Supplemental Material for the AISTATS 2014 paper "Decontamination of Mutually Contaminated Models"

Gilles Blanchard and Clayton Scott

A Proofs for Section 3

A.1 κ^* and $\widehat{\kappa}$ are well-defined

Lemma A.1. *The maximum operation in the definition of κ^* and $\widehat{\kappa}$ is well-defined, that is, the outside supremum is attained at at least one point.*

We prove the statement for

$$\kappa^* = \max_{\boldsymbol{\mu}} \inf_{C \in \mathcal{C}: F_{\boldsymbol{\mu}}(C) > 0} \frac{F_0(C)}{F_{\boldsymbol{\mu}}(C)}.$$

The argument for $\widehat{\kappa}$ is similar. Denote $G(\boldsymbol{\mu}) = \kappa^*(F_0 | F_{\boldsymbol{\mu}}) = \inf_{C \in \mathcal{C}: F_{\boldsymbol{\mu}}(C) > 0} \frac{F_0(C)}{F_{\boldsymbol{\mu}}(C)}$ the maximum proportion of the mixture $F_{\boldsymbol{\mu}}$ in the distribution F_0.

We argue that G is an upper semicontinuous function. To see this, define for each $C \in \mathcal{C}$ the function $g_C : S_M \to [0, \infty]$ as

$$g_C(\boldsymbol{\mu}) := \begin{cases} \frac{F_0(C)}{F_{\boldsymbol{\mu}}(C)} & \text{if } F_{\boldsymbol{\mu}}(C) > 0\,; \\ +\infty & \text{if } F_{\boldsymbol{\mu}}(C) = 0. \end{cases}$$

Then f_C is an upper semicontinuous function: if $\boldsymbol{\mu} \in S_M$ is such that $F_{\boldsymbol{\mu}}(C) > 0$, then f_C is continuous at point $\boldsymbol{\mu}$. Otherwise, $f_C(\boldsymbol{\mu}) = \infty$ and f_C is trivially upper semicontinuous at point $\boldsymbol{\mu}$. Clearly, one has $G(\boldsymbol{\mu}) = \inf_{C \in \mathcal{C}} f_C(\boldsymbol{\mu})$; as an infimum of upper semicontinuous functions, it is itself upper semicontinuous, and therefore attains its maximum on the compact set S_M.

A.2 Proof of Proposition 2

Point (a): We apply condition **P1** for all k, i with $\delta_{k,i} = c\delta_i / k^2$. By the union bound, with probability at least $1 - \sum_{i=0}^{M} \delta_i$, it holds simultaneously for all $k \geq 1$ and $i = 0, \ldots, M$ that

$$\forall k \geq 1\,, \qquad \forall i \in \{0, \ldots, M\} \ : \qquad \sup_{C \in \mathcal{C}_k} \left| F_i(C) - \widehat{F}_i(C) \right| \leq \epsilon_i^k (c\delta_i k^{-2}) \quad \text{(S.1)}$$

Recall the notation (from the proof of Lemma A.1) $G(\boldsymbol{\mu}) = \inf_{C\in\mathcal{C}:F_{\boldsymbol{\mu}}(C)>0} \frac{F_0(C)}{F_{\boldsymbol{\mu}}(C)}$ and introduce

$$\widehat{G}(\boldsymbol{\mu}) := \inf_k \inf_{C\in\mathcal{C}_k} \frac{\widehat{F}_0(C) + \epsilon_0^k(c\delta_0 k^{-2})}{\left(\widehat{F}_{\boldsymbol{\mu}}(C) - \sum_i \nu_i \epsilon_i^k(c\delta_i k^{-2})\right)_+}.$$

Observe that when (S.1) is satisfied, this implies that for all $\boldsymbol{\mu} \in S_M$, one has $G(\boldsymbol{\mu}) \leq \widehat{G}(\boldsymbol{\mu})$. Taking the maximum over $\boldsymbol{\mu}$ yields the first point.

