Sequential Cooperative Bayesian Inference: Supplementary Material

Junqi Wang [1] **Pei Wang** [1] **Patrick Shafto** [1]

Proof of Theorem 3.1

Theorem 3.1. *[(Miescke and Liese, 2008, Theorem 7.115)] In BI, the sequence of posteriors (S_k) is strongly consistent at $\widehat{\theta} = \delta_h$ for each $h \in \mathcal{H}$, with arbitrary choice of an interior point $\theta_0 \in (\mathcal{P}(\mathcal{H}))^\circ$ (i.e. $\theta_0(h) > 0$ for all $h \in \mathcal{H}$) as prior.*

Proof. We follow the same line as discussed right after this theorem in the paper. Let $\theta_0 = (\theta_0(1), \theta_0(2), \ldots, \theta_0(n))$ be the original prior, and let $\theta_k = (\theta_k(1), \theta_k(2), \ldots, \theta_k(m))$ be the posterior after having k data points $d_1, d_2, \ldots, d_k$. Then for $l \leq k$ and $h \in \mathcal{H}$, the posterior $\theta_l(i) = \left(\mathscr{N}_{\text{vec}}(\text{diag}(\mathbf{M}_{(d_l,\text{-})})\theta_{l-1})\right)(i)$ by Bayes' rule. In other words,

$$\theta_l(i) = \frac{\mathbf{M}_{(d_l,i)}[\theta_{(l-1)}(i)]}{\sum_{j=1}^m \mathbf{M}_{(d_l,j)}[\theta_{(l-1)}(j)]}. \tag{1}$$

This is a recursive formula, so we may move forward to calculate $\theta_l(i)$ from a smaller round index $\theta_t(i)$ with $t < l$:

$$\theta_l(i) = \frac{\left[\prod_{s=t}^l \mathbf{M}_{(d_s,i)}\right]\theta_{(t-1)}(i)}{\sum_{j=1}^m \left[\prod_{s=t}^l \mathbf{M}_{(d_s,j)}\right]\theta_{(t-1)}(j)}.$$

This recursion stops at prior θ_0, so we have an explicit expression of θ_k:

$$\theta_k(i) = \frac{\left[\prod_{s=1}^k \mathbf{M}_{(d_s,i)}\right]\theta_0(i)}{\sum_{j=1}^m \left[\prod_{s=1}^k \mathbf{M}_{(d_s,j)}\right]\theta_0(j)}. \tag{2}$$

It can be seen that for each hypothesis i, the denominator of the k-th posterior on i are the same, so we have

$$\frac{\theta_k(i)}{\theta_k(h)} = \frac{\left[\prod_{s=1}^k \mathbf{M}_{(d_s,i)}\right]\theta_0(i)}{\left[\prod_{s=1}^k \mathbf{M}_{(d_s,h)}\right]\theta_0(h)}. \tag{3}$$

So we define $\alpha_k(d)$ to be the frequency of the occurrence of data d in the first k rounds of a episode. And then

$$\log\left(\frac{\theta_k(i)}{\theta_k(h)}\right) = \log\left(\frac{\theta_0(i)}{\theta_0(h)}\right) + \sum_{d=1}^n \alpha_k(d)\log\left(\frac{\mathbf{M}_{(d,i)}}{\mathbf{M}_{(d,h)}}\right). \tag{4}$$

Since we know that the data (d_i) in the model is sampled following the i.i.d. with distribution $\mathbf{M}_{(\text{-},h)}$, then for a fixed k, $\alpha_k(i)$ follows the multinomial distribution with parameter $\mathbf{M}_{(\text{-},h)}$.

By the strong law of large numbers, $\frac{\alpha_k(i)}{k} \to \mathbf{M}_{(i,h)}$ almost surely as $k \to \infty$. Thus, when we rewrite the sample values to random variable version,

$$\frac{1}{k}\log\left(\frac{\Theta_k(i)}{\Theta_k(h)}\right) \to \sum_{d=1}^n \mathbf{M}_{(d,h)}\log\left(\frac{\mathbf{M}_{(d,i)}}{\mathbf{M}_{(d,h)}}\right) \quad \text{a.s.} \tag{5}$$

That is,

$$\frac{1}{k}\log\left(\frac{\Theta_k(i)}{\Theta_k(h)}\right) \longrightarrow -\text{KL}\left(\mathbf{M}_{(\text{-},h)}, \mathbf{M}_{(\text{-},i)}\right) \quad \text{a.s.} \tag{6}$$

By the assumption in Section 2 of the paper that $\mathbf{M}$ has distinct columns, the KL divergence between the i-th column and the h-th column is strictly positive, thus almost surely, $\log\left(\frac{\Theta_k(i)}{\Theta_k(h)}\right) \to -\infty$, or equivalently, $\frac{\Theta_k(i)}{\Theta_k(h)} \to 0$, for any $i \neq h$.

Therefore, $\theta_k = (\theta_k(1), \theta_k(2), \ldots, \theta_k(m)) \to \delta_h$ almost surely, equivalently, BI at $\widehat{\theta}$ is strongly consistent. □

Proof of Theorem 3.2

Theorem 3.2. *In BI, with $\widehat{\theta} = \delta_h$ for some $h \in \mathcal{H}$, let $\Theta_k(h)(\text{D}_1, \ldots, \text{D}_k) := S_k(h|\text{D}_1, \ldots, \text{D}_k)$ be the h-component of posterior given $\text{D}_1, \ldots, \text{D}_k$ as random variables valued in $\mathcal{D}$. Then $\frac{1}{k}\log\left(\frac{\Theta_k(h)}{1-\Theta_k(h)}\right)$ converges to a constant $\min_{h' \neq h}\left\{\text{KL}(\mathbf{M}_{(\text{-},h)}, \mathbf{M}_{(\text{-},h')})\right\}$ almost surely.*

Proof. Follow the previous proof. First recall that $\frac{1}{k}\log\left(\frac{\Theta_k(i)}{\Theta_k(h)}\right) \to -\text{KL}(\mathbf{M}_{(\text{-},h)}, \mathbf{M}_{(\text{-},i)})$ almost surely. Let $\eta := \text{argmin}_{i \neq h}\left\{\text{KL}(\mathbf{M}_{(\text{-},h)}, \mathbf{M}_{(\text{-},i)})\right\}$, then $\Theta_k(\eta)$ decays slowest among $\{\Theta_k(i) : i \neq h\}$ almost surely.

Therefore, for the sample values θ_k's, asymptotically,

$$\frac{1}{k}\log\left[\frac{\theta_k(\eta)}{\theta_k(h)}\right] \leq \frac{1}{k}\log\left[\frac{1-\theta_k(h)}{\theta_k(h)}\right] \leq \frac{1}{k}\log\left[\frac{(m-1)\theta_k(\eta)}{\theta_k(h)}\right].$$

So when we are taking limits $k \to \infty$, with probability one, we have

$$\begin{aligned}
&-\mathrm{KL}(\mathbf{M}_{(_,h)}, \mathbf{M}_{(_,\eta)}) \leq \lim_{k\to\infty} \frac{1}{k} \log\left[\frac{1-\theta_k(h)}{\theta_k(h)}\right] \\
&\leq \lim_{k\to\infty} -\mathrm{KL}(\mathbf{M}_{(_,h)}, \mathbf{M}_{(_,\eta)}) + \frac{1}{k}\log(m-1) \\
&= -\mathrm{KL}(\mathbf{M}_{(_,h)}, \mathbf{M}_{(_,\eta)}). \qquad (7)
\end{aligned}$$

□

Proof of Theorem 3.5

To prove Theorem 3.5, we need the following lemmas.

Lemma S.1. *Given a fixed hypothesis $h \in \mathcal{H}$, for any $\mu \in \mathcal{P}(\Delta^{m-1})$,*

$$\mathbb{E}_{\mu}(\theta(h)) \leq \mathbb{E}_{\Psi(h)(\mu)}(\theta(h)). \qquad (8)$$

equality happens when $\mathbf{M}^{\langle n\mathbf{x}\rangle}_{(i,h)} = \mathbf{M}^{\langle n\mathbf{x}\rangle}_{(j,h)}$ for any i, j and μ-almost everywhere for $\mathbf{x} \in \Delta^{m-1}$.

Remark 1. This lemma shows that the expectation of $\theta(h)$, in each round is increasing, thus the sequence obtained from all the rounds has an limit since the sequence is monotonic and upper bounded by 1. To prove the theorem we, then just need to show the limit is 1.

Proof. We start from the right hand side of Eq. 8. Let Δ denote Δ^{m-1} for short.

$$\begin{aligned}
&\mathbb{E}_{\Psi(h)(\mu)}(\theta(h)) \\
&= \int_{\Delta} \theta(h)\mathrm{d}(\Psi(h)(\mu))(\theta) \\
&= \int_{\Delta} \sum_{d=1}^{n} \tau_d(T_d^{-1}(\theta))\theta(h)\mathrm{d}(T_{d*}(\mu))(\theta) \\
&= \sum_{d=1}^{n} \int_{\Delta} \tau_d(\theta)(T_d(\theta))(h)\mathrm{d}\left(T_{d*}(\mu)\right)(T_d(\theta)) \\
&= \sum_{d=1}^{n} \int_{\Delta} \tau_d(\theta)(T_d(\theta))(h)\mathrm{d}\mu(\theta) \\
&= \sum_{d=1}^{n} \int_{\Delta} \frac{T_d(\theta)(h)}{n\theta(h)} T_d(\theta)(h)\mathrm{d}\mu(\theta) \\
&= \int_{\Delta} \sum_{d=1}^{n} \frac{T_d(\theta)(h)^2}{n\theta(h)}\mathrm{d}\mu(\theta)
\end{aligned}$$

In the calculation, the bijectivity of T_d and the formula $(T_{d*}(\mu))(E) = \mu(T_d^{-1}(E))$ is used (and will be used repetitively later).

Consider that by definition of the bijection T_d, the sum $\sum_{d=1}^{n} T_d(\theta)(h) = n\theta(h)$ (T_d is the d-th row of Sinkhorn scaling by column sums $n\theta$). Thus

$$\begin{aligned}
\mathbb{E}_{\Psi(h)(\mu)}(\theta(h)) &= \int_{\Delta} \frac{\sum_{d=1}^{n} T_d(\theta)(h)^2}{\sum_{d=1}^{n} T_d(\theta)(h)}\mathrm{d}\mu(\theta) \\
&\geq \int_{\Delta} \frac{\left(\sum_{d=1}^{n} T_d(\theta)(h)\right)^2}{n\sum_{d=1}^{n} T_d(\theta)(h)}\mathrm{d}\mu(\theta) \\
&= \int_{\Delta} \frac{1}{n}\sum_{d=1}^{n} T_d(\theta)(h)\mathrm{d}\mu(\theta) \\
&= \int_{\Delta} \theta(h)\mathrm{d}\mu(\theta) \\
&= \mathbb{E}_{\mu}(\theta(h)), \qquad (9)
\end{aligned}$$

where $\sum_{d=1}^{n} T_d(\theta)(h)^2 \geq \frac{1}{n}\left(\sum_{d=1}^{n} T_d(\theta)(h)\right)^2$ by Cauchy-Schwarz inequality, with equality achieved if and only if $T_d(\theta)(h)$ is constant on d. Therefore, the equality of Eq. (9) is achieved when $\mathbf{M}^{\langle n\mathbf{x}\rangle}_{(d,h)}$ is constant on d, μ-almost everywhere for $\mathbf{x} \in \Delta^{m-1}$.

