Improving Uncertainty Calibration of Deep Neural Networks via Truth Discovery and Geometric Optimization Supplementary Material

Chunwei Ma[1] **Ziyun Huang**[2] **Jiayi Xian**[1] **Mingchen Gao**[1] **Jinhui Xu**[1]

[1]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[2]Computer Science and Software Engineering, Penn State Erie, Erie, PA, USA

A APPENDIX

A.1 PROOF OF THEOREM 3.1

Proof. For the updated truth vector $\mathbf{z}^*$ in each iteration, we want it to maintain the prediction accuracy. That is to say, the index of the maximum component of $\mathbf{z}^*$ could not be different from that of $\mathbf{z}_{ens}$, i.e., $\arg\max_l z_l^* = \arg\max_l (z_{ens})_l$. Let $c = \arg\max_l (z_{ens})_l$. Then, we want to find the projection of $\mathbf{z}^*$ onto the accuracy-preserving simplex $\triangle_a : \{z_c > z_l, \forall l \neq c; 0 \leq z_{l'} \leq 1, \forall l'; \sum_{l'=1}^{L} z_{l'} = 1\}$. This projection can be found through the following constrained optimization problem:

$$\begin{aligned} \underset{\{z_1, z_2, ..., z_L\}}{\text{minimize}} \quad & (z_1 - z_1^*)^2 + (z_2 - z_2^*)^2 + ... + (z_L - z_L^*)^2 \\ \text{s.t.} \quad & z_l < z_c \quad \forall l \neq c \\ & 0 \leq z_{l'} \leq 1 \quad \forall l' \\ & \sum_{l'=1}^{L} z_{l'} = 1. \end{aligned} \tag{1}$$

We solve the above optimization problem using Lagrange Multipliers with Karush-Kuhn-Tucker (KKT) Conditions. Let $\lambda_0, \lambda_1, ..., \lambda_{c-1} \geq 0, \eta_1, ..., \eta_{c-1} \in \mathbb{R}$ and define

$$G = \sum_{l'=1}^{L} (z_{l'} - z_{l'}^*)^2 + \sum_{\forall l \neq c} \lambda_l (z_1 - z_c + \eta_l^2) + \lambda_0 (z_1 + ... + z_L - 1).$$

Taking the derivative with respect to $\mathbf{z}$ and the multipliers gives

$$\begin{aligned} \frac{\partial G}{\partial z_c} &= 2(z_c - z_c^*) + \lambda_0 + \lambda_1 + ... + \lambda_{c-1} \\ \frac{\partial G}{\partial z_l} &= 2(z_l^* - z_l) + \lambda_0 - \lambda_l \quad \forall l \neq c \\ \frac{\partial G}{\partial \lambda_l} &= z_l - z_c + \eta_l^2 \quad \forall l \neq c \\ \frac{\partial G}{\partial \eta_l} &= 2\lambda_l \eta_l \quad \forall l \neq c. \end{aligned}$$

Setting the gradient of the Lagrangian G to 0 and rearranging gives us $\lambda_l = 2(z_1^* - z_1)$, $z_l - z_c + \eta_l^* = 0$, and $\lambda_l \eta_l = 0$, $\forall l \neq c$. For every $\lambda_l \leq 0$, there are two possibilities: (I) if $\lambda_l = 0$, then $z_l = z_l^*$ which means that the l^{th} components will not be changed; (II) if $\lambda_l > 0$, then z_l will be decreased to be the same as z_c. In Algorithm 2, we first sort the components of

$\mathbf{z}^*$, denoted as $\{z_l^{*\prime}\}_{l=1}^L$, then the components less than z_c^* should be fixed, or otherwise it will contradict the above two conditions. Then, we iterate through the top $\tilde{l}$ components of the sorted components until we find an $\tilde{l}$ such that the following condition holds:

$$\frac{1}{\tilde{l}+1}(z_c^* + \sum_{l'=1}^{\tilde{l}} z_{l'}^{*\prime}) > z_{\tilde{l}+1}^{*\prime}. \tag{2}$$