Point (b): let $\epsilon > 0$ be an arbitrary positive constant. For any $\boldsymbol{\mu} \in S_M$, let $C_{\boldsymbol{\mu}} \in \mathcal{C}$ with $F_{\boldsymbol{\mu}}(C_{\boldsymbol{\mu}}) > 0$ be such that $\frac{F_0(C_{\boldsymbol{\mu}})}{F_{\boldsymbol{\mu}}(C_{\boldsymbol{\mu}})} \leq \kappa^* + \epsilon/4$.

By continuity of the function $\boldsymbol{\mu} \mapsto F_{\boldsymbol{\mu}}(C)$ for any fixed C, there exists for each $\boldsymbol{\mu} \in S_M$ an open neighborhood $N_{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ for which both of the following conditions are realized for all $\boldsymbol{\mu'} \in N_{\boldsymbol{\mu}}$:

$$\frac{F_0(C_{\boldsymbol{\mu}})}{F_{\boldsymbol{\mu'}}(C_{\boldsymbol{\mu}})} \leq \kappa^* + \frac{\epsilon}{2}, \tag{S.2}$$

$$\text{and } F_{\boldsymbol{\mu'}}(C_{\boldsymbol{\mu}}) \geq \frac{1}{2} F_{\boldsymbol{\mu}}(C_{\boldsymbol{\mu}}). \tag{S.3}$$

(For the second condition, we have used the fact that $F_{\boldsymbol{\mu}}(C_{\boldsymbol{\mu}}) > 0$). By compactness of S_M, there exists a finite subset S_M^ϵ of S_M such that $(N_{\boldsymbol{\mu}})_{\boldsymbol{\mu}\in S_M^\epsilon}$ covers S_M.

Denote $F_{\min}^\epsilon := \frac{1}{2}\min_{\boldsymbol{\mu}\in S_M^\epsilon} F_{\boldsymbol{\mu}}(C_{\boldsymbol{\mu}})$; it is a positive quantity since $F_{\boldsymbol{\mu}}(C_{\boldsymbol{\mu}}) > 0$ for any $\boldsymbol{\mu}$, and S_M^ϵ is finite. For each $\boldsymbol{\mu} \in S_M$, denote $\zeta(\boldsymbol{\mu})$ an arbitrary element of the finite net S_M^ϵ such that $\boldsymbol{\mu} \in N_{\zeta(\boldsymbol{\mu})}$. By property (S.2), we have

$$\sup_{\boldsymbol{\mu}\in S_M} \frac{F_0(C_{\zeta(\boldsymbol{\mu})})}{F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})})} \leq \max_{\boldsymbol{\mu}\in S_M^\epsilon} \sup_{\boldsymbol{\mu'}\in N_{\boldsymbol{\mu}}} \frac{F_0(C_{\boldsymbol{\mu}})}{F_{\boldsymbol{\mu'}}(C_{\boldsymbol{\mu}})} \leq \kappa^* + \frac{\epsilon}{2}, \tag{S.4}$$

and by property (S.3):

$$\inf_{\boldsymbol{\mu}\in S_M} F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})}) \geq \min_{\boldsymbol{\mu}\in S_M^\epsilon} \inf_{\boldsymbol{\mu'}\in N_{\boldsymbol{\mu}}} F_{\boldsymbol{\mu'}}(C_{\boldsymbol{\mu}}) \geq F_{\min}^\epsilon. \tag{S.5}$$

Denote $\mathcal{C}_\epsilon := \{C_{\boldsymbol{\mu}}\,,\ \boldsymbol{\mu} \in S_M^\epsilon\}$. Let $\eta \in (0, F_{\min}^\epsilon/2)$ be another arbitrary positive constant. Consider the distribution $Q = \frac{1}{M+1}\sum_{i=0}^M F_i$, to which we apply condition **P2**. This entails that for each individual $C \in \mathcal{C}$ there exists a k_C and $\widetilde{C} \in \mathcal{C}_{k_C}$ with

$$Q(C\Delta\widetilde{C}) \leq \frac{\eta}{M+1},$$

implying for all $i \in \{0, \ldots, M\}$:

$$\left|F_i(C) - F_i(\widetilde{C})\right| \leq F_i(C\Delta\widetilde{C}) \leq (M+1)Q(C\Delta\widetilde{C}) \leq \eta,$$

and then also for all $\boldsymbol{\mu} \in S_M$:

$$\left|F_{\boldsymbol{\mu}}(\widetilde{C}) - F_{\boldsymbol{\mu}}(C)\right| \leq \sum_{i=1}^M \mu_i \left|F_i(C) - F_i(\widehat{C})\right| \leq \eta.$$