□

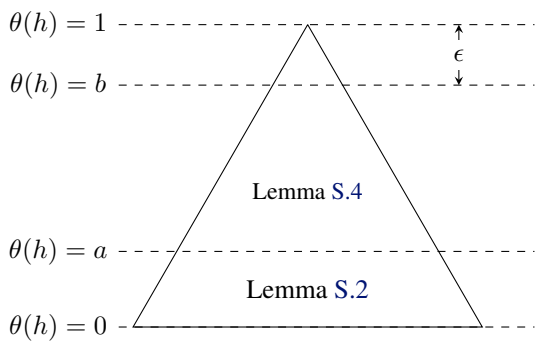

Figure 1. Sketch of Δ^{m-1}, for a general θ, its y-coordinate is $\theta(h)$. The levels are compatible with proof of Theorem 3.5. Lemma S.2 and Lemma S.4 are located where they contribute to prove the vanishing of measure in the limit.

The following lemmas helps showing that the measure μ_k of the complement of a neighborhood of $\delta_h \in \mathcal{P}(\mathcal{H})$ has limit 0.

Lemma S.2. *Given $\mathbf{M}$, $h \in \mathcal{H}$ and prior $\mu_0 \in \mathcal{P}(\Delta^{m-1})$ satisfying assumptions, we have*

$$\mathbb{E}_{\mu_k}\left(\frac{\theta(h')}{\theta(h)}\right) = \mathbb{E}_{\mu_0}\left(\frac{\theta(h')}{\theta(h)}\right) \qquad (10)$$

for any $k \geq 0$ and any $h' \neq h$.

Proof. It suffices to prove the $k = 1$ case for a general μ_0 (then we have the rest by induction).

$$\begin{aligned}
&\mathbb{E}_{\mu_1}\left(\frac{\theta(h')}{\theta(h)}\right) = \int_\Delta \left(\frac{\theta(h')}{\theta(h)}\right) \mathrm{d}\mu_1(\theta) \\
=& \int_\Delta \left(\frac{\theta(h')}{\theta(h)}\right) \mathrm{d}(\Psi(h)(\mu_0))(\theta) \\
=& \int_\Delta \sum_{d\in\mathcal{D}} \tau_d(T_d^{-1}(\theta))\frac{\theta(h')}{\theta(h)}\mathrm{d}(T_{d*}(\mu_0))(\theta) \\
=& \sum_{d\in\mathcal{D}} \int_\Delta \tau_d(\theta)\frac{T_d(\theta)(h')}{T_d(\theta)(h)}\mathrm{d}(T_{d*}(\mu_0))(T_d(\theta)) \\
=& \sum_{d\in\mathcal{D}} \int_\Delta \frac{T_d(\theta)(h)}{n\theta(h)}\frac{T_d(\theta)(h')}{T_d(\theta)(h)}\mathrm{d}(\mu_0)(\theta) \\
=& \int_\Delta \sum_{d\in\mathcal{D}} \frac{T_d(\theta)(h')}{n\theta(h)}\mathrm{d}(\mu_0)(\theta) \\
=& \int_\Delta \frac{\sum_{d\in\mathcal{D}} T_d(\theta)(h')}{n\theta(h)}\mathrm{d}(\mu_0)(\theta) \\
=& \int_\Delta \frac{n\theta(h')}{n\theta(h)}\mathrm{d}(\mu_0)(\theta) \\
=& \mathbb{E}_{\mu_0}\left(\frac{\theta(h')}{\theta(h)}\right). \qquad (11)
\end{aligned}$$

□

Lemma S.3. *The operator $\Psi(h)$ preserves convex combinations of probability measures, i.e., for positive $a_1, a_2, \ldots, a_l$ with $\sum_{i=1}^l a_i = 1$ and probability measures $\mu_1, \mu_2, \ldots, \mu_l$,*

$$\Psi(h)\left(\sum_{i=1}^l a_i\mu_i\right) = \sum_{i=1}^l a_i\Psi(h)(\mu_i).$$

Proof. By definition, for any measurable set E in Borel σ algebra $\mathfrak{A}$,

$$\Psi(h)(\mu)(E) := \int_E \sum_{d=1}^n \tau_d(T_d^{-1}(\theta))\mathrm{d}(T_{d*}(\mu))(\theta).$$

where every summand commutes with convex combination. □

Lemma S.4. *Given $\mathbf{M}$, h, and μ_0 satisfying the assumptions, then for any $0 < a < b < 1$,*

$$\lim_{k\to\infty} \mu_k(\{\theta \in \Delta^{m-1} : a \le \theta(h) \le b\}) = 0 \qquad (12)$$

Proof. We first show a property of μ on the set $\Delta_{[a,b]} := \{\theta \in \Delta^{m-1} : a \le \theta(h) \le b\}$.

For any μ supported on $\Delta_{[a,b]}$ (that is, $\mu(\Delta_{[a,b]}) = 1$), there is a positive number ϵ_0, such that

$$\mathbb{E}_{\Psi^2(h)(\mu)}(\theta(h)) - \mathbb{E}_\mu(\theta(h)) \ge \epsilon_0. \qquad (13)$$

According to the calculation in Lemma S.1, especially the first step of Eq. (9),

$$\begin{aligned}
&\mathbb{E}_{\Psi^2(h)(\mu)}(\theta(h)) \\
=& \int_\Delta \frac{\sum_{d=1}^n T_d(\theta)(h)^2}{\sum_{d=1}^n T_d(\theta)(h)}\mathrm{d}(\Psi(h)(\mu))(\theta) \\
=& \int_\Delta \sum_{e=1}^n \tau_e(T_e^{-1}(\theta))\frac{\sum_{d=1}^n T_d(\theta)(h)^2}{\sum_{d=1}^n T_d(\theta)(h)}\mathrm{d}(T_{e*}(\mu))(\theta) \\
=& \int_\Delta \sum_{e=1}^n \tau_e(\theta)\frac{\sum_{d=1}^n T_d(T_e(\theta))(h)^2}{\sum_{d=1}^n T_d(T_e(\theta))(h)}\mathrm{d}\mu(\theta) \qquad (14)
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}_{\Psi^2(h)(\mu)}(\theta(h)) - \mathbb{E}_\mu(\theta(h)) \\
=& \int_\Delta \sum_{e=1}^n \tau_e(\theta)\frac{\sum_{d=1}^n T_d(T_e(\theta))(h)^2}{\sum_{d=1}^n T_d(T_e(\theta))(h)} - \theta(h)\mathrm{d}\mu(\theta) \qquad (15)
\end{aligned}$$

To show the claim, it suffices to find a positive lower bound of the integrand of Eq. (15), $\mathfrak{I}(\theta) := \sum_{e=1}^n \tau_e(\theta)\frac{\sum_{d=1}^n T_d(T_e(\theta))(h)^2}{\sum_{d=1}^n T_d(T_e(\theta))(h)} - \theta(h)$, for all $\theta \in \Delta_{[a,b]}$. Moreover, since $\Delta_{[a,b]}$ is compact, we just need to show $\mathfrak{I}(\theta) > 0$ on $\Delta_{[a,b]}$.

With Cauchy-Schwarz inequality used in Lemma S.1, we know

$$\begin{aligned}
\mathfrak{I}(\theta) &= \sum_{e=1}^n \tau_e(\theta)\frac{\sum_{d=1}^n T_d(T_e(\theta))(h)^2}{\sum_{d=1}^n T_d(T_e(\theta))(h)} - \theta(h) \\
&\ge \sum_{e=1}^n \tau_e(\theta)\frac{1}{n}\left(\sum_{d=1}^n T_d(T_e(\theta))(h)\right) - \theta(h) \\
&= \sum_{e=1}^n \tau_e(\theta)T_e(\theta)(h) - \theta(h) \\
&= \sum_{e=1}^n \frac{T_e(\theta)(h)}{n\theta(h)}T_e(\theta)(h) - \theta(h) \\
&\ge \frac{1}{n}T_e(\theta)(h) - \theta(h) \\
&= \theta(h) - \theta(h) = 0 \qquad (16)
\end{aligned}$$

$\mathfrak{I}(\theta)$ vanishes if and only if both line 2 and line 5 has equality, and we will discuss why these can not happen simultaneously.

The equality in line 5 requires that $T_e(\theta)(h)$ are identical for all $e \in \mathcal{D}$, or more precisely, the vector $\mathbf{M}^{\langle n\theta\rangle}_{(\_,h)} = t\mathbf{e}_n$ has identical components. Further if equality in line 2 holds, the terms $T_d(T_e(\theta))(h)$ are the same for all $d \in \mathcal{D}$. That is, by condition $\sum_{e=1}^n T_e(\theta)(h) = n\theta(h)$ and $\sum_{d=1}^n T_d(T_e(\theta))(h) = nT_e(\theta)(h)$, $\mathfrak{I}(\theta)$ vanishes if and only if $T_d(T_e(\theta))(h) = T_e(\theta)(h) = \theta(h)$ for all $d, e \in \mathcal{D}$.

We analyze the Sinkhorn scaled matrices in detail: Let $\mathbf{M}^\star = \mathbf{M}^{\langle n\theta\rangle}$ be the scaled matrix whose e-th row is $T_e(\theta)$, and let $\mathbf{M}^{(e)} = \mathbf{M}^{\langle nT_e(\theta)\rangle}$ be the scaled matrix whose d-th row is $T_d(T_e(\theta))$. Since $\mathbf{M}^\star$ and each $\mathbf{M}^{(e)}$ has the same h-th column, there are diagonal matrices $\mathbf{D}^{(e)} = \mathrm{diag}(\frac{n\mathbf{M}^\star_{(e,1)}}{\sum_{i=1}^n \mathbf{M}^\star_{(i,1)}}, \frac{n\mathbf{M}^\star_{(e,2)}}{\sum_{i=1}^n \mathbf{M}^\star_{(i,2)}}, \dots, \frac{n\mathbf{M}^\star_{(e,n)}}{\sum_{i=1}^n \mathbf{M}^\star_{(i,n)}})$ such that $\mathbf{M}^{(e)} = \mathbf{M}^\star\mathbf{D}^{(e)}$. Since $\mathbf{M}^\star$ and all $\mathbf{M}^{(e)}$ are row-normalized to $\mathbf{e}$ (i.e., their row sums are 1), we have the following equations from the row sums:

$$\mathfrak{S}(d,e) := \sum_{j=1}^m \mathbf{M}^\star_{(d,j)} \frac{n\mathbf{M}^\star_{(e,j)}}{\sum_{i=1}^n \mathbf{M}^\star_{(i,j)}} = 1 \tag{17}$$

for all $d, e \in \mathcal{D}$ representing the d-th row-sum of $\mathbf{M}^{(e)}$.