Since $\{z_l^{*\prime}\}_{l=1}^L$ is sorted, Inequal. (2) implies that all conditions in problem (1) hold. Therefore, the solution

$$\tilde{z}_{l'} = \begin{cases} \frac{1}{\tilde{l}+1}(z_c^* + \sum_{l'=1}^{\tilde{l}} z_{l'}^{*\prime}) & \forall l' \leq \tilde{l} \\ z_{l'}^{*\prime} & \forall l' > \tilde{l} \end{cases}$$

given by Algorithm 3.1 will not alter the index of the maximum component. □

A.2 EXPERIMENTAL DETAILS

In this section we describe the detailed experimental setups for post-hoc calibration. Three architectures ResNet18, DenseNet121, and ResNeXt29 are trained on CIFAR10/100 for 200 epochs with a batch size of 128. The learning rate is 0.1 initially and multiplied by 0.2 at the 60^{th}, 120^{th}, and 160^{th} epoch. For all experiment SGD optimizer is used with momentum at 0.9 and weight decay at 5e-4. To obtain as many sources as possible in reasonable time, we treat all models from 200 snapshots as sources for truth discovery and for the computation of the Entropy based Geometric Variance (HV). In the post-hoc calibration step, all the original testing sets are further randomly split into two equal-sized calibration/evaluation datasets, i.e., $N_c = N_e = 5000$ for CIFAR10/100 and $N_c = N_e = 25000$ for ImageNet. This random split is performed 5 times using different seeds, and the resulting means and standard deviations are reported. For all the experiments, the number of bins B is set at 15 and their endpoints are determined by evenly distributing all calibration samples into the 15 bins. Then the number and locations of the bins are fixed during the optimizations. As for the hyperparameters in Eq. (15), we simply set $\alpha_1 = 1, \alpha_2 = \psi_\kappa$ without manually tuning, in order to keep minimal human intervention.

A.3 ADDITIONAL TABLES & FIGURES

Table A1: Comparison between DE/TDE/aTDE on PreResNet110. The best results are highlighted.

