Appendix

A. Proof

We give the proofs of the theorems proposed in the paper.

A.1. Proof of Theorem 1

When $\alpha = 0$ and $\beta \geq 0$, the optimization problem can be formalized with constraints as

$$\begin{aligned} &\min_{\mathbb{F}} \mathcal{L}_{\text{ECE}} - \beta \cdot \log\left[\det(\tilde{M}_{\setminus y}^{\top}\tilde{M}_{\setminus y})\right] \\ &\text{s.t. } 0 \leq F_j^k \leq 1, \\ &\quad \sum_{j\in[L]} F_j^k = 1, \end{aligned}$$

where $k \in [K],\ j \in [L]$. Note that in the objective function, the first term $\mathcal{L}_{\text{ECE}}$ depends on the predictions on label y, i.e., F_y^k. The second term $-\log\left[\det(\tilde{M}_{\setminus y}^{\top}\tilde{M}_{\setminus y})\right]$ depends on the normalized non-maximal predictions $\tilde{F}_{\setminus y}^k$. Because $\forall i, j$, F_y^i and F_y^j are mutually independent, F_y^i and $\tilde{F}_{\setminus y}^i$ are also mutually independent (since the normalization), the two terms in the objective function can separately achieve their own minimum. Therefore, the optimal solution of the objective function will tend to satisfies the equations $F^k = 1_y$, where $k \in [K]$.

□

A.2. Proof of Theorem 2

When $\alpha > 0$ and $\beta = 0$, the optimization problem can be formalized with constraints as

$$\begin{aligned} &\min_{\mathbb{F}} \mathcal{L}_{\text{ECE}} - \alpha \cdot \mathcal{H}(\mathcal{F}) \\ &\text{s.t. } 0 \leq F_j^k \leq 1, \\ &\quad \sum_{j\in[L]} F_j^k = 1, \end{aligned}$$

where $k \in [K],\ j \in [L]$. Then the Lagrangian is

$$\begin{aligned} L = \mathcal{L}_{\text{ECE}} - \alpha \cdot \mathcal{H}(\mathcal{F}) + \sum_{k\in[K]} \omega_k (1 - \sum_{j\in[L]} F_j^k) \\ + \sum_{k\in[K]} \sum_{j\in[L]} \left[\beta_{k,j} F_j^k + \gamma_{k,j}(1 - F_j^k)\right], \end{aligned}$$

where $\beta_{k,j} \leq 0$, $\gamma_{k,j} \leq 0$. The partial derivatives for F_j^k are

$$\begin{aligned} \frac{\partial L}{\partial F_y^k} &= -\frac{1}{F_y^k} + \frac{\alpha}{K}\left[1 + \log \mathcal{F}_y\right] - \omega_k + \beta_{k,y} - \gamma_{k,y}, \\ \frac{\partial L}{\partial F_j^k} &= \frac{\alpha}{K}\left[1 + \log \mathcal{F}_j\right] - \omega_k + \beta_{k,j} - \gamma_{k,j}, \forall j \neq y. \end{aligned}$$

According to the KKT conditions for the optimal solution, we have $\forall k \in [K]$, $j \in [L]$,

$$\begin{aligned} &\frac{\partial L}{\partial F_j^k} = 0, \\ &\beta_{k,j} F_j^k = 0, \\ &\gamma_{k,j}(1 - F_j^k) = 0. \end{aligned}$$

Consider the optimal solutions in $(0, 1)^{L\times K}$, then all $\beta_{k,j}$ and $\gamma_{k,j}$ equal to zero. Now we have

$$\begin{aligned} &-\frac{1}{F_y^k} + \frac{\alpha}{K}\left[1 + \log \mathcal{F}_y\right] = \omega_k, \\ &\frac{\alpha}{K}\left[1 + \log \mathcal{F}_j\right] = \omega_k, \forall j \neq y, \end{aligned}$$

and from the second equations, we can derive $\forall j \neq y$,

$$\begin{aligned} \mathcal{F}_j = \exp\left(\frac{\omega_k K}{\alpha} - 1\right) &\Longrightarrow \sum_{j\neq y} \mathcal{F}_j = \sum_{j\neq y} \exp\left(\frac{\omega_k K}{\alpha} - 1\right) \\ &\Longrightarrow 1 - \mathcal{F}_y = (L-1)\exp\left(\frac{\omega_k K}{\alpha} - 1\right) \\ &\Longrightarrow \omega_k = \frac{\alpha}{K}\left[1 + \log\left(\frac{1 - \mathcal{F}_y}{L-1}\right)\right], \end{aligned}$$

and this also shows that $\forall i, j \neq y$, there is $\mathcal{F}_i = \mathcal{F}_j = \frac{1-\mathcal{F}_y}{L-1}$. There further is $\forall k \in [K]$,

$$\frac{1}{F_y^k} = \frac{\alpha}{K} \log\left[\frac{\mathcal{F}_y(L-1)}{1 - \mathcal{F}_y}\right].$$