Then we calculate $(n-1)\sum_{e=1}^n \mathfrak{S}(e,e) - \sum_{d\neq e} \mathfrak{S}(d,e)$. On the right hand side, since $\mathfrak{S}(d,e) = 1$ for every d, e, we have

$$(n-1)\sum_{e=1}^n \mathfrak{S}(e,e) - \sum_{d\neq e} \mathfrak{S}(d,e) = (n-1)n - (n^2-n) = 0.$$

Meanwhile,

$$\begin{aligned}
&(n-1)\sum_{e=1}^n \mathfrak{S}(e,e) - \sum_{d\neq e} \mathfrak{S}(d,e) \\
&= \sum_{j=1}^m \frac{n}{\sum_{i=1}^n \mathbf{M}^\star_{(i,j)}} \left(\sum_{e=1}^n (n-1)(\mathbf{M}^\star_{(e,j)})^2 \right. \\
&\qquad \left. - \sum_{d\neq e} \mathbf{M}^\star_{(d,j)}\mathbf{M}^\star_{(e,j)} \right) \\
&= \sum_{j=1}^m \frac{n}{\sum_{i=1}^n \mathbf{M}^\star_{(i,j)}} \left(\sum_{d<e} (\mathbf{M}^\star_{(d,j)} - \mathbf{M}^\star_{(e,j)})^2 \right) \\
&= 0
\end{aligned} \tag{18}$$

Therefore, $\mathbf{M}^\star_{(d,j)} = \mathbf{M}^\star_{(e,j)}$ for any d, e, and j. Therefore, the rows of $\mathbf{M}^\star$ are identical, so the columns of $\mathbf{M}^\star$ are all parallel (or say, collinear as vectors, i.e. one is a scalar-multiple of the other) to each other.

By Sinkhorn scaling theory (Fienberg et al., 1970), the cross-ratios are invariant. Since $\mathbf{M}$ is a positive matrix and has distinct (non-parallel) columns, the 2×2 cross-ratios are not identically 1, however, $\mathbf{M}^\star$ — a scaled matrix of $\mathbf{M}$ — has cross ratios identically 1. Therefore our assumption that $\mathfrak{I}(\theta) = 0$ cannot happen, and by compactness of $\Delta_{[a,b]}$ and continuity of $\mathfrak{I}(\theta)$, we can conclude that $\mathfrak{I}(\theta)$ has a lower bound $\epsilon_0 > 0$ on $\Delta_{[a,b]}$.

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\Psi^2(h)(\mu)}(\theta(h)) - \mathbb{E}_\mu(\theta(h)) \\
&= \int_\Delta \mathfrak{I}(\theta)\mathrm{d}\mu(\theta) \\
&\geq \int_\Delta \epsilon_0 \mathrm{d}\mu(\theta) \\
&= \epsilon_0\mu(\Delta) = \epsilon_0.
\end{aligned} \tag{19}$$

Thus we prove the property Eq. (13).

We prove the lemma by contradiction:

Suppose the limit does not exist or the limit is nonzero. In either case, there exists a positive real number $\epsilon > 0$, such that there are infinitely many integers, or say a sequence (k_i) such that

$$\mu_{k_i}(\{a \leq \theta(h) \leq b\}) > \epsilon.$$

We may assume k_i contains no consecutive elements, i.e., $k_{i+1} - k_i > 1$ for all i, otherwise, we can always find a subsequence satisfying this (for example, choose the sequence of all odd or even k_i's, at least one of them is infinite, so we have a sequence).

For a μ-measurable set E, let $\mu|_E$ be the restriction of μ on E, which can be treated as a measure on Δ by setting the measure of the complement E^c zero (but the measure of Δ is no longer 1). We scale it to $\widehat{\mu}|_E := (\mu(E))^{-1}\mu|_E$ to make it a probability measure, then $\mu_{k_i} = [\mu_{k_i}(\Delta_{[a,b]})]\widehat{\mu}_{k_i}|_{\Delta_{[a,b]}} + [1 - \mu_{k_i}(\Delta_{[a,b]})]\widehat{\mu}_{k_i}|_{(\Delta-\Delta_{[a,b]})}$. Thus according to Lemma S.3,

$$\begin{aligned}
&\mathbb{E}_{\Psi^2(h)(\mu_{k_i})}(\theta(h)) \\
=\ &\mu_{k_i}(\Delta_{[a,b]})\mathbb{E}_{\Psi^2(h)(\widehat{\mu}_{k_i}|_{\Delta_{[a,b]}})}(\theta(h)) \\
&+(1-\mu_{k_i}(\Delta_{[a,b]}))\mathbb{E}_{\Psi^2(h)(\widehat{\mu}_{k_i}|_{(\Delta-\Delta_{[a,b]})})}(\theta(h)) \\
\geq\ &\mu_{k_i}(\Delta_{[a,b]})(\mathbb{E}_{\widehat{\mu}_{k_i}|_{\Delta_{[a,b]}}}(\theta(h)) + \epsilon_0) \\
&+(1-\mu_{k_i}(\Delta_{[a,b]}))\mathbb{E}_{\widehat{\mu}_{k_i}|_{(\Delta-\Delta_{[a,b]})}}(\theta(h)) \\
\geq\ &\epsilon\epsilon_0 + \mathbb{E}_{\mu_{k_i}}(\theta(h))
\end{aligned} \tag{20}$$

By Lemma S.1, we can see that $\mathbb{E}_{\mu_{k+1}}[\theta(h)] \geq \mathbb{E}_{\mu_k}[\theta(h)]$, and there is a sequence (k_i) such that $\mathbb{E}_{\mu_{k_i+2}}[\theta(h)] \geq \mathbb{E}_{\mu_{k_i}}[\theta(h)] + \epsilon_0\epsilon$. Thus $\mathbb{E}_{\mu_{k_i+2}}[\theta(h)] \geq \mathbb{E}_{\mu_0}[\theta(h)] + i\epsilon_0\epsilon$, so $\lim_{k\to\infty} \mathbb{E}_{\mu_k}[\theta(h)] = \infty$.

However, $\theta(h) \leq 1$, we have $\mathbb{E}_{\mu_k}(\theta(h)) \leq 1$ for all k, which is a contradiction. Therefore, we know that Eq. (12) holds. □

Theorem 3.5. *In SCBI, let* $\mathbf{M}$ *be a positive matrix. If the teacher is teaching one hypothesis* h *(i.e.,* $\widehat{\theta} = \delta_h \in \mathcal{P}(\mathcal{H})$*), and the prior distribution* $\mu_0 \in \mathcal{P}(\Delta^{m-1})$ *satisfies* $\mu_0 = \delta_{\theta_0}$ *with* $\theta_0(h) > 0$*, then the estimator sequence* (S_k) *is consistent, for each* $h \in \mathcal{H}$*, i.e., the posterior random variables* $(\Theta_k)_{k\in\mathbb{N}}$ *converge to the constant random variable* $\widehat{\theta}$ *in probability.*

Some notions used in the proof are visualized in Fig. 1.

Proof. Let Z_0 be a random variable with sample space Δ^{m-1} such that the $\mathrm{Law}(Z_0) = \mu_0$. This is the initial state in SCBI. The posteriors in the following rounds are determined by the sequence of data taught by teacher, which makes the posteriors random variables as well. Let Z_k be the random variable representing the posterior after k-rounds of SCBI, the law of Z_k is given by $\mathrm{Law}(Z_k) = \mu_k = [\Psi(h)]^k(\mu_0)$ according to the definition of $\Psi(h)$.

The consistency mentioned in the theorem is equivalent to that the sequence (Z_k) converges to $\widehat{Z}$ with $\mathrm{Law}(\widehat{Z}) = \widehat{\mu}$ in probability where $\widehat{\mu} = \delta_{\widehat{\theta}}$.

We prove the theorem by contradiction. Suppose $Z_k \to \widehat{Z}$ in probability is not valid, i.e., there exists $\epsilon > 0$ such that

$$\lim_{k\to\infty} \Pr(d(Z_k, \widehat{Z}) > \epsilon) \tag{21}$$

does not exist or the limit is positive, where the metric d on Δ^{m-1} is the Euclidean distance inherited from $\mathbb{R}^m$. In either case, there is a real number $C > 0$ such that

$$\Pr(d(Z_{k'}, \widehat{Z}) > \epsilon) > C \tag{22}$$

for a subsequence $(Z_{k'})$ of (Z_k).

Let $R := \mathbb{E}_{\mu_0}\left[\frac{1-\theta(h)}{\theta(h)}\right] = \frac{1-\theta_0(h)}{\theta_0(h)}$, let $a = \frac{1}{4R/C+1}$ and $b = 1-\epsilon$. By Lemma S.4, there exists $N > 0$ such that for all $k > N$,

$$\mu_k(\Delta_{[a,b]}) < C/2.$$

Therefore, for all the terms in (k') satisfying $k' > N$, $\mu_{k'}(\{\theta : \theta(h) < a\}) > C/2$. Furthermore,

$$\begin{aligned}
&\mathbb{E}_{\mu_{k'}}\left[\frac{1-\theta(h)}{\theta(h)}\right] \\
\geq\ & \int_{\{\theta:\theta(h)<a\}} \left[\frac{1-\theta(h)}{\theta(h)}\right] \mathrm{d}\mu_{k'}(\theta) \\
\geq\ & \left[\frac{1-a}{a}\right]\frac{C}{2} \\
=\ & \left[\frac{4R}{C}\right]\frac{C}{2} \\
=\ & 2R > R.
\end{aligned} \tag{23}$$

However, by Lemma S.2,

$$\begin{aligned}
&\mathbb{E}_{\mu_{k'}}\left[\frac{1-\theta(h)}{\theta(h)}\right] \\
=\ & \mathbb{E}_{\mu_{k'}}\left[\frac{\sum_{h'\neq h}\theta(h')}{\theta(h)}\right] \\
=\ & \sum_{h'\neq h}\mathbb{E}_{\mu_{k'}}\left[\frac{\theta(h')}{\theta(h)}\right] \\
=\ & \sum_{h'\neq h}\mathbb{E}_{\mu_0}\left[\frac{\theta(h')}{\theta(h)}\right] \\
=\ & \mathbb{E}_{\mu_0}\left[\frac{\sum_{h'\neq h}\theta(h')}{\theta(h)}\right] \\
=\ & \mathbb{E}_{\mu_0}\left[\frac{1-\theta(h)}{\theta(h)}\right] \\
=\ & R,
\end{aligned} \tag{24}$$

which is a contradiction. Therefore,

$$\lim_{k\to\infty} \Pr(d(Z_k, \widehat{Z}) > \epsilon) = 0. \tag{25}$$

And the sequence of SCBI estimators is consistent at $\widehat{\theta}$.

