Counterfactual Inference of Second Opinions (Supplementary material)

Nina L. Corvelo Benz[1,2] **Manuel Gomez Rodriguez**[1]

[1]Max Planck Institute for Software Systems, Kaiserslautern, Germany
[2]Department of Biosystems Science and Engineering, ETH Zurich, Zurich, Switzerland

1 PROOFS

Proof of Theorem 2.

Let $\zeta \subseteq \zeta' \subseteq \mathcal{H}$ and both non-empty. Then, for any $x \in \mathcal{X}, u \in \mathcal{U}$, we have that

$$(f(x, \zeta', u))_\zeta = ((f_h(x, u))_{h \in \zeta'})_\zeta = (f_h(x, u))_{h \in \zeta} = f(x, \zeta, u).$$

Proof of Theorem 3. Let $\mathcal{M}'$ be constructed from $\mathcal{M}$ by changing causal mechanism f with f'. To prove equivalence of $\mathcal{M}$ and $\mathcal{M}'$, we only need to show that, for any $u \in \mathcal{U}, x \in \mathcal{X}, \zeta \in \mathcal{P}(\mathcal{H}) \setminus \emptyset$, it holds that

$$\mathbf{y} = f(x, \zeta, u) \iff \mathbf{y} = f'(x, \zeta, u), \tag{1}$$

as only the causal mechanism was altered in the construction of $\mathcal{M}'$. Let $h \in \zeta$ be an arbitrary expert, then

$$(f'(x, \zeta, u))_h \overset{def.}{=} f(x, \{h\}, u) = (f(x, \zeta, u))_h,$$

where the last equality holds because f is set invariant. Thus, $f(x, \zeta, u) = f'(x, \zeta, u)$ for all $x \in \mathcal{X}, \zeta \in \mathcal{H}, u \in \mathcal{U}$.

Proof of Theorem 4. For clarity, we explicitly write $\mathbf{Y}_{Z=\{h'\}}$ and $\mathbf{Y}_{Z=\zeta}$ to better distinguish the two interventional outcomes. For discrete probability distribution $P(U)$ the right probability is given by

$$P^{M;do[Z:=\zeta']}(\mathbf{Y}_\zeta = \mathbf{y} \mid X = x) = \sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta', u))_\zeta = \mathbf{y})],$$

whereas the left is given by

$$P^{M;do[Z:=\zeta]}(\mathbf{Y} = \mathbf{y} \mid X = x) = \sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta, u) = \mathbf{y}].$$

Because f is a set invariant mechanism over Z, $(f(x, \zeta', u))_\zeta = f(x, \zeta, u)$, thus,

$$\sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta, u) = \mathbf{y}] = \sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta', u))_\zeta = \mathbf{y}].$$

The proof is analogous for the continuous probability distributions $P(U)$.

Proof of Corollary 1. Choose $\zeta = \{h\}$ in Theorem 4 and note that abusing notation $\mathbf{Y}_{Z=\zeta}$ is in this case equivalent to Y_h.

Proof of Corollary 2. Using the definition of counterfactual distributions, the proof is analogous to the proof of Theorem 4 but using the posterior distribution $P(U \mid X = x, Z = \{h\}, Y_h = c)$. Let $\zeta \subseteq H$ be so that $h, h' \in \zeta$. For all $c \in \mathcal{Y}$, by definition, we have that

$$P^{M|X=x,Z=\{h\},\mathbf{Y}=c;do[Z:=\{h'\}]}(\mathbf{Y} = c') = \sum_{u \in \mathcal{U}} P(U = u \mid X, (\mathbf{Y}_{Z=\{h\}})_h = c) \cdot \mathbb{1}[f(x, \{h'\}, u) = c']. \tag{2}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Using that $\mathcal{M}$ is set invariant we get that

$$(\mathbf{Y}_{Z=\{h\}})_h = f(x,\{h\},U) = (f(x,\zeta,U))_h = (\mathbf{Y}_{Z=\zeta})_h \quad \text{and} \quad \mathbb{1}[f(x,\{h'\},u)) = c'] = \mathbb{1}[(f(x,\zeta,u))_{h'} = c'].$$