Thus for $\forall k, l \in [K]$, $F_y^k = F_y^l = \mathcal{F}_y$, and finally

$$\frac{1}{\mathcal{F}_y} = \frac{\alpha}{K} \log\left[\frac{\mathcal{F}_y(L-1)}{1 - \mathcal{F}_y}\right].$$

□

A.3. Proof of Corollary 1

It is easy to see that the negative LED part $-\log(\mathbb{ED})$ achieves its minimum if and only if the non-maximal predictions of each individual network are mutually orthogonal. According to the conclusion in Theorem 2, if there is $K \mid (L-1)$, the optimal solution in Corollary 1 can simultaneously make the two terms of the ADP regularizer achieve their own minimum.

□

B. More Analyses

In this section, we provide more details on the theoretical and practical analyses mentioned in the paper.

B.1. JS-divergence as the Diversity

JS-divergence among individual predictions is a potentially plausible definition of ensemble diversity for DNNs. Specifically, we consider in the output space of classifiers, where F^k represents a vector variable in $\mathbb{R}^L$. The JS-divergence of K elements in $\mathbb{F}$ is defined as

$$\text{JSD}(\mathbb{F}) = \mathcal{H}(\mathcal{F}) - \frac{1}{K}\sum_{k\in[K]}\mathcal{H}(F^k). \tag{1}$$

To encourage high values of JS-divergence , we can add a regularization term of it in the objective function. However, when minimizing the objective function, there is neither a closed form solution nor an intuitively reasonable solution, as formally stated in the following theorem:

Theorem* 1. *Given $\lambda > 0$, (x, y) be an input-label pair. The minimization problem is defined as*

$$\min_{\mathbb{F}} \mathcal{L}_{ECE} - \lambda \cdot JSD. \tag{2}$$

Then the problem has no solution in $(0,1)^{L\times K}$.

Proof. The optimization problem can be formalized with constraints as

$$\begin{aligned}&\min_{\mathbb{F}} \mathcal{L}_{\text{ECE}} - \lambda \cdot \text{JSD}(\mathbb{F})\\ &\text{s. t. } 0 \le F_j^k \le 1,\\ &\quad \sum_{j\in[L]} F_j^k = 1,\end{aligned}$$

where $k \in [K],\ j \in [L]$. Then the Lagrangian is

$$\begin{aligned}L = \mathcal{L}_{\text{ECE}} - \lambda \cdot \text{JSD}(\mathbb{F}) + \sum_{k\in[K]}\omega_k(1 - \sum_{j\in[L]} F_j^k)\\ + \sum_{k\in[K]}\sum_{j\in[L]}\left[\beta_{k,j}F_j^k + \gamma_{k,j}(1-F_j^k)\right],\end{aligned}$$

where $\beta_{k,j} \le 0$, $\gamma_{k,j} \le 0$. The partial derivatives for F_j^k are

$$\begin{aligned}\frac{\partial L}{\partial F_y^k} &= -\frac{1}{F_y^k} + \frac{\lambda}{K}\left[\log\mathcal{F}_y - \log F_y^k\right] - \omega_k + \beta_{k,y} - \gamma_{k,y},\\ \frac{\partial L}{\partial F_j^k} &= \frac{\lambda}{K}\left[\log\mathcal{F}_j - \log F_j^k\right] - \omega_k + \beta_{k,j} - \gamma_{k,j}, \forall j \ne y.\end{aligned}$$

According to the KKT conditions for the optimal solution, similar to the proof of Theorem 2, we can derive $\forall j \ne y$,

$$\begin{aligned}\mathcal{F}_j = F_j^k \exp(\frac{\omega_k K}{\lambda}) &\Longrightarrow \sum_{j\ne y}\mathcal{F}_j = \sum_{j\ne y} F_j^k \exp(\frac{\omega_k K}{\lambda})\\ &\Longrightarrow 1-\mathcal{F}_y = (1-F_y^k)\exp(\frac{\omega_k K}{\lambda})\\ &\Longrightarrow \omega_k = \frac{\lambda}{K}\log(\frac{1-\mathcal{F}_y}{1-F_y^k}),\end{aligned}$$

further combine with the first equation, there is

$$\frac{\lambda}{K}\log\left[\frac{\mathcal{F}_y(1-F_y^k)}{F_y^k(1-\mathcal{F}_y)}\right] = \frac{1}{F_y^k}. \tag{3}$$

Since $\mathcal{F}_y = \frac{1}{K}\sum_{k\in[K]} F_y^k$, Eq. (3) cannot holds for all $k \in [K]$, there is no optimal solution in $(0,1)^{L\times K}$.