□

Proof of Theorem 3.6

Theorem 3.6. *With matrix* $\mathbf{M}$*, hypothesis* $h \in \mathcal{H}$*, and a prior* $\mu_0 = \delta_{\theta_0} \in \mathcal{P}(\Delta^{m-1})$ *same as in Theorem. 3.5, let* θ_k *denote a sample value of the posterior* Θ_k *after* k *rounds of SCBI, then*

$$\lim_{k\to\infty} \mathbb{E}_{\mu_k}\left[\frac{1}{k}\log\left(\frac{\theta_k(h)}{1-\theta_k(h)}\right)\right] = \mathfrak{R}^s(\mathbf{M}; h) \tag{26}$$

where $\mathfrak{R}^s(\mathbf{M}; h) := \min_{h\neq h'} \mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)}, \mathbf{M}^\sharp_{(\_,h')}\right)$ *with* $\mathbf{M}^\sharp = \mathscr{N}_{col}(\mathrm{diag}(\mathbf{M}_{(\_,h)})^{-1}\mathbf{M})$*. Thus we call* $\mathfrak{R}^s(\mathbf{M}; h)$ *the asymptotic rate of convergence (RoC) of SCBI.*

Proof. We treat θ_k as random variables, then

$$\mathbb{E}_{\mu_{k+1}}\left(\log\left[\frac{\theta_{k+1}(h')}{\theta_{k+1}(h)}\right]\right) = \mathbb{E}_{\mu_k}\left(\log\left[\frac{\theta_k(h')}{\theta_k(h)}\right]\right) + W_k^{h'},$$

where

$$W_k^{h'} = -\mathbb{E}_{\mu_k}\left[\mathrm{KL}\left(\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta_k\rangle}_{(\_,h)}), \mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta_k\rangle}_{(\_,h')})\right)\right].$$

We can get it from the following calculation (Δ represents

the simplex Δ^{m-1}):

$$\begin{aligned}
&\mathbb{E}_{\mu_{k+1}}\left(\log\left[\frac{\theta_{k+1}(h')}{\theta_{k+1}(h)}\right]\right)\\
=&\int_\Delta \log\left[\frac{\theta(h')}{\theta(h)}\right]\mathrm{d}\mu_{k+1}(\theta)\\
=&\int_\Delta \sum_{d=1}^n \tau_d(T_d^{-1}(\theta))\log\left[\frac{\theta(h')}{\theta(h)}\right]\mathrm{d}(T_{d*}(\mu_{k+1}))(\theta)\\
=&\sum_{d=1}^n\int_\Delta \tau_d(\theta)\log\left[\frac{T_d(\theta)(h')}{T_d(\theta)(h)}\right]\mathrm{d}(\mu_k)(\theta)\\
=&\int_\Delta \sum_{d=1}^n \frac{T_d(\theta)(h)}{n\theta(h)}\log\left[\frac{T_d(\theta)(h')}{T_d(\theta)(h)}\right]\mathrm{d}(\mu_k)(\theta)\\
=&\int_\Delta \sum_{d=1}^n \frac{T_d(\theta)(h)}{n\theta(h)}\left\{\log\left[\frac{T_d(\theta)(h')}{n\theta(h')}\frac{n\theta(h)}{T_d(\theta)(h)}\right]\right.\\
&\left.+\log\left[\frac{n\theta(h')}{n\theta(h)}\right]\right\}\mathrm{d}(\mu_k)(\theta)\\
=&\int_\Delta -\mathrm{KL}\left(\mathscr{N}(\mathbf{M}^{\langle n\theta\rangle}_{(_,h)}),\mathscr{N}(\mathbf{M}^{\langle n\theta\rangle}_{(_,h')})\right)\mathrm{d}\mu_k\\
&+\int_\Delta \log\left[\frac{\theta(h')}{\theta(h)}\right]\mathrm{d}\mu_k\\
=&\ W_k^{h'}+\mathbb{E}_{\mu_k}\left(\log\left[\frac{\theta_k(h')}{\theta_k(h)}\right]\right).
\end{aligned}\tag{27}$$

Next, we show

$$\lim_{k\to\infty}\mathbb{E}_{\mu_k}\frac{1}{k}\log\left(\frac{\theta_k(h')}{\theta_k(h)}\right)=-\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right),\tag{28}$$

and then by a similar argument in the proof of Theorem 3.2, we can show the result in this theorem.

To show Eq. (28), we can make use of Eq. (27). By showing that $W_k^{h'}$ converges to $-\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)$, we can conclude that $\mathbb{E}_{\mu_k}\frac{1}{k}\log\left(\frac{\theta_k(h')}{\theta_k(h)}\right)$, as the average of $(W_i^{h'})$ on the first k-terms, converges to $-\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)$ as well.

To prove this, we need the following result from direct calculation:

Lemma S.5. *Given a $n\times 2$ positive matrix $[\mathbf{a},\mathbf{b}]$ with columns as n-vectors $\mathbf{a}=(a_1,a_2,\dots,a_n)^\top$ and $\mathbf{b}=(b_1,b_2,\dots,b_n)^\top$ with $\sum_{i=1}^n a_i=\sum_{i=1}^n b_i=1$, consider the 2×2 cross-ratios: $C_i:=CR(1,2;1,i)=\frac{a_1b_i}{a_ib_1}$, then $\mathrm{KL}(\mathbf{a},\mathbf{b})=\log\left(\sum_{i=1}^n a_iC_i\right)-\sum_{i=1}^n a_i\log C_i$. With fixed $C_i\in(0,\infty)$ for $i=1,2,\dots,n$, $\mathrm{KL}(\mathbf{a},\mathbf{b})$ is continuous and bounded about $\mathbf{a}\in\Delta^{n-1}$.*

Proof of Lemma S.5. The formula of $\mathrm{KL}(\mathbf{a},\mathbf{b})$ is from direct calculation.

The KL-divergence is continuous and bounded since by the formula, every part is continuous and bounded given the restrictions on $\mathbf{a}$ and C_i. □

Now we continue to prove Theorem 3.6:

By continuity of the KL-divergence given fixed cross-ratios, for any $\epsilon>0$, we find a number $\delta>0$ such that for any $\theta\in\Delta^{m-1}$ with $\theta(h)>1-\delta$,

$$\left|\mathrm{KL}\left(\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h)}),\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h')})\right)-\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)\right|<\frac{\epsilon}{3}.\tag{29}$$

Further, according to Theorem 3.5, and the boundedness from Lemma S.5, we can find a number $N>0$, such that for any $k>N$, we have $\mu_k(\{\theta:\theta(h)<1-\delta\})<C$ where C satisfies

$$C\cdot\sup_{\theta\in\Delta^{m-1}}\left\{\mathrm{KL}\left(\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h)}),\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h')})\right)\right\}<\frac{\epsilon}{3}.\tag{30}$$

The expectation $W_k^{h'}$ can be split into two parts, $W_k^{h'}=-W_>-W_<$ where

$$W_>=\int_{\theta(h)>1-\delta}\mathrm{KL}\left(\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h)}),\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h')})\right)\mathrm{d}\mu_k(\theta),\tag{31}$$

and

$$W_<=\int_{\theta(h)\le 1-\delta}\mathrm{KL}\left(\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h)}),\mathscr{N}_{\mathrm{vec}}(\mathbf{M}^{\langle n\theta\rangle}_{(\_,h')})\right)\mathrm{d}\mu_k(\theta).\tag{32}$$

Similarly, since μ_k is a probability measure, $\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)=K_>+K_<$ where

$$K_>=\int_{\theta(h)>1-\delta}\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)\mathrm{d}\mu_k(\theta),\tag{33}$$

and

$$K_<=\int_{\theta(h)\le 1-\delta}\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)\mathrm{d}\mu_k(\theta).\tag{34}$$

Then we have

$$\left|W_k^{h'}+\mathrm{KL}\left(\mathbf{M}^\sharp_{(\_,h)},\mathbf{M}^\sharp_{(\_,h')}\right)\right|\le|K_>-W_>|+|W_<|+|K_<|.\tag{35}$$

The choice of δ can make a good estimate of the integral on $\theta(h)>1-\delta$.

$$\begin{aligned}
&|K_>-W_>|\\
\le&\ \frac{\epsilon}{3}(1-C)\\
<&\ \frac{\epsilon}{3}.
\end{aligned}\tag{36}$$

For the other two terms, directly from condition Eq. (30), we have $|K_<|<\frac{\epsilon}{3}$ and $|W_<|<\frac{\epsilon}{3}$, and hence $|K_>-W_>|+|K_<|+|W_<|<\epsilon$.

Therefore, $W_k^{h'}$ converges to $-\mathrm{KL}\left(\mathbf{M}^{\sharp}_{(\_,h)}, \mathbf{M}^{\sharp}_{(\_,h')}\right)$.

□

An Example on a 2 by 2 Matrix

Let $\mathcal{H} = \{h_1, h_2\}$, $\mathcal{D} = \{d_1, d_2\}$, and the shared joint distribution be $\mathbf{M}^{JD} = \begin{matrix} \\ d_1 \\ d_2 \end{matrix} \begin{matrix} h_1 & h_2 \\ \end{matrix} \begin{pmatrix} 0.3 & 0.3 \\ 0.1 & 0.3 \end{pmatrix}$. Further assume that the learner has uniform prior on $\mathcal{H}$, i.e. $S_0 = \theta_0 = (0.5, 0.5)$ and the true hypothesis given to the teachers is h_1. In round 1, the BI teacher will sample a data from $\mathcal{D}$ according to the first column of $\mathbf{M} = \begin{pmatrix} 0.75 & 0.5 \\ 0.25 & 0.5 \end{pmatrix}$, which is obtained by column normalizing $\mathbf{M}^{JD}$. On the contrast, the SCBI teacher will form his likelihood matrix by first doing $(\mathbf{r}_1, \mathbf{c}_1)$-Sinkhorn scaling on $\mathbf{M}$, then column normalization if needed, where $\mathbf{r}_1 = (1, 1)$ and $\mathbf{c}_1 = (1, 1)$ based on the uniform priors. The resulting limit matrix (with precision of three decimals) is $\mathbf{M}_1^* = \begin{pmatrix} 0.634 & 0.366 \\ 0.366 & 0.634 \end{pmatrix}$, which is already column normalized. Hence the SCBI teacher will teach according to the first column of the $\mathbf{M}_1 = \mathbf{M}_1^*$. Suppose that d_1 is sampled by both teachers. The posterior for the BI learner is $S_1^{\mathfrak{b}}(d_1) = (0.6, 0.4)$ (normalizing the first row of $\mathbf{M}$). The posterior for the SCBI learner is $S_1^{\mathfrak{s}}(d_1) = (0.643, 0.366)$ (the first row of $\mathbf{M}_1$).