In what follows we use the shortened notation $\varepsilon_i^k \equiv \epsilon_i^k(c\delta_i k^{-2})$, and further define $\underline{\varepsilon}(\epsilon,\eta) := \max_i \max_{C\in\mathcal{C}_\epsilon} \varepsilon_i^{k_C}$. For fixed (ϵ,η), the quantity $\underline{\varepsilon}(\epsilon,\eta)$ is defined as a maximum of a finite number of functions decreasing to 0 as $\boldsymbol{n}\to\infty$, and therefore $\underline{\varepsilon}$ also decreases to zero. Below, we assume that all components of $\boldsymbol{n}$ are chosen big enough so that $F_{\min}^{\epsilon} - \eta - 2\underline{\varepsilon}(\epsilon,\eta) > 0$. It holds with probability $1-\sum_{i=0}^{M}\delta_i$ that

$$\begin{aligned}
\widehat{\kappa} &\le \sup_{\boldsymbol{\mu}\in S_M} \inf_k \inf_{C\in\mathcal{C}_k} \frac{F_0(C)+2\varepsilon_0^k}{\left(F_{\boldsymbol{\mu}}(C) - 2\sum_i \mu_i \varepsilon_i^k\right)_+} \\
&\le \sup_{\boldsymbol{\mu}\in S_M} \inf_{C\in\mathcal{C}} \frac{F_0(\widetilde{C})+2\varepsilon_0^{k_C}}{\left(F_{\boldsymbol{\mu}}(\widetilde{C}) - 2\sum_i \mu_i \varepsilon_i^{k_C}\right)_+} \\
&\le \sup_{\boldsymbol{\mu}\in S_M} \inf_{C\in\mathcal{C}} \frac{F_0(C)+\eta+2\varepsilon_0^{k_C}}{\left(F_{\boldsymbol{\mu}}(C) - \eta - 2\sum_i \mu_i \varepsilon_i^{k_C}\right)_+} \\
&\le \sup_{\boldsymbol{\mu}\in S_M} \frac{F_0(C_{\zeta(\boldsymbol{\mu})})+\eta+2\varepsilon_0^{k_{C_{\zeta(\boldsymbol{\mu})}}}}{\left(F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})}) - \eta - 2\sum_i \mu_i \varepsilon_i^{k_{C_{\zeta(\boldsymbol{\mu})}}}\right)_+} \\
&\le \sup_{\boldsymbol{\mu}\in S_M} \frac{F_0(C_{\zeta(\boldsymbol{\mu})})+\eta+2\underline{\varepsilon}(\epsilon,\eta)}{\left(F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})}) - \eta - 2\underline{\varepsilon}(\epsilon,\eta)\right)_+} \\
&\le \left(\sup_{\boldsymbol{\mu}\in S_M} \frac{F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})})}{\left(F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})}) - \eta - 2\underline{\varepsilon}(\epsilon,\eta)\right)_+}\right) \sup_{\boldsymbol{\mu}\in S_M} \frac{F_0(C_{\zeta(\boldsymbol{\mu})})+\eta+2\underline{\varepsilon}(\epsilon,\eta)}{F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})})} \\
&\le \left(\frac{F_{\min}^{\epsilon}}{\left(F_{\min}^{\epsilon} - \eta - 2\underline{\varepsilon}(\epsilon,\eta)\right)_+}\right)\left(\sup_{\boldsymbol{\mu}\in S_M}\frac{F_0(C_{\zeta(\boldsymbol{\mu})})}{F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})})} + \sup_{\boldsymbol{\mu}\in S_M}\frac{\eta+2\underline{\varepsilon}(\epsilon,\eta)}{F_{\boldsymbol{\mu}}(C_{\zeta(\boldsymbol{\mu})})}\right) \\
&\le \left(\frac{F_{\min}^{\epsilon}}{\left(F_{\min}^{\epsilon} - \eta - 2\underline{\varepsilon}(\epsilon,\eta)\right)_+}\right)\left(\kappa^* + \frac{\epsilon}{2}\right) + \frac{\eta+2\underline{\varepsilon}(\epsilon,\eta)}{\left(F_{\min}^{\epsilon} - \eta - 2\underline{\varepsilon}(\epsilon,\eta)\right)_+},
\end{aligned}$$