		CIFAR100						CIFAR10					
	#sources	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**
DE	10	2.22	2.10	0.638	0.252	1.61	82.32	1.10	0.50	0.113	0.054	0.19	96.28
TDE		2.27	3.06	0.648	0.254	2.78	82.21	1.38	1.35	0.119	0.056	1.35	96.32
aTDE		2.37	3.16	0.648	0.254	2.63	82.32	1.43	1.38	0.119	0.056	1.38	96.28
DE	20	2.65	2.43	0.618	0.247	2.07	82.67	1.25	0.45	0.110	0.053	0.23	96.42
TDE		1.68	1.90	0.621	0.247	1.48	82.61	1.23	1.00	0.112	0.054	0.99	96.39
aTDE		1.66	1.94	0.620	0.247	1.41	82.67	1.21	0.97	0.112	0.054	0.96	96.42
DE	30	2.93	2.69	0.610	0.245	2.33	82.92	1.27	0.43	0.107	0.052	0.25	96.42
TDE		1.59	1.92	0.612	0.245	0.85	82.87	1.15	0.83	0.109	0.053	0.82	96.42
aTDE		1.61	1.97	0.612	0.245	0.80	82.92	1.19	0.83	0.109	0.053	0.82	96.42
DE	40	2.83	2.48	0.606	0.244	2.17	82.79	1.33	0.41	0.107	0.052	0.25	96.41
TDE		1.61	2.05	0.606	0.244	0.72	82.86	1.09	0.78	0.109	0.052	0.77	96.39
aTDE		1.54	1.98	0.606	0.244	0.71	82.79	1.09	0.76	0.109	0.052	0.75	96.41
DE	50	2.89	2.50	0.602	0.244	2.20	82.83	1.31	0.48	0.106	0.052	0.23	96.38
TDE		1.61	1.99	0.602	0.243	0.78	82.89	1.06	0.76	0.107	0.052	0.75	96.38
aTDE		1.55	1.93	0.602	0.243	0.72	82.83	1.07	0.76	0.107	0.052	0.75	96.38
DE	60	3.06	2.62	0.600	0.243	2.39	83.00	1.31	0.38	0.106	0.052	0.23	96.45
TDE		1.75	1.98	0.600	0.242	0.92	82.96	1.05	0.71	0.107	0.052	0.69	96.40
aTDE		1.79	2.02	0.600	0.242	0.96	83.00	1.09	0.66	0.107	0.052	0.64	96.45
DE	70	3.08	2.63	0.598	0.243	2.41	83.02	1.24	0.39	0.106	0.052	0.21	96.42
TDE		1.85	2.02	0.599	0.242	1.03	83.04	1.04	0.70	0.107	0.052	0.68	96.38
aTDE		1.82	2.00	0.599	0.242	1.01	83.02	1.04	0.66	0.107	0.052	0.64	96.42
DE	80	3.06	2.63	0.596	0.243	2.38	83.00	1.21	0.39	0.105	0.052	0.25	96.46
TDE		1.76	2.00	0.597	0.242	1.01	83.00	1.03	0.59	0.106	0.052	0.59	96.45
aTDE		1.78	2.00	0.597	0.242	1.01	83.00	1.01	0.58	0.106	0.052	0.58	96.46
DE	90	3.06	2.59	0.596	0.243	2.39	83.03	1.18	0.39	0.105	0.052	0.22	96.44
TDE		1.83	2.03	0.596	0.242	1.10	83.06	0.99	0.57	0.106	0.052	0.57	96.46
aTDE		1.80	2.00	0.596	0.242	1.07	83.03	0.98	0.59	0.106	0.052	0.59	96.44
DE	100	3.07	2.58	0.595	0.242	2.38	83.02	1.15	0.44	0.106	0.052	0.19	96.41
TDE		1.81	1.95	0.595	0.241	1.05	83.01	1.02	0.60	0.106	0.052	0.60	96.42
aTDE		1.78	1.96	0.595	0.241	1.06	83.02	1.03	0.61	0.106	0.052	0.61	96.41

Table A2: Comparison between DE/TDE/aTDE on PreResNet164. The best results are highlighted.