Similarly, in round 2, the SCBI teacher would update his likelihood matrix by first doing $(\mathbf{r}_2, \mathbf{c}_2)$-Sinkhorn scaling on $\mathbf{M}_1$, where $\mathbf{r}_2 = (1, 1)$ and $\mathbf{c}_2 = (0.643, 0.366) \times 2 = (1.268, 0.732)$. The resulting limit matrix is $\mathbf{M}_2^* = \begin{pmatrix} 0.758 & 0.242 \\ 0.51 & 0.49 \end{pmatrix}$. Then through column normalizing $\mathbf{M}_2^*$, a updated likelihood matrix $\mathbf{M}_2 = \begin{pmatrix} 0.60 & 0.33 \\ 0.4 & 0.67 \end{pmatrix}$ is obtained. The SCBI teacher will teach according to the first column of the $\mathbf{M}_2$. Whereas the BI teacher will again sample another data according to the the first column of $\mathbf{M}$. Suppose that d_1 is sampled for both teachers. The posteriors for BI and SCBI learners are $S_2^{\mathfrak{b}}(d_1, d_1) = (0.692, 0.308)$ and $S_1^{\mathfrak{s}}(d_1, d_1) = (0.758, 0.242)$ respectively.

Although same teaching points are assumed, the SCBI learner's posterior on the true hypothesis h_1 is higher than the BI learner in both rounds. Moreover, notice that the KL divergence between h_1 and h_2 is increasing as the likelihood matrix is updating through the SCBI. This will eventually lead much faster convergence for the SCBI learner.

Calculations about Sample Efficiency

Here we compute the expectation $\mathfrak{E}$ mentioned in Sec. 4.1 of the paper, for matrices of size $n \times 2$.

We first calculate the average of RoC for a particular matrix $\mathbf{M}$. For simplicity, let $\mathbf{x} = \mathbf{M}_{(\_,1)}$ and $\mathbf{y} = \mathbf{M}_{(\_,2)}$.

$$\begin{aligned} & \frac{1}{2}\sum_{h=1}^{2} \mathfrak{R}^{\mathfrak{b}}(\mathbf{M}; h) \\ = \; & \frac{1}{2}\left(\mathrm{KL}(\mathbf{M}_{(\_,1)}, \mathbf{M}_{(\_,2)}) + \mathrm{KL}(\mathbf{M}_{(\_,2)}, \mathbf{M}_{(\_,1)})\right) \\ = \; & \frac{1}{2}\left(\mathrm{KL}(\mathbf{x}, \mathbf{y}) + \mathrm{KL}(\mathbf{y}, \mathbf{x})\right) \\ = \; & \frac{1}{2}\left(\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{y}_i)(\ln \mathbf{x}_i - \ln \mathbf{y}_i)\right). \end{aligned} \tag{37}$$

To calculate that for SCBI, denote $\mathbf{x}/\mathbf{y} = \mathscr{N}_{vec}(\mathbf{v})$ the normalization of vector $\mathbf{v}$ with$\mathbf{v}_i = \mathbf{x}_i/\mathbf{y}_i$.

$$\begin{aligned} & \frac{1}{2}\sum_{h=1}^{2} \mathfrak{R}^{\mathfrak{s}}(\mathbf{M}; h) \\ = \; & \frac{1}{2}\left[\mathrm{KL}\left(\frac{\mathbf{e}}{n}, \mathbf{x}/\mathbf{y}\right) + \mathrm{KL}\left(\frac{\mathbf{e}}{n}, \mathbf{y}/\mathbf{x}\right)\right] \\ = \; & \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\left(-2\ln n - \ln\frac{\mathbf{x}_i}{\mathbf{y}_i} + \ln\left(\sum_{j=1}^{n}\frac{\mathbf{x}_j}{\mathbf{y}_j}\right)\right.\right. \\ & \left.\left. - \ln\frac{\mathbf{y}_i}{\mathbf{x}_i} + \ln\left(\sum_{j=1}^{n}\frac{\mathbf{y}_j}{\mathbf{x}_j}\right)\right)\right] \\ = \; & \frac{1}{2}\left[\ln\left(\sum_{j=1}^{n}\frac{\mathbf{x}_j}{\mathbf{y}_j}\right) + \ln\left(\sum_{j=1}^{n}\frac{\mathbf{y}_j}{\mathbf{x}_j}\right)\right] - \ln n. \end{aligned} \tag{38}$$

The simulation of $\mathfrak{P}$ is based on the above calculations. For $\mathfrak{E}$, the above expressions can be further simplified.

Given $\mathbf{M} = (\mathbf{x}, \mathbf{y})$ uniformly distributed in $(\Delta^{n-1})^2$, with measure $\nu \otimes \nu$ where ν is the measure of uniform probability distribution on Δ^{n-1}, we can calculate the expected value,

$$\begin{aligned} & \mathbb{E}\left[\frac{1}{2}\sum_{h=1}^{2} \mathfrak{R}^{\mathfrak{b}}(\mathbf{M}; h)\right] \\ = \; & \frac{1}{2}\int_{(\Delta^{n-1})^2} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{y}_i)(\ln \mathbf{x}_i - \ln \mathbf{y}_i)\mathrm{d}\nu(\mathbf{x})\mathrm{d}\nu(\mathbf{y}) \\ = \; & n\int_{\Delta^{n-1}} \mathbf{x}_1 \ln \mathbf{x}_1 \mathrm{d}\nu(\mathbf{x}) - n\int_{(\Delta^{n-1})^2} \mathbf{x}_1 \ln \mathbf{y}_1 \mathrm{d}\nu(\mathbf{x})\mathrm{d}\nu(\mathbf{y}) \\ = \; & n\int_0^1 x(n-1)(1-x)^{n-2}\ln x \mathrm{d}x + \\ & n\int_0^1\int_0^1 x(n-1)^2(1-x)^{n-2}(1-y)^{n-2}\ln y \mathrm{d}x\mathrm{d}y \\ = \; & \frac{n-1}{n}. \end{aligned} \tag{39}$$

Here we use the fact that

$$\int_{\{\theta\in\Delta^{n-1}:\theta(h)=a\}} \mathrm{d}\mathbf{x} = (n-1)(1-a)^{n-2}.$$

Furthermore, since integral on $(\Delta^{n-1})^2$ with measure $\nu\otimes\nu$ is symmetric on $\mathbf{x}$ and $\mathbf{y}$, we have

$$\mathfrak{E} = \int_{(\Delta^{n-1})^2} \ln\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i}{\mathbf{y}_i}\right)\mathrm{d}\nu(\mathbf{x})\mathrm{d}\nu(\mathbf{y}) - \ln n - \frac{n-1}{n}. \quad (40)$$

In general, we calculate the integral in Eq. (40) by Monte Carlo method since other numerical integral methods we tried becomes slow dramatically as n grows. In particular, when $n=2$, an expression related to the dilogarithm Li_2 can be obtained (can be easily checked in Wolfram software).

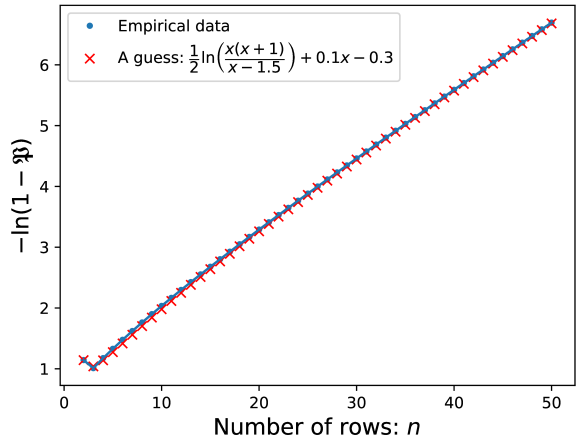

Figure 2. A guess of the $\mathfrak{P}$ values for each n together with empirical data

And we have an empirical formula to describe the relation between $\mathfrak{P}$ and n, shown in Fig. 2:

$$\mathfrak{P}_{(n,2)} = 1 - \sqrt{\frac{x-1.5}{x(x+1)}}e^{-0.1x+0.3} \quad (41)$$

Proof of Proposition 5.1

Proposition 5.1. *Given a sequence of identical independent $\mathcal{D}$-valued random variables $(\mathrm{D}_i)_{i\geq 1}$ following uniform distribution. Let $\mu_0\in\mathcal{P}(\Delta^{m-1})$ be a prior distribution on Δ^{m-1}, and $\mu_{k+1}=\Psi^{\mathbf{L}}_{\mathrm{D}_k}(\mu_k)$, then μ_k converges, in probability, to $\sum_{i\in\mathcal{H}} a_i\delta_i$, where $a_i=\mathbb{E}_{\mu_0}[\theta(i)]$.*

To show the above proposition, we need the following lemma:

Lemma S.6. *Given the conditions in Proposition 5.1, then for any $k\in\mathbb{N}$ and $h\in\mathcal{H}$,*

$$\mathbb{E}_{\mu_k}(\theta(h)) = \mathbb{E}_{\mu_0}(\theta(h)). \quad (42)$$

Proof. It suffices to show $\mathbb{E}_{\mu_{k+1}}(\theta(h)) = \mathbb{E}_{\mu_k}(\theta(h))$ for any k.

$$\begin{aligned}
\mathbb{E}_{\mu_{k+1}}(\theta(h)) &= \int_{\Delta^{m-1}}\theta(h)\mathrm{d}(\mu_{k+1})(\theta)\\
&= \sum_{d\in\mathcal{D}}\mathrm{D_k}(d)\int_{\Delta^{m-1}}T_d(\theta)(h)\mathrm{d}\mu_k(\theta)\\
&= \int_{\Delta^{m-1}}\frac{1}{n}\sum_{d\in\mathcal{D}}T_d(\theta)(h)\mathrm{d}(\mu_k)(\theta)\\
&= \int_{\Delta^{m-1}}\theta(h)\mathrm{d}(\mu_k)(\theta) = \mathbb{E}_{\mu_k}(\theta(h)).
\end{aligned}$$

□

Proof of Proposition 5.1. We first show the following result:

For any $\epsilon>0$, let

$$\Delta_\epsilon := \left\{\theta\in\Delta^{m-1} : \theta(i)\leq 1-\epsilon, \forall i=1,2,\ldots,m\right\},$$

then $\lim_{k\to\infty}\mu_k(\Delta_\epsilon)=0$.