Thus, Eq. (2) is equal to

$$\sum_{u\in\mathcal{U}} P(U=u \mid X, (\mathbf{Y}_{Z=\zeta})_h = c)\cdot \mathbb{1}[(f(x,\zeta,u))_{h'} = c'] \overset{def.}{=} P^{M;do[Z:=\zeta]}(Y_{h'} = c' \mid X = x, Y_h = c).$$

Proof of Theorem 7. Note that for all $\zeta_1,\zeta_2,\zeta_3 \subseteq H$ so that $h\in\zeta_1, h'\in\zeta_2, h,h'\in\zeta_3$ holds that

$$P^{\mathcal{M}|X=x,Z=\zeta_1,Y_h=c;do[Z=\zeta_2]}(Y_{h'}=c') = 0 \iff P^{\mathcal{M}|X=x,Z=\{h\},Y_h=c;do[Z=\{h'\}]}(Y_{h'}=c') = 0, \tag{3}$$

$$P^{\mathcal{M}|X=x,Z=\{h\},Y_h=c;do[Z=\{h'\}]}(Y_{h'}=c') = 0 \iff P^{\mathcal{M};do[Z=\zeta_3]}(Y_{h'}=c' \mid X=x, Z=\zeta_3, Y_h=c) = 0, \tag{4}$$

where Eq. (3) follows from the definition of set invariance and Eq. (4) follows from Corollary 2. Recall that $p_\zeta(h,c) := P^{\mathcal{M};do[Z=\zeta]}(Y_h = c \mid X)$ and $p_h(c) := P^{\mathcal{M};do[Z=\{h\}]}(Y_h = c \mid X)$.

It follows from Corollary 1 that, for all $h\in H$, $c\in\mathcal{Y}$ and ζ,ζ' so that $h\in\zeta$ and $h\in\zeta'$, we have that

$$p_\zeta(h,c) = p_h(h,c) = p_{\zeta'}(h,c).$$

Thus, following implications hold

$$\frac{p_{\zeta_2}(h',c)}{p_{\zeta_1}(h,c)} \geq \frac{p_{\zeta_2}(h',c')}{p_{\zeta_1}(h,c')} \iff \frac{p_{h'}(h',c)}{p_h(h,c)} \geq \frac{p_{h'}(h',c')}{p_h(h,c')} \iff \frac{p_{\zeta_3}(h',c)}{p_{\zeta_3}(h,c)} \geq \frac{p_{\zeta_3}(h',c')}{p_{\zeta_3}(h,c')}.$$

With these set of implications, it is straight forward to imply one statement from the other.

Proof of Theorem 10. Let ψ be a subgroup in Ψ so that $|\psi|\geq 2$. Let h and h' denote two arbitrary experts in subgroup ψ. As the Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ is set invariant, it is enough to show that pairwise conditional stability condition is satisfied for pair h and h'. Analogously to Oberst and Sontag [2019], we proceed by proving the contrapositive, that for all sets ζ, so that $h,h'\in\zeta$, and $c\neq c'$

$$P^{\mathcal{M}(\Psi);do[Z=\zeta]}(Y_{h'}=c' \mid X, Y_h = c) \neq 0 \implies \frac{p_\zeta(h',c)}{p_\zeta(h,c)} < \frac{p_\zeta(h',c')}{p_\zeta(h,c')}.$$

If the conditional probability is positive, almost surely there must exist Gumbel noise variables $g_{\psi,c}$ and $g_{\psi,c'}$ such that

$$\log P(Y_h = c \mid X) + g_{\psi,c} > \log P(Y_h = c' \mid X) + g_{\psi,c'}$$
$$\log P(Y_{h'} = c \mid X) + g_{\psi,c} < \log P(Y_{h'} = c' \mid X) + g_{\psi,c'},$$

as the sub-mechanisms f_h and $f_{h'}$ of each expert share the noise vector of the subgroup ψ.

Recall that, by set invariance, $p_\zeta(h,c) = P(Y_h = c \mid X)$ for all ζ, h, c. Hence, we can substitute the probabilities in both inequalities. Then, we further subtract the first inequality from the second which cancels out the Gumbel noises. Finally, using the properties of the logarithm function, the inequality is rearranged deriving the implication.

$$\log p_\zeta(h',c) - \log p_\zeta(h,c) < \log p_\zeta(h',c') - \log p_\zeta(h,c'),$$
$$\frac{p_\zeta(h',c)}{p_\zeta(h,c)} < \frac{p_\zeta(h',c')}{p_\zeta(h,c')}.$$

This proves that the Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ satisfies the pairwise conditional stability condition

$$\frac{p_\zeta(h',c)}{p_\zeta(h,c)} \geq \frac{p_\zeta(h',c')}{p_\zeta(h,c')} \implies P^{\mathcal{M}(\Psi);do[Z=\zeta]}(Y_{h'}=c' \mid X, Y_h = c) = 0,$$

for any two experts in the same subgroup in Ψ.