		CIFAR100						CIFAR10					
	#sources	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**
DE	10	1.90	1.86	0.618	0.242	1.00	82.95	0.93	0.32	0.108	0.052	0.17	96.57
TDE		2.66	3.27	0.632	0.245	3.14	83.02	1.25	1.25	0.115	0.053	1.24	96.55
aTDE		2.60	3.18	0.632	0.245	3.18	82.95	1.28	1.22	0.115	0.053	1.22	96.57
DE	20	2.05	1.87	0.595	0.238	1.49	83.30	1.14	0.28	0.105	0.051	0.12	96.54
TDE		1.73	2.45	0.602	0.239	1.86	83.30	1.24	1.06	0.109	0.052	1.06	96.49
aTDE		1.67	2.46	0.602	0.239	1.86	83.30	1.24	1.01	0.109	0.052	1.01	96.54
DE	30	2.33	2.03	0.588	0.236	1.75	83.49	1.16	0.26	0.103	0.050	0.22	96.62
TDE		1.46	2.01	0.593	0.237	1.27	83.47	1.11	0.87	0.105	0.050	0.86	96.56
aTDE		1.47	2.03	0.593	0.237	1.25	83.49	1.14	0.80	0.105	0.050	0.80	96.62
DE	40	2.50	2.24	0.585	0.236	1.87	83.61	1.27	0.31	0.102	0.050	0.25	96.66
TDE		1.49	1.90	0.589	0.236	0.95	83.58	1.03	0.72	0.104	0.050	0.71	96.64
aTDE		1.52	1.93	0.589	0.236	0.92	83.61	1.07	0.70	0.104	0.050	0.69	96.66
DE	50	2.39	2.09	0.583	0.235	1.81	83.52	1.33	0.35	0.102	0.050	0.28	96.68
TDE		1.32	1.73	0.586	0.235	0.90	83.49	1.14	0.65	0.104	0.050	0.65	96.66
aTDE		1.34	1.76	0.586	0.235	0.86	83.52	1.14	0.63	0.104	0.050	0.63	96.68
DE	60	2.56	2.16	0.580	0.234	1.86	83.58	1.25	0.31	0.101	0.049	0.24	96.66
TDE		1.59	1.75	0.584	0.234	0.70	83.57	1.10	0.64	0.103	0.050	0.64	96.65
aTDE		1.59	1.76	0.584	0.234	0.69	83.58	1.11	0.63	0.103	0.050	0.63	96.66
DE	70	2.56	2.17	0.579	0.234	1.90	83.64	1.23	0.27	0.101	0.049	0.25	96.66
TDE		1.48	1.81	0.582	0.234	0.75	83.67	1.04	0.62	0.102	0.049	0.62	96.63
aTDE		1.49	1.78	0.582	0.234	0.72	83.64	1.05	0.59	0.102	0.049	0.59	96.66
DE	80	2.52	2.03	0.577	0.234	1.83	83.56	1.21	0.25	0.101	0.049	0.19	96.59
TDE		1.49	1.67	0.581	0.233	0.63	83.54	1.07	0.61	0.102	0.049	0.61	96.61
aTDE		1.50	1.69	0.581	0.233	0.65	83.56	1.08	0.63	0.102	0.049	0.63	96.59
DE	90	2.42	2.01	0.576	0.233	1.79	83.51	1.31	0.30	0.100	0.049	0.25	96.66
TDE		1.39	1.58	0.579	0.233	0.64	83.52	1.11	0.55	0.102	0.049	0.54	96.67
aTDE		1.44	1.57	0.579	0.233	0.63	83.51	1.11	0.56	0.102	0.049	0.55	96.66
DE	100	2.47	2.00	0.576	0.233	1.81	83.55	1.32	0.35	0.100	0.049	0.27	96.67
TDE		1.44	1.61	0.578	0.233	0.68	83.54	1.08	0.52	0.102	0.049	0.51	96.68
aTDE		1.42	1.62	0.578	0.233	0.69	83.55	1.09	0.53	0.102	0.049	0.52	96.67

Table A3: Comparison between DE/TDE/aTDE on WideResNet28x10. The best results are highlighted.