We prove this by contradiction. Suppose the limit does not exist or is not 0, then there is a positive number C and a subsequence $(\mu_{k_i})_{i\in\mathbb{N}}$ such that $\mu_{k_i}(\Delta_\epsilon)>C$ for all i.

We define a linear functional $\mathscr{L}(\mu) := \mathbb{E}_\mu f(\theta)$, where $f(\theta)=||\theta-u||_2^2$ with $u=\frac{\mathbf{e}}{m}$ the center of Δ^{m-1}.

By definition, for a random variable following uniform distribution on $\mathcal{D}$, $\mathscr{L}\left(\Psi^{\mathbf{L}}_{\mathrm{D}}(\mu)\right)=\mathbb{E}_\mu\left(\mathbb{E}_{d\sim\mathrm{D}}f(T_d(\theta))\right)$.

Consider that f is a strictly convex function, by Jensen's inequality, $\mathbb{E}_{d\sim\mathrm{D}}f(T_d(\theta))\geq f(\mathbb{E}_{d\sim\mathrm{D}}T_d(\theta))=f(\theta)$, with equality if and only if $T_d(\theta)=\theta$ for all $d\in\mathcal{D}$, equivalently by the assumptions on matrix $\mathbf{L}$, $\theta=\delta_h$ for some $h\in\mathcal{H}$. (This is because we assume $\mathbf{L}$ have distinct columns, thus not all 2-by-2 cross-ratios are 1, for any pair of columns. however, after T_d all 2-by-2 cross-ratios are 1, indicating the existence of degeneration on every pair of columns. This can only happen when $\theta=\delta_h$ for some $h\in\mathcal{H}$.)

Thus for any $\theta\in\Delta_\epsilon$, $\mathbb{E}_{d\sim\mathrm{D}}f(T_d(\theta))>f(\theta)$. As Δ_ϵ is compact, there is a lower bound $B>0$, such that

$\mathbb{E}_{d\sim \mathrm{D}} f(T_d(\theta)) - f(\theta) > B$ for all $\theta \in \Delta_\epsilon$. Thus

$$\begin{aligned}
&\mathscr{L}(\mu_{k_i+1}) \\
=& \int_{\Delta^{m-1}\setminus\Delta_\epsilon} \mathbb{E}_{d\sim\mathcal{D}} f(T_d(\theta))\mathrm{d}\mu_{k_i} + \int_{\Delta_\epsilon} \mathbb{E}_{d\sim\mathcal{D}} f(T_d(\theta))\mathrm{d}\mu_{k_i} \\
>& \int_{\Delta^{m-1}\setminus\Delta_\epsilon} f(\theta)\mathrm{d}\mu_{k_i} + \int_{\Delta_\epsilon} f(\theta)\mathrm{d}\mu_{k_i} + BC \\
=& \mathscr{L}(\mu_{k_i}) + BC
\end{aligned}$$

for all $i \in \mathbb{N}$ and simply $\mathscr{L}(\mu_{k+1}) \geq \mathscr{L}(\mu_k)$ for general k.

Therefore, $\mathscr{L}(\mu_k)$ is unbounded as $k \to \infty$ since there is at least a $BC > 0$ increment at each k_i.

However, by definition, f is bounded by m since $\sqrt{m}$ is the diameter of Δ^{m-1} under 2-norm, thus $\mathscr{L}(\mu) \leq m$.

Such a contradiction shows that the opposite of our assumption, $\lim_{k\to\infty} \mu_k(\Delta_\epsilon) = 0$, is valid.

Consider that ϵ is arbitrary, and in Lemma S.6 we show that $\mathbb{E}_{\mu_k}\theta(h)$ is invariant, thus μ_k approaches $\sum_{i\in\mathcal{H}} a_i\delta_i$ with $a_i = \mathbb{E}_{\mu_0}\theta(i)$ in probability. □

Empirical Data for Stability: Perturbation on Prior

We sample 5 matrices of size 3×3, each of them are column-normalized, and their columns are sampled independently and uniformly on Δ^2, listed below:

$$\begin{aligned}
\mathbf{M}_1 &= \begin{pmatrix} 0.6559 & 0.5505 & 0.7310 \\ 0.1680 & 0.3359 & 0.0403 \\ 0.1760 & 0.1136 & 0.2287 \end{pmatrix} \\
\mathbf{M}_2 &= \begin{pmatrix} 0.2461 & 0.6600 & 0.4310 \\ 0.6785 & 0.0655 & 0.2325 \\ 0.0754 & 0.2746 & 0.3365 \end{pmatrix} \\
\mathbf{M}_3 &= \begin{pmatrix} 0.7286 & 0.1937 & 0.7620 \\ 0.0739 & 0.4786 & 0.1999 \\ 0.1974 & 0.3277 & 0.0382 \end{pmatrix} \\
\mathbf{M}_4 &= \begin{pmatrix} 0.4745 & 0.2024 & 0.5946 \\ 0.2898 & 0.7499 & 0.1313 \\ 0.2357 & 0.0477 & 0.2741 \end{pmatrix} \\
\mathbf{M}_5 &= \begin{pmatrix} 0.2207 & 0.5466 & 0.1605 \\ 0.3828 & 0.3807 & 0.5697 \\ 0.3965 & 0.0727 & 0.2698 \end{pmatrix}
\end{aligned} \tag{43}$$

And the 5 sampled priors are:

$$\begin{aligned}
\theta_1 &= (0.3333, 0.3333, 0.3333)^\top \\
\theta_2 &= (0.1937, 0.4291, 0.3771)^\top \\
\theta_3 &= (0.4544, 0.0814, 0.4641)^\top \\
\theta_4 &= (0.5955, 0.2995, 0.1051)^\top \\
\theta_5 &= (0.4771, 0.0593, 0.4636)^\top
\end{aligned} \tag{44}$$

These names (including the rank 4 samples below) are overriding the previously defined identical symbols in this part and in the corresponding subsection in the main paper. In the 4×4 cases, we sample 3 matrices in the same way as in 3×3 case.

$$\begin{aligned}
\mathbf{M}'_1 &= \begin{pmatrix} 0.3916 & 0.2306 & 0.0460 & 0.0404 \\ 0.1408 & 0.6350 & 0.2139 & 0.2310 \\ 0.2375 & 0.0275 & 0.1667 & 0.2412 \\ 0.2301 & 0.1068 & 0.5734 & 0.4874 \end{pmatrix} \\
\mathbf{M}'_2 &= \begin{pmatrix} 0.3744 & 0.6892 & 0.0112 & 0.3200 \\ 0.3204 & 0.2320 & 0.4498 & 0.3530 \\ 0.0291 & 0.0688 & 0.3865 & 0.0653 \\ 0.2761 & 0.0100 & 0.1526 & 0.2618 \end{pmatrix} \\
\mathbf{M}'_3 &= \begin{pmatrix} 0.2885 & 0.0873 & 0.2319 & 0.1009 \\ 0.0653 & 0.2239 & 0.0575 & 0.2584 \\ 0.5934 & 0.3276 & 0.2283 & 0.3925 \\ 0.0529 & 0.3612 & 0.4823 & 0.2482 \end{pmatrix}
\end{aligned} \tag{45}$$

And 3 corresponding priors are sampled:

$$\begin{aligned}
\theta'_1 &= (0.2500, 0.2500, 0.2500, 0.2500)^\top \\
\theta'_2 &= (0.1789, 0.3664, 0.2915, 0.1632)^\top \\
\theta'_3 &= (0.4460, 0.4676, 0.0821, 0.0043)^\top
\end{aligned} \tag{46}$$

The value we use to test the effectiveness of perturbed SCBI is called the successful rate, which is $\mathbb{E}\left[\theta^L_\infty(h)\right] = \mathbb{E}_{\mu^L_\infty}\left[\theta(h)\right]$ where h is the true hypothesis that the teacher teaches (Definition 5.2). Successful rate is well defined, i.e. the limit exists, according to the convergences in probability (Theorem 3.5 and Proposition 5.1) with an ϵ discussion based on them. To find the successful rate, we use Monte-Carlo method on 10^4 teaching sequences, and use Proposition 5.1 to accelerate the simulation.

We can estimate an upper bound of the standard deviation (precision) of the empirical successful rate calculated based on Proposition 5.1. The successful rate of a single teaching sequence is between 0 and 1, thus with a standard deviation smaller than 1. So the standard deviation of the empirical successful rate is bounded by $(N)^{-1/2}$ where N is the number of sample sequences. Actually the precision is much smaller since the successful rate for a single sequence is much more stable.

Our first simulation is shown in Fig. 3, where we take θ^L_0 evenly on a series of concentric circles centered at θ^T_0. There are 14 such circles with radius 0.005 to 0.07. On the i-th layer (smallest circle is the first layer) we take $6i$ many points evenly separated, the upper right figure in Fig. 4 shows how the points are taken in detail.

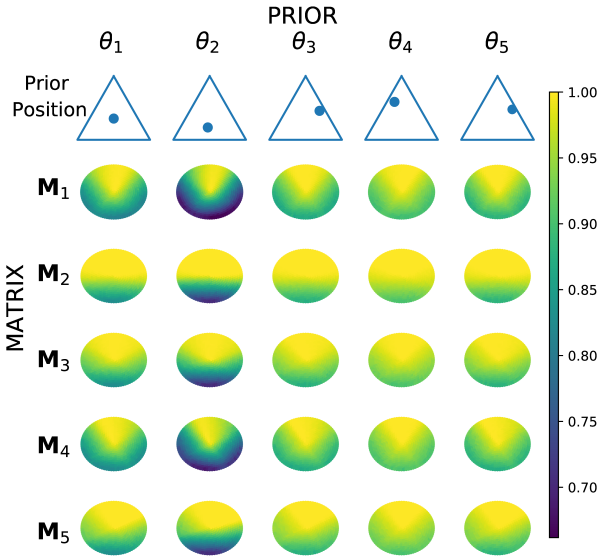

Figure 3. Successful rate of SCBI perturbed on prior. Each entry corresponds to a pair $\mathbf{M}$ and θ_0^T. The first row shows for each prior θ_0^T, the position it locates in Δ^2 and the range of θ_0^L in the simulation. The 5 rows below are the zoom-in version of the shaded area in each case, whose color at each point represents the successful rate when θ_0^L locates at that point.

Thus we have 6 groups of points each distributed along a ray. We plot the successful rate versus the distance from the center along each ray in Fig. 4 for all the 25 combinations of $\mathbf{M}$ and θ_0^T.