2 RANDOMIZED GREEDY ALGORITHM FOR THE CLIQUE PARTITIONING PROBLEM

The idea behind the simple greedy randomized Algorithm 1 is to sequentially grow a clique starting from a random vertex in $\mathcal{G}$ until no vertices can be added, remove this clique from the graph and repeat this process on the remaining graph until no vertices are left. For the current clique ψ the set of vertices that can be added, called candidate set, consists of vertices that have edges to all the vertices in ψ so that the sum these edge weights is non-positive. The expert with minimum sum is added next to the clique, and the candidate set is updated.

The updated candidate set is a subset of the previous set, thus, the sum of edge weights of a vertex connected to the updated clique can be computed in constant time by considering the previous value and the weight of the edge to the newly added vertex. If no edge exists, the vertex can be removed from the candidate set. Algorithm 1 can thus be implemented in $O(|\mathcal{E}|)$.

Algorithm 1 Greedy Algorithm for the Clique Partitioning Problem, $N(\psi)$ denotes the set of vertices not in ψ with edges to all vertices in ψ

Input: weighted, undirected graph $\mathcal{G} = (\mathcal{H}, \mathcal{E}, w)$
while $\mathcal{G}$ not empty **do**
  pick random vertex h
  $\psi \leftarrow h$
  while $N(\psi)$ not empty **do**
    $h^* \leftarrow \arg\min_{h' \in N(\psi)} \sum_{h \in \psi} w(\{h', h\})$
    if $\sum_{h \in \psi} w(\{h^*, h\}) \leq 0$ **then**
      $\psi \leftarrow h^*$
    else
      break
    end if
  end while
  $\Psi \leftarrow \psi$
  delete ψ from $\mathcal{G}$
  $\psi = \emptyset$
end while
return Ψ

We note that, since the algorithm minimizes the sum of weights for each clique sequentially, it's performance in recovering a partition minimizing the overall sum depends on the sequence of sampled vertices which we start each clique from. To stabilize the algorithm's performance, one can rerun the algorithm a few times and choose among the returned partitions the one minimizing the overall sum of edge weights between vertices in the same set (see objective function in optimization problem 7). In the experiments on real (synthetic) data, we run Algorithm 1 10 (5) times.

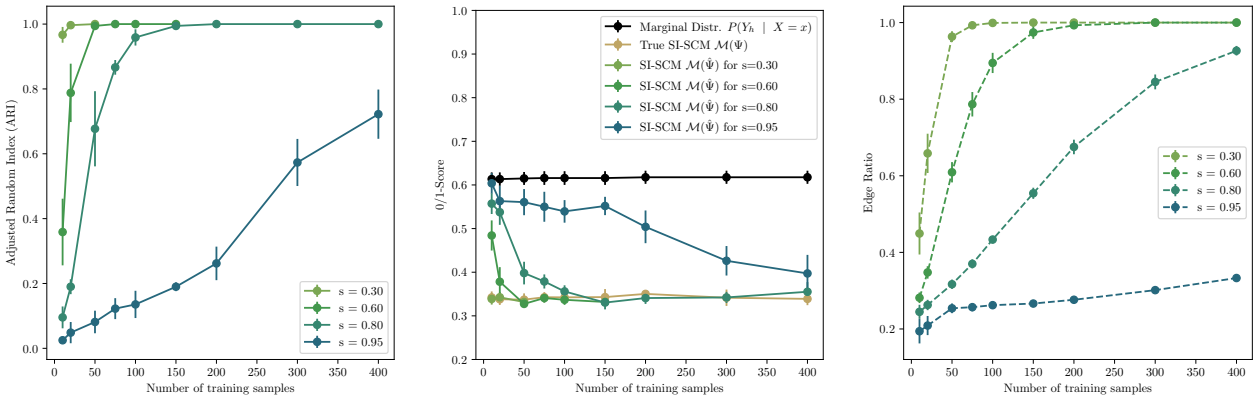
3 EXPERIMENTS ON SYNTHETIC DATA

In this section, we assess the performance of Algorithm 1 at recovering the groups of mutually similar experts underpinning our Gumbel-Max SI-SCM using synthetic data.