		CIFAR100						CIFAR10					
	#sources	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**
DE	10	6.00	5.41	0.629	0.235	5.22	83.97	1.04	0.24	0.094	0.045	0.21	96.93
TDE		4.73	4.69	0.641	0.237	4.01	84.10	1.28	1.06	0.099	0.047	1.05	96.93
aTDE		4.61	4.59	0.641	0.237	3.91	83.97	1.28	1.06	0.099	0.047	1.04	96.93
DE	20	6.27	5.70	0.621	0.233	5.49	84.21	0.96	0.23	0.092	0.045	0.11	97.02
TDE		5.05	4.62	0.628	0.234	4.25	84.23	1.19	0.80	0.096	0.046	0.78	97.00
aTDE		5.04	4.61	0.628	0.234	4.24	84.21	1.18	0.77	0.096	0.046	0.76	97.02
DE	30	6.28	5.73	0.617	0.232	5.49	84.22	1.07	0.33	0.091	0.045	0.24	97.15
TDE		5.20	4.69	0.623	0.233	4.35	84.26	1.15	0.60	0.094	0.045	0.57	97.12
aTDE		5.16	4.66	0.623	0.233	4.32	84.22	1.17	0.57	0.094	0.045	0.54	97.15
DE	40	6.39	5.75	0.616	0.232	5.58	84.28	1.09	0.28	0.091	0.044	0.21	97.14
TDE		5.39	4.86	0.622	0.232	4.54	84.31	1.25	0.53	0.093	0.045	0.51	97.14
aTDE		5.36	4.83	0.622	0.232	4.51	84.28	1.27	0.53	0.093	0.045	0.51	97.14
DE	50	6.45	5.86	0.615	0.232	5.64	84.38	1.07	0.35	0.091	0.044	0.24	97.17
TDE		5.48	5.04	0.621	0.232	4.62	84.37	1.15	0.49	0.093	0.045	0.47	97.15
aTDE		5.50	5.05	0.621	0.232	4.63	84.38	1.14	0.47	0.093	0.045	0.45	97.17
DE	60	6.39	5.72	0.614	0.232	5.57	84.30	1.05	0.34	0.091	0.044	0.21	97.14
TDE		5.53	5.00	0.619	0.232	4.66	84.36	1.14	0.49	0.093	0.045	0.47	97.13
aTDE		5.47	4.94	0.619	0.232	4.60	84.30	1.13	0.48	0.093	0.045	0.46	97.14
DE	70	6.42	5.75	0.614	0.232	5.59	84.31	1.10	0.32	0.090	0.044	0.23	97.15
TDE		5.55	5.02	0.619	0.232	4.70	84.34	1.10	0.46	0.092	0.045	0.44	97.14
aTDE		5.52	4.99	0.619	0.232	4.67	84.31	1.11	0.45	0.092	0.045	0.43	97.15
DE	80	6.44	5.80	0.615	0.232	5.62	84.30	1.14	0.33	0.090	0.044	0.26	97.16
TDE		5.54	5.00	0.619	0.232	4.70	84.27	1.10	0.45	0.092	0.044	0.41	97.15
aTDE		5.57	5.03	0.619	0.232	4.74	84.30	1.11	0.44	0.092	0.044	0.40	97.16
DE	90	6.43	5.83	0.614	0.232	5.63	84.30	1.20	0.42	0.090	0.044	0.29	97.19
TDE		5.59	5.11	0.619	0.232	4.75	84.31	1.12	0.39	0.092	0.044	0.36	97.19
aTDE		5.58	5.10	0.619	0.232	4.74	84.30	1.12	0.39	0.092	0.044	0.36	97.19
DE	100	6.41	5.79	0.614	0.231	5.61	84.28	1.20	0.41	0.090	0.044	0.30	97.20
TDE		5.58	5.08	0.618	0.232	4.74	84.29	1.11	0.40	0.092	0.044	0.36	97.17
aTDE		5.57	5.07	0.618	0.232	4.73	84.28	1.13	0.42	0.092	0.044	0.33	97.20

Table A4: Comparison between DE/TDE/aTDE on ImageNet. The best results are highlighted.