where we have used (S.4) and (S.5) for the last inequality. By choosing first η small enough, then all components of $\boldsymbol{n}_0$ big enough, the r.h.s. of the above inequality can be made smaller than $\kappa^*+\epsilon$, for all $\boldsymbol{n}\succ\boldsymbol{n}_0$ ($\succ$ indicates the inequality holds for all components). Since $\sum_{i=0}^{M}\delta_i \to 0$ as $\boldsymbol{\mu}\to 0$, this implies the second part of the proposition.

For the last point of the proposition, consider an arbitrary open set Ω containing the set $\mathcal{B}^*$. Then $\Omega^c := S_M \setminus \Omega$ is a compact set; therefore, the function $G(\boldsymbol{\mu}) := \inf_{C\in\mathcal{C},F_{\boldsymbol{\mu}}(C)>0} \frac{F_0(C)}{F_{\boldsymbol{\mu}}(C)}$, being upper semicontinuous (see proof of Lemma A.1), attains its supremum $\widetilde{\kappa}$ on Ω^c. Observe that $\widetilde{\kappa} > \kappa^*$ must hold, otherwise we would have a contradiction with the definition of $\mathcal{B}^*$. Finally, we have:

$$\begin{aligned}
\mathbb{P}\left[\widehat{\boldsymbol{\mu}} \notin \Omega\right] &\le \mathbb{P}\left[\widehat{\boldsymbol{\mu}} \notin \Omega; G(\widehat{\boldsymbol{\mu}}) \le \widehat{G}(\widehat{\boldsymbol{\mu}})\right] + \mathbb{P}\left[G(\widehat{\boldsymbol{\mu}}) > \widehat{G}(\widehat{\boldsymbol{\mu}})\right] \\
&\le \mathbb{P}\left[\widehat{\kappa} \ge \widetilde{\kappa}\right] + \sum_{i=1}^{M}\delta_i,
\end{aligned}$$

where we have used that $\widehat{\kappa} = \widehat{G}(\widehat{\boldsymbol{\mu}})$ by definition, and the argument used in the proof of point (a). By point (b), the first probability converges to 0 as $\boldsymbol{\mu} \to \infty$. Thus, the probability that $\widehat{\boldsymbol{\mu}} \in \Omega$ must converge to 1 as $\boldsymbol{n} \to \infty$. This applies in particular to any open set of the form $\Omega_\epsilon := \{\boldsymbol{\mu} : d(\boldsymbol{\mu}, \mathcal{B}^*) < \epsilon\}$, hence the conclusion.

B Proofs for Section 4

B.1 Proof of Lemma 1

Suppose the first condition does not hold, so that

$$\sum_{i \in I} \epsilon_i P_i = \alpha \left(\sum_{i \notin I} \epsilon_i P_i \right) + (1-\alpha) H.$$

Then $\sum_i \gamma_i P_i = H$, where $\gamma_i = \frac{\epsilon_i}{1-\alpha}$ for $i \in I$, and $\gamma_i = -\frac{\alpha \epsilon_i}{1-\alpha}$ for $i \notin I$. Since $\sum_{i \notin I} \epsilon_i = 1$, at least one $\gamma_i < 0$, so the second condition is violated.

Now suppose the second condition is violated, say $\sum_i \gamma_i P_i = H$. Let $I = \{i \,|\, \gamma_i \geq 0\}$, which has fewer than K elements by assumption. Since $\sum_i \gamma_i = 1$, we also know $1 \leq |I|$ and further that $\Gamma := \sum_{i \in I} \gamma_i > 1$. A violation of the first condition is obtained by $\epsilon_i = \gamma_i / \Gamma$ for $i \in I$, $\epsilon_i = -\gamma_i/(\Gamma - 1)$ for $i \notin I$ (noting that $\sum_{i \notin I} (-\gamma_i) = \Gamma - 1$), and $\alpha = (\Gamma - 1)/\Gamma$.