Experimental setup. We consider a synthetic prediction task with $k = 5$ labels and 20 features per sample, whose values we sample uniformly at random from the interval $[0, 1]$, and a set $\mathcal{H}$ of 48 synthetic experts. These synthetic experts make label predictions according to a Gumbel-Max SI-SCM with five disjoint groups of mutually similar experts Ψ, *i.e.*, each expert within a group $\psi \in \Psi$ use the same Gumbel noise within the model[1]. Moreover, for each expert, the probability $P(Y_h = c \mid X = x)$ is given by a multinomial logit model with random weight coefficients $w = (w_1, \ldots, w_5)$, which we also sample uniformly at random from the interval $[0, 1]$ independently for each expert, *i.e.*,

$$P(Y_h = c \mid X = x) = \frac{\exp(w_c \cdot x)}{\sum_{j \in \mathcal{Y}} \exp(w_j \cdot x)}. \tag{5}$$

[1]The groups in the partition Ψ contain 6, 7, 11, 11 and 13 experts.



(a) Adjusted random index (ARI) (b) Average 0/1-loss on a held-out set (c) Edge ratio

Figure 1: Performance of Algorithm 1 at recovering the partition Ψ given the true probabilities $P(Y_h = c \mid X = x)$ under different amounts of data m and sparsity levels $s \in (0, 1)$. To compute the mean and the standard deviation in panels (a), (b) and (c), we run each experiment five times.

We measure the performance of Algorithm 1 at recovering the partition Ψ given the true probabilities $P(Y_h = c \mid X = x)$ under different amounts of training data m and sparsity levels $s \in (0, 1)$. The sparsity level s controls the average number of observed expert predictions per sample, *i.e.*, for each sample, all experts make a prediction but we only observe $\max\{2, (1 - s)|\mathcal{H}|\}$, picked at random. Here, note that, as the sparsity level s decreases (increases) and the amount of training data increases (decreases), it is easier (harder) to recover the partition Ψ.

As a measure of the difficulty of each inference problem, we will use the edge ratio r, defined as the fraction of pair of experts who belong to the same group $\psi \in \Psi$, among all pairs whose predictions did not violate conditional stability and were at least once observed for the same sample. As performance metrics, we will use:

- The adjusted random index (ARI), which measures similarity between the partition $\hat{\Psi}$ returned by Algorithm 1 and the true partition Ψ. Its value lies in the interval $[0, 1]$ where 1.0 means full recovery and 0.0 means a completely random partition was recovered with no similarity to the true one.
- The average $0/1$-loss on a held-out set (with 1000 samples) of a predictor that, given an observed label Y_h, returns the most likely label $Y_{h'}$ under the inferred counterfactual distribution $P^{\mathcal{M}(\hat{\Psi}) \mid X=x, Z=\{h\}\mathbf{Y}=y_h; \text{do}(Z=\{h'\})(\mathbf{Y})}$. Here, to estimate the inferred counterfactual distributions, we use 500 samples.

As a point of comparison, for the second performance metric, we also compute the average $0/1$-loss over the same held-out set of two other predictors that, given an observed label Y_h, return the most likely label $Y_{h'}$ under the true counterfactual distributions $P^{\mathcal{M}(\Psi) \mid X=x, Z=\{h\}\mathbf{Y}=y_h; \text{do}(Z=\{h'\})(\mathbf{Y})}$ and the counterfactual distribution $P^{\mathcal{M}(\mathcal{H}) \mid X=x, Z=\{h\}\mathbf{Y}=y_h; \text{do}(Z=\{h'\})(\mathbf{Y})}$, respectively.

Results. Figure 1 summarizes the results, which show that, as long as the edge ratio $r > 0.3$, the inferred partition $\hat{\Psi}$ is very similar to the true partition Ψ (*i.e.*, the value of ARI is very close to 1) and the $0/1$-losses of the predictors that use $\mathcal{M}(\hat{\Psi})$ and $\mathcal{M}(\Psi)$ respectively are very similar. Here, note however that, even the predictor that uses the true model $\mathcal{M}(\Psi)$ has a non zero $0/1$-loss is not error free because, given an observed expert prediction Y_h and feature vector x, the expert prediction $Y_{h'}$ is not deterministic.