		CIFAR100						CIFAR10					
	#sources	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**	**ECE^{KDE}↓**	**ECE↓**	**NLL↓**	**MSE↓**	**KS↓**	**ACC↑**
DE	5	2.32	2.45	0.831	0.302	2.04	78.65	2.00	2.18	0.866	0.313	2.18	77.79
TDE		1.99	2.43	0.838	0.304	0.90	78.38	1.99	2.46	0.874	0.315	2.46	77.61
aTDE		2.33	2.78	0.837	0.304	1.26	78.65	2.21	2.69	0.873	0.315	2.69	77.79
DE	10	2.94	2.94	0.808	0.297	2.60	79.15	2.12	2.32	0.843	0.307	2.32	78.25
TDE		1.96	2.31	0.809	0.297	1.41	79.02	1.78	2.17	0.846	0.307	2.17	78.17
aTDE		2.11	2.46	0.809	0.297	1.56	79.15	1.88	2.26	0.846	0.307	2.26	78.25
DE	15	2.96	2.92	0.802	0.295	2.61	79.15	1.98	2.20	0.837	0.304	2.20	78.37
TDE		2.04	2.32	0.801	0.295	1.60	79.17	1.74	2.06	0.839	0.305	2.06	78.32
aTDE		2.03	2.31	0.801	0.295	1.59	79.15	1.79	2.12	0.839	0.305	2.12	78.37
DE	20	3.10	3.07	0.799	0.295	2.76	79.26	1.89	2.14	0.834	0.304	2.14	78.42
TDE		2.14	2.36	0.798	0.294	1.75	79.21	1.65	2.01	0.837	0.304	2.01	78.35
aTDE		2.20	2.42	0.798	0.294	1.81	79.26	1.73	2.08	0.837	0.304	2.08	78.42
DE	25	3.11	3.12	0.797	0.295	2.79	79.25	1.86	2.12	0.833	0.303	2.12	78.46
TDE		2.17	2.42	0.796	0.294	1.81	79.22	1.69	2.09	0.836	0.303	2.09	78.45
aTDE		2.20	2.45	0.796	0.294	1.84	79.25	1.71	2.11	0.836	0.303	2.11	78.46
DE	30	3.11	3.09	0.795	0.294	2.78	79.27	1.82	2.08	0.832	0.303	2.08	78.48
TDE		2.24	2.47	0.794	0.293	1.88	79.26	1.72	2.11	0.835	0.303	2.11	78.45
aTDE		2.24	2.47	0.794	0.293	1.88	79.27	1.75	2.14	0.835	0.303	2.14	78.48
DE	35	3.14	3.10	0.794	0.294	2.80	79.27	1.79	2.07	0.831	0.303	2.07	78.50
TDE		2.28	2.49	0.793	0.293	1.92	79.26	1.73	2.14	0.834	0.303	2.14	78.50
aTDE		2.29	2.50	0.793	0.293	1.94	79.27	1.74	2.15	0.834	0.303	2.15	78.50
DE	40	3.17	3.12	0.794	0.294	2.82	79.30	1.77	2.06	0.830	0.302	2.06	78.51
TDE		2.31	2.50	0.793	0.293	1.97	79.29	1.72	2.13	0.833	0.302	2.13	78.51
aTDE		2.33	2.52	0.793	0.293	1.98	79.30	1.72	2.14	0.833	0.302	2.14	78.51
DE	45	3.21	3.14	0.793	0.294	2.86	79.34	1.78	2.06	0.830	0.302	2.06	78.54
TDE		2.35	2.54	0.792	0.293	2.01	79.31	1.72	2.18	0.833	0.302	2.18	78.52
aTDE		2.38	2.58	0.792	0.293	2.04	79.34	1.74	2.20	0.833	0.302	2.20	78.54
DE	50	3.24	3.17	0.792	0.294	2.89	79.37	1.73	2.04	0.830	0.302	2.04	78.52
TDE		2.41	2.58	0.791	0.293	2.07	79.35	1.66	2.11	0.833	0.302	2.11	78.50
aTDE		2.44	2.61	0.791	0.293	2.10	79.37	1.69	2.14	0.833	0.302	2.14	78.52