To have a similar directional data for matrices of size 4×4, we take a sample of 15 directions in $\mathbb{R}^3$ (showing in Fig. 5 in spherical coordinates centered at $(1/4, 1/4, 1/4, 1/4)^\top$, with $(1, 0, 0, 0)^\top$ as $\phi = 0$ axis and $(0, 1, 0, 0)^\top$ on the half-plane given by $\theta = 0$) and simulate the perturbations of θ_0^L in Δ^3 along the 15 directions. On each ray, we take 20 evenly placed θ_0^L with distance to the center θ_0^T from 0.005 to 0.100. Then we plot the successful rate versus the distance in Fig. 5 for all 9 cases as before.

Remark 2. This part provides evidences of linear influence of the perturbation distance on the successful rate along a fixed direction.

Next we explore the global behavior of perturbations on prior. Here we sample for each combination of $\mathbf{M}$ and θ_0^T a set of 300 points for θ_0^L evenly distributed in Δ^3.

In Fig. 6, we plot the successful rate versus the value of $\theta_0^L(h)$, for all 25 situations.

We plot in Fig. 7 the distance to center as x-coordinates, for 9 situations with matrices of size 4×4.

In this part, we observe that there is a lower bound of the successful rate which depends linearly on the distance to center, with slope bounded by $\frac{1}{\theta_0^T(h)}$ (Conjecture 5.3).

Empirical Data for Stability: Perturbation on Matrix

Fig. 8 shows the behavior of perturbations on all sampled 3×3 matrices in Section 5. Perturbations are taken only along the relevant column / irrelevant column, since a perturbation on the target column is equivalent of the combination of a perturbation on other two columns (they have the same set of Cross-ratios, which determines the SCBI behavior). The cycle path in each plot is the equi-normalized-KL path, with any point on the path having the same normalized-KL to the target column as that of the original matrix $\mathbf{T}$.

These graphs should not be confused with the ones occur in the prior perturbed part, as we are plotting each column of the matrix here (the simplex is actually $\mathcal{P}(\mathcal{D})$), while we were plotting the priors in previous discussion (the simplex is $\mathcal{P}(\mathcal{H})$).

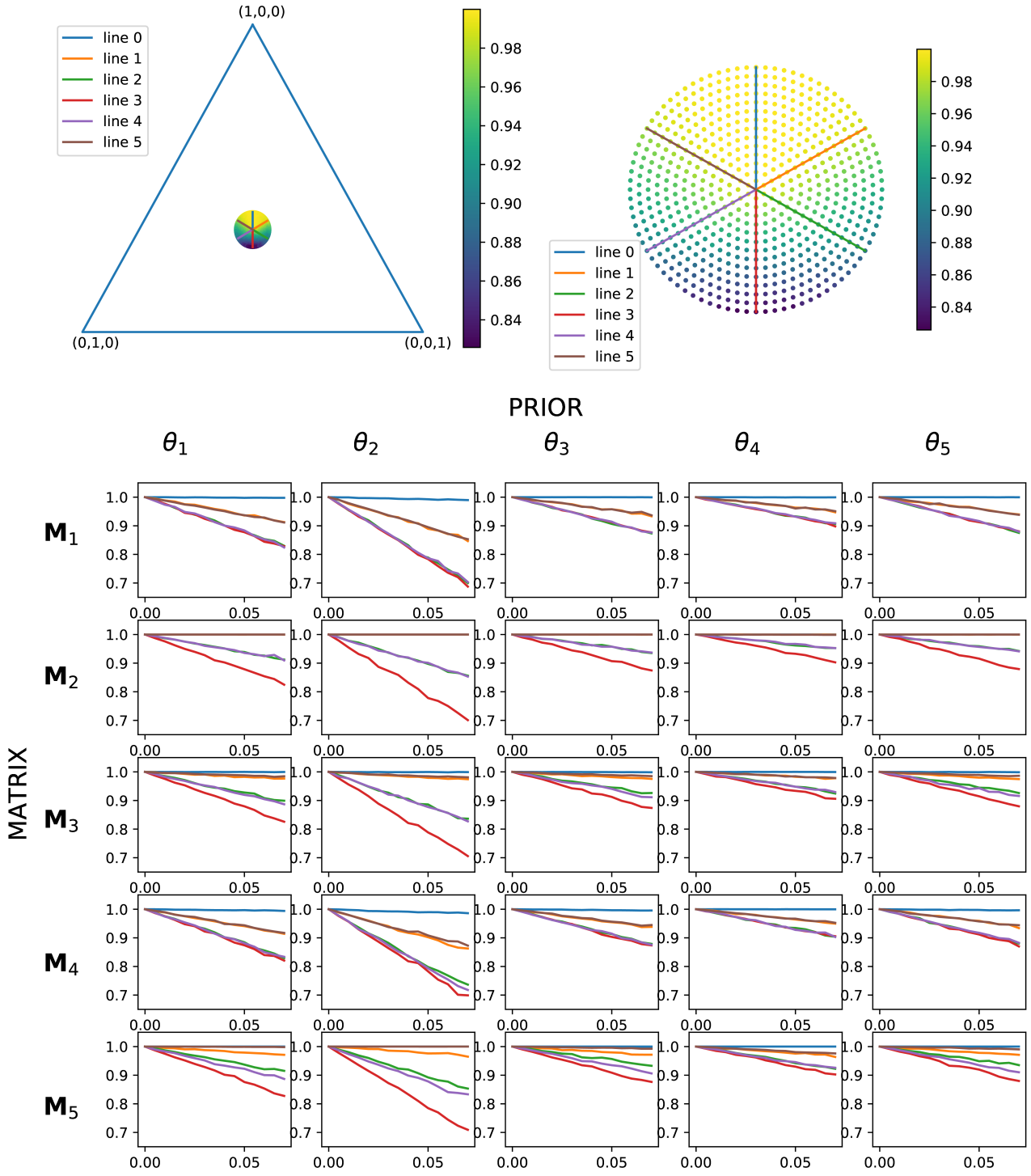

Figure 4. Upper-Left: the six rays at the center θ_1. Upper-Right: zoom-in figure of the six rays in general. Lower: Successful rate versus distance to center along 6 rays. Fig. 5 in the main paper contains the Row 3 Column 1 picture of the lower one (with a different y-scale).

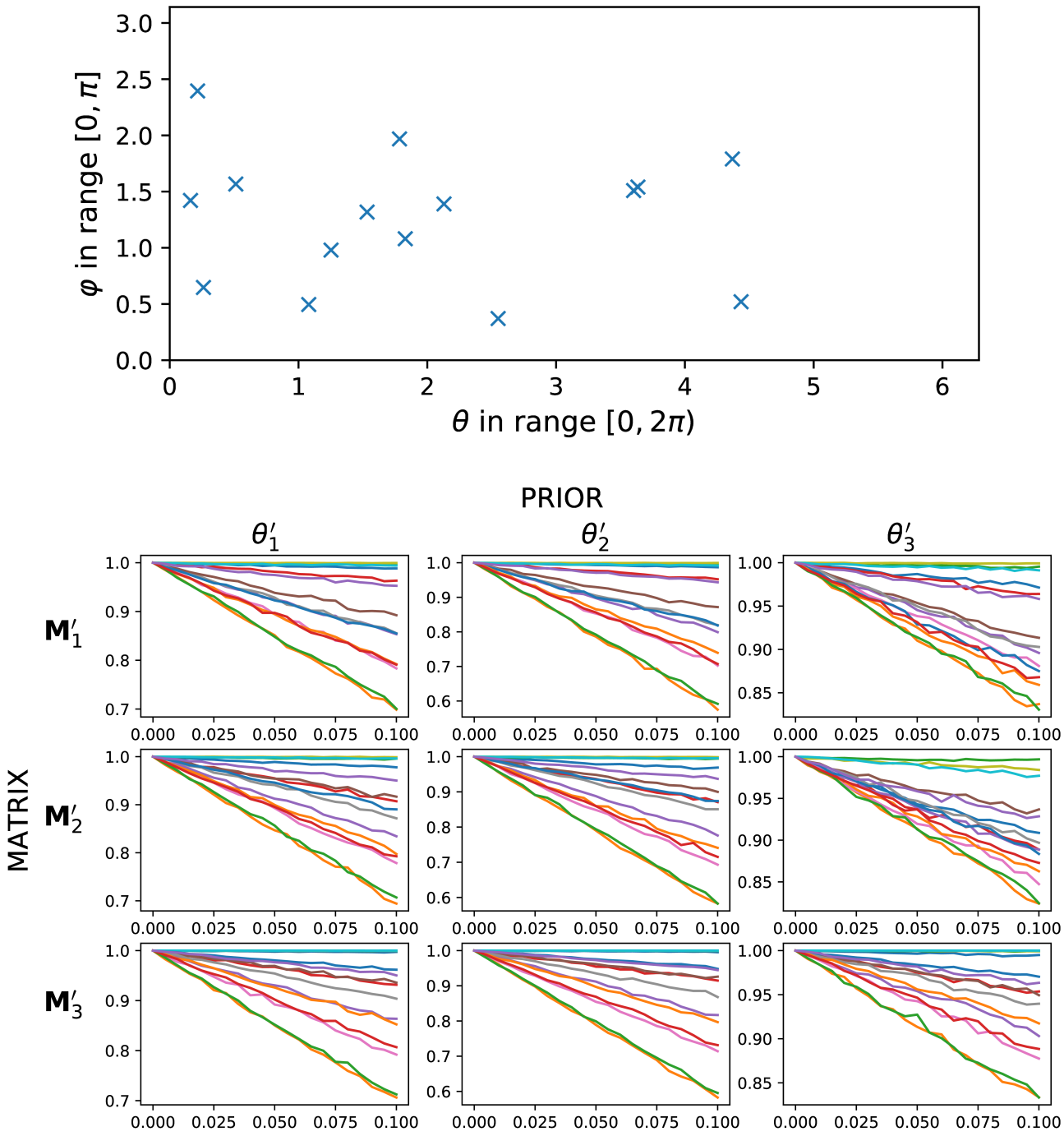

Figure 5. Upper: the sampled directions in spherical coordinates. Lower: Successful rate versus distance to center, along 15 rays, for all the 9 cases of matrices of size 4×4. The plot at Row 1 Column 1 appears in Fig. 5 of the main file.