4 ADDITIONAL FIGURES FOR EXPERIMENTS ON REAL DATA

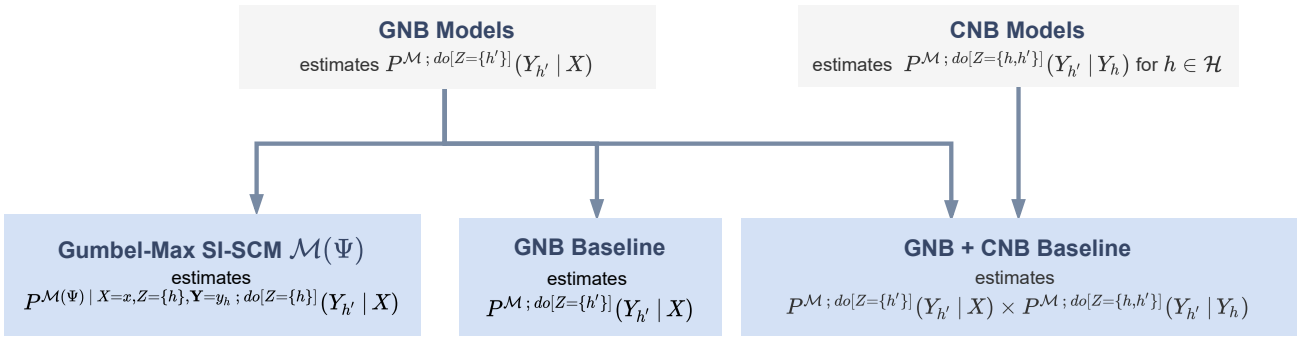


Figure 2: Different models used in our experiments on real data.

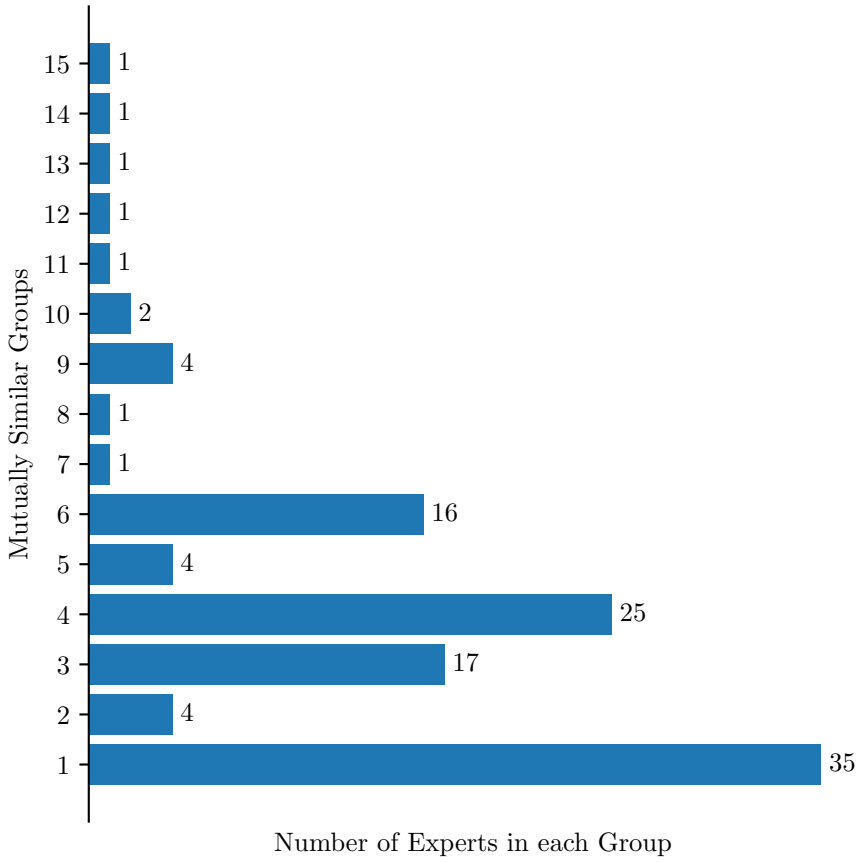
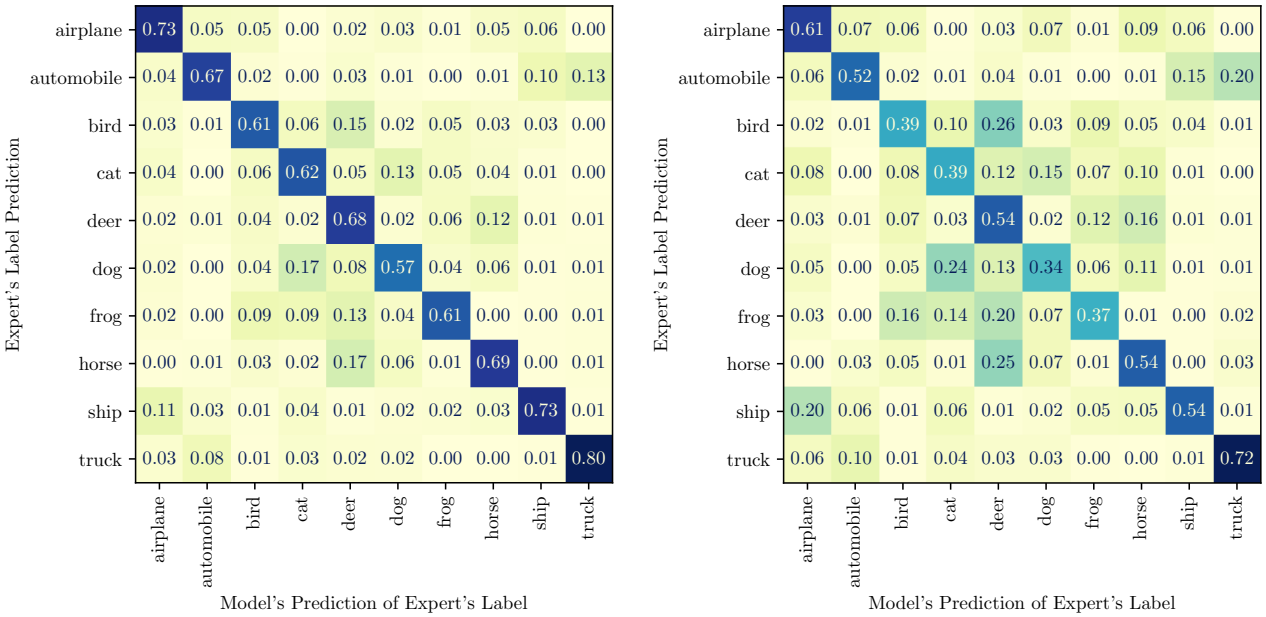


