0.017 & 0.022 & 0.021 & 0.018 & 0.819 & 0.02 \\
0.02 & 0.03 & 0.01 & 0.023 & 0.017 & 0.022 & 0.021 & 0.018 & 0.019 & 0.82
\end{bmatrix}$$

Uniform-high noise matrix:

$$\begin{bmatrix}
0.46 & 0.07 & 0.04 & 0.05 & 0.06 & 0.04 & 0.06 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.48 & 0.04 & 0.05 & 0.06 & 0.04 & 0.06 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.45 & 0.05 & 0.06 & 0.04 & 0.06 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.46 & 0.06 & 0.04 & 0.06 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.05 & 0.47 & 0.04 & 0.06 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.05 & 0.06 & 0.45 & 0.06 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.05 & 0.06 & 0.04 & 0.47 & 0.07 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.05 & 0.06 & 0.04 & 0.06 & 0.48 & 0.08 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.05 & 0.06 & 0.04 & 0.06 & 0.07 & 0.49 & 0.07 \\
0.05 & 0.07 & 0.04 & 0.05 & 0.06 & 0.04 & 0.06 & 0.07 & 0.08 & 0.48
\end{bmatrix}$$

Random 0.5 noise matrix:

$$\begin{bmatrix}
0.55 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.56 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.55 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.06 & 0.54 & 0.05 & 0.05 & 0.05 & 0.04 & 0.06 & 0.06 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.56 & 0.05 & 0.05 & 0.05 & 0.04 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.54 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.04 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.55 & 0.05 & 0.05 & 0.05 \\
0.04 & 0.04 & 0.05 & 0.05 & 0.06 & 0.05 & 0.04 & 0.56 & 0.05 & 0.05 \\
0.06 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.55 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.55
\end{bmatrix}$$

Random 0.7 noise matrix:

$$\begin{bmatrix}
0.37 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.06 & 0.38 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.08 & 0.07 \\
0.07 & 0.07 & 0.36 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.08 \\
0.07 & 0.07 & 0.07 & 0.37 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.07 & 0.08 & 0.07 & 0.37 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.08 & 0.07 & 0.07 & 0.07 & 0.36 & 0.07 & 0.07 & 0.07 & 0.06 \\
0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.37 & 0.07 & 0.07 & 0.07 \\
0.07 & 0.06 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.37 & 0.07 & 0.07 \\
0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.38 & 0.06 \\
0.07 & 0.07 & 0.07 & 0.08 & 0.07 & 0.07 & 0.07 & 0.07 & 0.06 & 0.37
\end{bmatrix}$$

C.4 NOISE TRANSITION MATRIX FOR CIFAR-100 DATASET

For sparse noise matrix, we randomly divide 100 classes into 50 disjoint pairs, the flipping probability (T_{ji}, T_{ij}) in each pair is randomly chosen from $(0.05, 0.75)$, $(0.1, 0.70)$, $(0.15, 0.65)$, $(0.2, 0.6)$.

C.5 PARAMETER SETTINGS ON NOISED DATASET

MNIST, Fashion MNIST, CIFAR-10 For experiments on MNIST and Fashion-MNIST datasets, we use the convolutional neural network used in DMI for DMI, PL and f-divergences. All the experiments are performed with batch size 128. PL and f-divergences adopt two kinds of learning rate setting and trained for 80 epochs, either with initial learning rate 5e-4 or 1e-3, then decay 0.2, 0.5, 0.2 every 20 epochs. We choose the default learning rate setting for DMI, BLC and FLC. For DMI's convolutional neural network, Adam(Kingma & Ba (2014)) with default parameters is used as the optimizer, while for loss-correction's fully-connected neural network case we use AdaGrad(Duchi et al. (2010)) in order to be consistent with their works.

CIFAR-10 and CIFAR-100 For all methods and both datasets, we unify the model to be an 18-layer PreAct Resnet (He et al. (2016)) and train it using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. All methods firstly train with CE warm-up for 120 (CIFAR-10) or 240 (CIFAR-100) epochs on CIFAR-10 and CIFAR-100 respectively. For DMI, BLC and FLC, we use the default learning rate settings. For PL and f-divergences, we train 100 epochs after the warm-up with initial learning rate 0.01, and decays 0.1 every 30 epochs.

Clothing 1M For clothing 1M, we use pre-trained ResNet50, SGD optimizer with momentum 0.9 and weight decay 1e-3. The initial learning rate is 0.002. All mentioned f-divergences trained 40 epochs, after 10 epochs, the learning rate becomes 5e-5. Then it decays 0.2, 0.5 consequently for every 5 epochs. We compare with reported best result for all our baseline methods.

C.6 PARAMETER SETTINGS ON CLEAN DATASET

We adopt the same setting (except for the number of epochs and the learning rate setting) as used in the noised dataset for each dataset.

MNIST, Fashion MNIST For CE, we trained the model for 40 epochs. The initial learning rate is 5e-4, and it decays 0.2 after 20 epochs. For D_f measures, the learning rate setting is the same as that in the noised dataset.

CIFAR-10 For CE, we trained the model for 300 epochs. Learning rate is 0.1 for first 150 epochs. From 150-th epoch to 250-th epoch, the learning rate is 0.01. Then, 0.001 till the end. For D_f measures, we trained the model for 240 epochs. The initial learning rate is 0.1, and it decays 0.1 for every 60 epochs.

CIFAR-100 For CE, we trained the model for 200 epochs. Learning rate is 0.1 for first 60 epochs. From 61-th epoch to 120-th epoch, the learning rate is 0.02 (save the model at 120-th epoch as a warm-up model for D_f measures). From 121-th epoch to 160-th epoch, the learning rate is 0.004. Then, 0.0008 till the end. For D_f measures, we load pre-trained CE model and trained for another 100 epochs. The initial learning rate is 0.01, and it decays 0.1 for every 30 epochs.

C.7 COMPUTING INFRASTRUCTURE

In our experiments, we use a GPU cluster (8 TITAN V GPUs and 16 GeForce GTX 1080 GPUs) for training and evaluation.