□

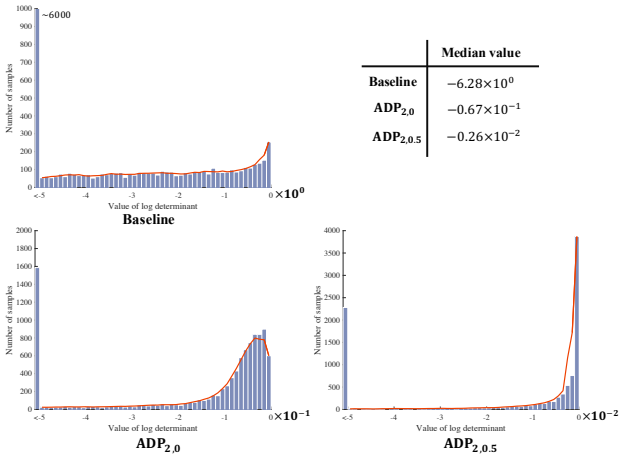
Therefore, it is difficult to appropriately select λ and a balance between accuracy and diversity, which makes it unsuitable to directly define the ensemble diversity as JS-divergence. Note that this dilemma is mainly caused by the second term in the definition of JS-divergence (Eq. (1)) since it can override the ECE term and lead to the wrong prediction with a low value of the total loss.

B.2. Temperature Scaling

Guo et al. (2017) propose the temperature scaling (TS) method to calibrate the predictions of neural networks. The TS method simply use a temperature $T > 0$ on the logits, and return the predictions as

$$F = \mathbb{S}(\frac{z}{T}),$$

where $\mathbb{S}(\cdot)$ is the softmax function, z is the logits. Usually the temperature T is set to be 1 in the training phase, and be larger than 1 in the test phase (Hinton et al., 2015). To solve the numerical obstacle in the ADP training procedure when L is large, we propose to apply the TS method in an opposite way. Namely, in the training phase, we apply a high value of T to increase the values of non-maximal predictions, then in the test phase we apply $T = 1$ to give final predictions. Note that the non-maximal predictions in Corollary 1 equal to $\frac{K(1-\mathcal{F}_y)}{L-1}$. Using the TS method is equivalent to reduce



	Median value
Baseline	-6.28×10^{0}
$\text{ADP}_{2,0}$	-0.67×10^{-1}
$\text{ADP}_{2,0.5}$	-0.26×10^{-2}

Figure 1. The histogram of the ensemble diversity values on the test set of CIFAR-10. There are totally 10,000 samples. The tiny table shows the median values of the ensemble diversity.

$\mathcal{F}_y$ in the training phase. Other possible ways to solve the numerical obstacle can be increasing the number of members in the ensemble (increasing K) or performing a dropout sampling in the ensemble entropy term of the ADP regularizer (decreasing $L-1$). Further investigation of these solutions is one of our future work.

B.3. Histogram of the Ensemble Diversity

To further investigate the relationship between ensemble diversity and robustness, we plot the histogram of the logarithm values of ensemble diversity on the test set of CIFAR-10 in Fig. 1. The median values of different training methods are shown in the top-right panel. An interesting phenomenon is that when the LED part is inactive as in the $\text{ADP}_{2,0}$ setting, the learned ensemble diversity is still much larger than the baseline. This is because even though the ensemble entropy part of ADP regularizer does not explicitly encourage ensemble diversity, it does expand the feasible space of the optimal solution. Due to the degrees of freedom on the optimal individual predictions F_j^k in the feasible space, the ensemble diversity is unlikely to be small. However, the existence of the LED part further explicitly encourage the ensemble diversity, as shown in Fig. 1.

References

Guo, Chuan, Pleiss, Geoff, Sun, Yu, and Weinberger, Kilian Q. On calibration of modern neural networks. *International Conference on Machine Learning (ICML)*, 2017.

Hinton, Geoffrey, Vinyals, Oriol, and Dean, Jeff. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.