Figure 3: Size of the mutually similar expert groups returned by Algorithm 1 for the preprocessed CIFAR-10H dataset.



(a) Gumbel-Max SI-SCM (b) GNB

(c) GNB+CNB

Figure 4: Confusion matrices of the counterfactual predictions of our model and the predictions of the two baselines for expert's labels on the test dataset.

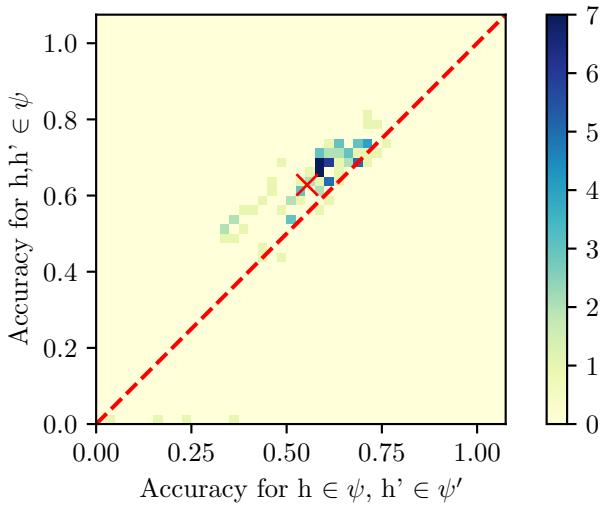


Figure 5: Per-expert test accuracy achieved by the baseline GNB+CNB on the preprocessed CIFAR-10H dataset. For each expert h', the y-axis measures the test accuracy whenever the observed expert h belongs to the same group of mutually similar experts as h' and the x-axis measures the test accuracy whenever h does not belong to the same group. For each cell, the darkness is proportional to the number of experts with the corresponding test accuracies.

References

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.