References

Stephen E Fienberg et al. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

Klaus-J. Miescke and Friedrich Liese. *Statistical Decision*

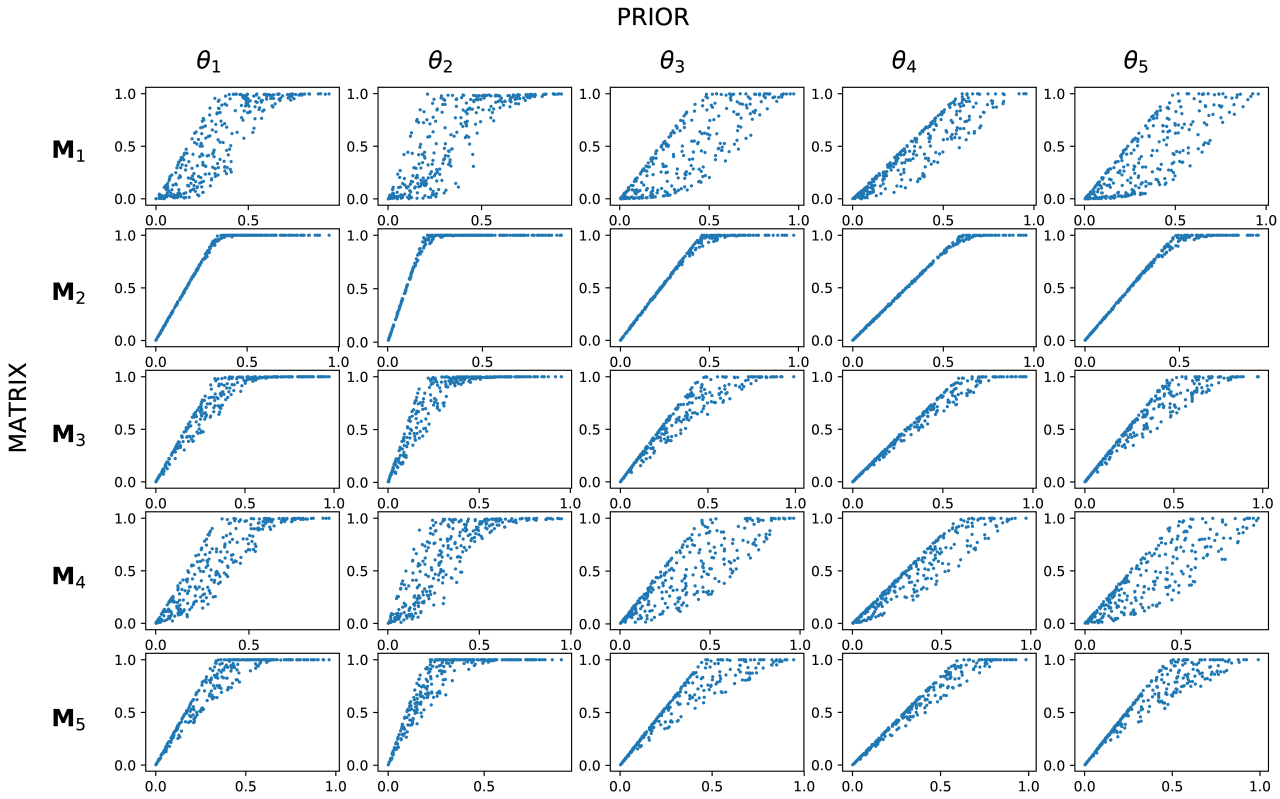

Figure 6. The 25 cases of 3×3 matrices, with successful rate versus $\theta_0^T(h)$ plotted. Plot at Row 3 Column 1 appears in Fig. 5 of the main file.

Theory: Estimation, Testing, and Selection. Springer, 2008. doi: https://doi.org/10.1007/978-0-387-73194-0.

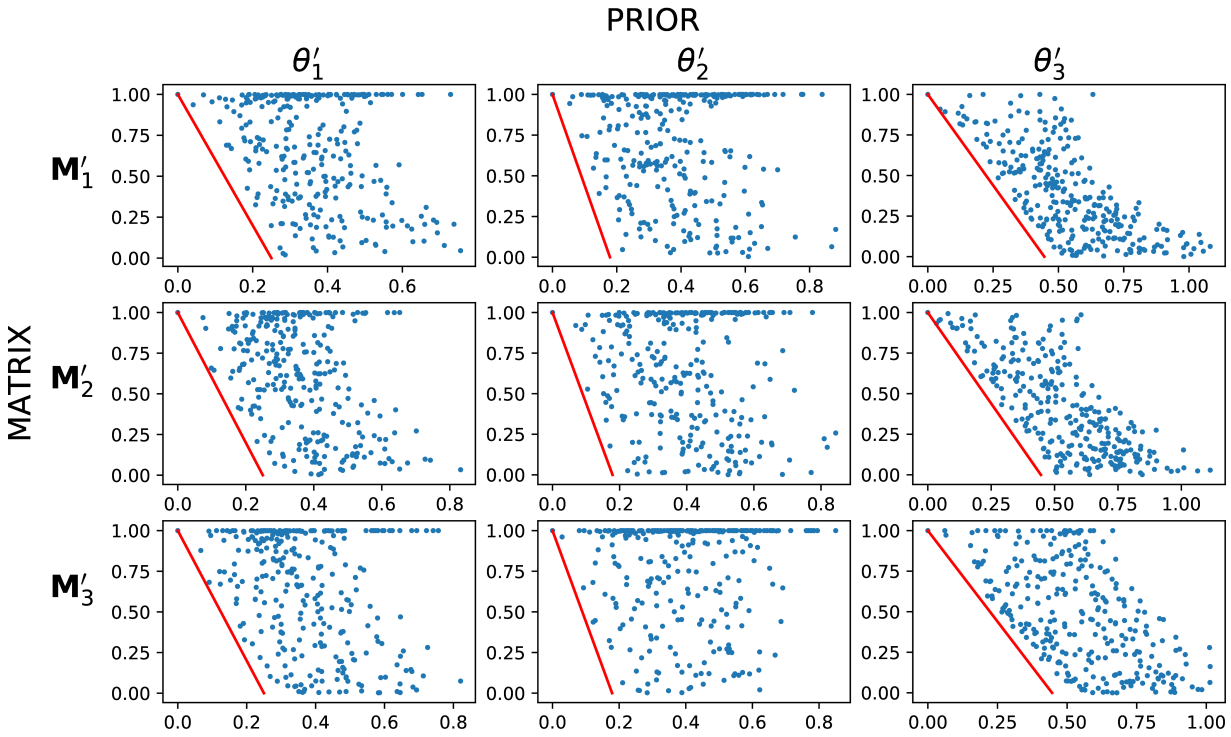

Figure 7. The 9 cases of 4×4 matrices, with successful rate versus the distance to the center plotted. Plot at Row 1 Column 1 appears in Fig. 5 of the main file. Red line is the lower bound given in Conjecture 5.3.

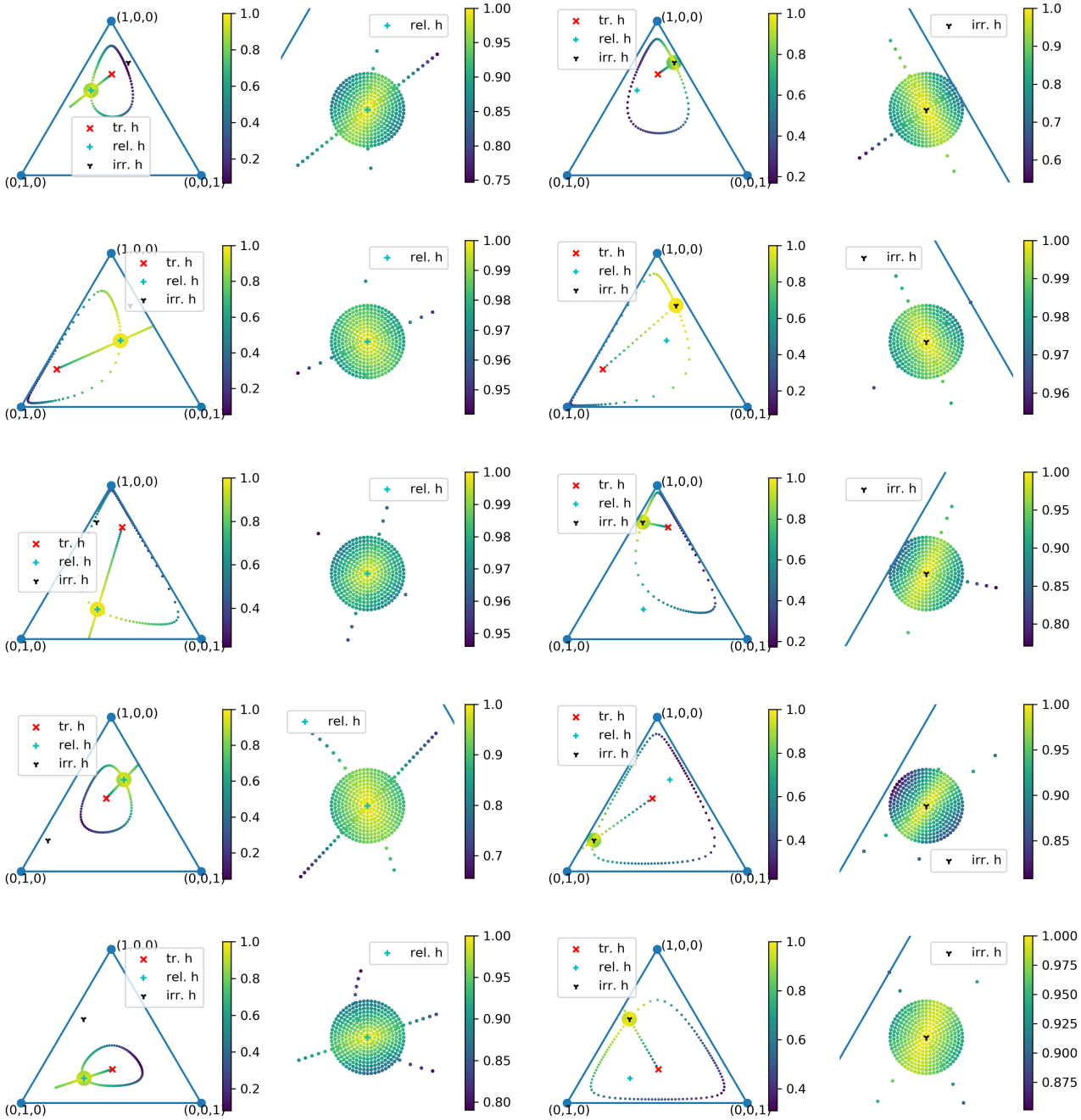

Figure 8. Perturbations on matrix $\mathbf{L}$. First column: Perturbations on the irrelevant column of $\mathbf{L}$. Second column: zoom-in of the first row. Third column: Perturbations on the relevant column of $\mathbf{L}$. Last column: zoom-in of the third column. The scales of color in the zoomed figures are different from that of the original ones. Fig. 6 in the main paper is the third row here.