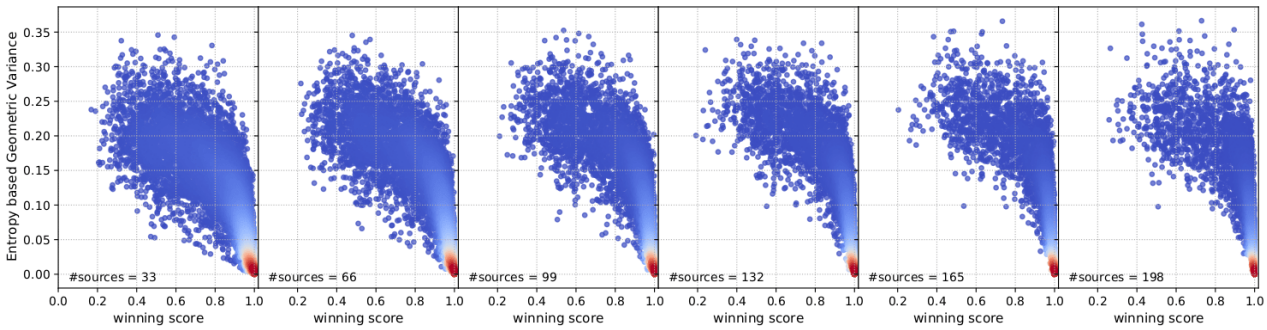

Figure A1: Correlation between Entropy based Geometric Variance (HV) and winning score of DenseNet121 trained on CIFAR10. Winning score stands for the score of predicted label, i.e. the maximum score of the network's output.

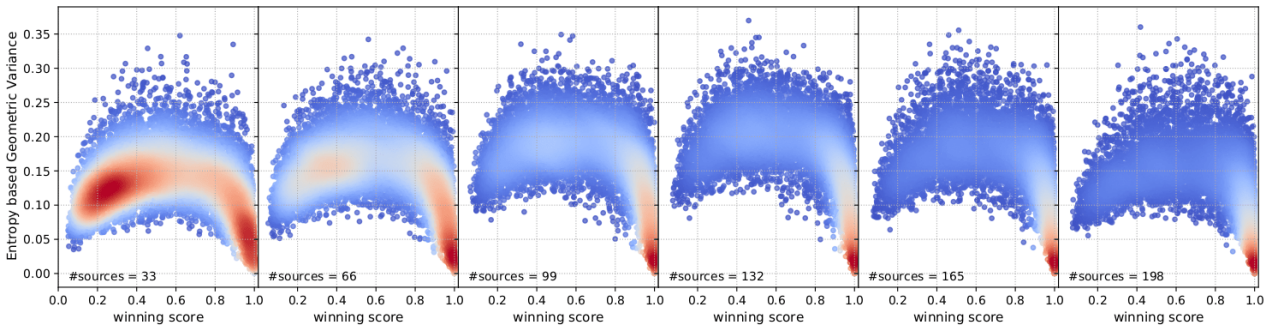

Figure A2: Correlation between Entropy based Geometric Variance (HV) and winning score of ResNet18 trained on CIFAR100.

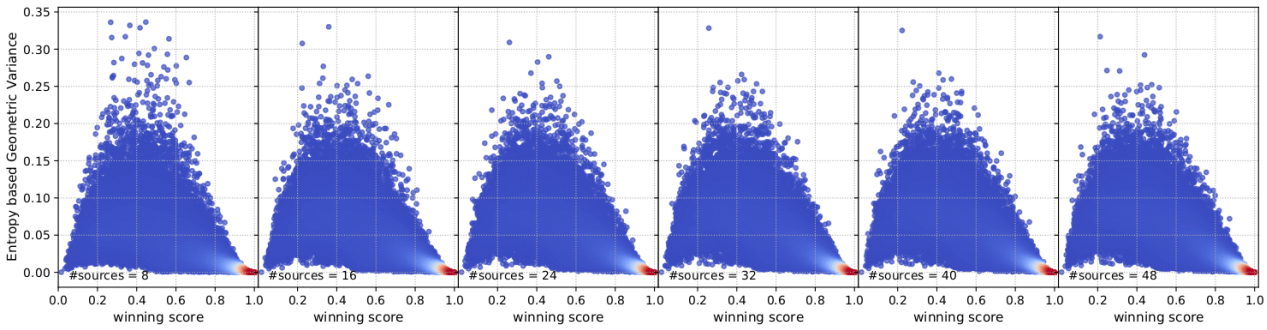

Figure A3: Correlation of Entropy based Geometric Variance (HV) and winning score of ResNet50 trained on ImageNet using DE.