B.2 Proof of Lemma 2

(a) $\Rightarrow$ (b): Follows immediately from the definition of the residue.

(b) $\Rightarrow$ (c): By assumption, there exists $\kappa > 0$ such that $\boldsymbol{\pi}_1 = \kappa \boldsymbol{e}_1 + (1-\kappa)\boldsymbol{\eta}_1$, where $\boldsymbol{\eta}_1 = \sum_{i=2}^{L} \mu_i \boldsymbol{\pi}_i$, with $\mu_i \geq 0$, for all $2 \leq i \leq L$. Thus,

$$\boldsymbol{e}_1 = \kappa^{-1} \boldsymbol{\pi}_1 - \sum_{i=2}^{L} \frac{(1-\kappa)}{\kappa} \mu_i \boldsymbol{\pi}_i \,;$$

a similar relation holds for all rows. This implies that Π is invertible and allows to identify (for instance) the first row of Π^{-1} as $(\kappa^{-1}, -\frac{(1-\kappa)}{\kappa}\mu_2, \ldots, -\frac{(1-\kappa)}{\kappa}\mu_L)$. This implies (c).

(c) $\Rightarrow$ (a): Without loss of generality, consider $\ell = 1$ and the problem of identifying $\kappa^*(\pi_1 | (\pi_i)_{2 \leq i \leq L})$, and the associated residue (if it exists). According to characterization (9), this corresponds to the optimization problem

$$\max_{\boldsymbol{\nu}, \boldsymbol{\gamma}} \sum_{i=2}^{L} \nu_i \;\; s.t. \;\; \boldsymbol{\pi}_1 = (1 - \sum_{i \geq 2} \nu_i) \boldsymbol{\gamma} + \sum_{i \geq 2} \nu_i \boldsymbol{\pi}_i,$$

over $\boldsymbol{\gamma} \in S_L$ and $\boldsymbol{\nu} = (\nu_2, \ldots, \nu_L) \in \Delta_{L-1} = \Big\{ (\nu_2, \ldots, \nu_L) | \nu_i \geq 0; \sum_{i=2}^{L} \nu_i \leq 1 \Big\}$.

We now reformulate this problem. First, note that the constraint implies that admissible $\boldsymbol{\nu}$ are such that $\sum_{i\geq 2}\nu_i < 1$, otherwise we would have a linear relation between the $\boldsymbol{\pi}_i$, contradicting invertibility of Π.

Then for an admissible $\boldsymbol{\nu}$, denote $\boldsymbol{\eta}(\boldsymbol{\nu}) := (1-\sum_{i\geq 2}\nu_i)^{-1}(1,-\nu_2,\ldots,-\nu_L)$. Observe that the constraint of the optimization problem is equivalent to $\Pi^T\boldsymbol{\eta}=\boldsymbol{\gamma}$, or $\boldsymbol{\eta}=(\Pi^T)^{-1}\boldsymbol{\gamma}$. The inverse mapping of $\boldsymbol{\eta}$ to $\boldsymbol{\nu}$ is $\boldsymbol{\nu}(\boldsymbol{\eta})=\eta_1^{-1}(-\eta_2,\ldots,-\eta_L)$, so that the objective of the optimization can be rewritten as

$$-\frac{\sum_{i=2}^L \eta_i}{\boldsymbol{e}_1^T\boldsymbol{\eta}} = -\frac{\mathbf{1}^T\boldsymbol{\eta}}{\boldsymbol{e}_1^T\boldsymbol{\eta}}+1 = 1-\frac{1}{\boldsymbol{e}_1^T\boldsymbol{\eta}} = 1-\frac{1}{\boldsymbol{e}_1^T(\Pi^T)^{-1}\boldsymbol{\gamma}},$$

where $\mathbf{1}$ denotes a L-dimensional vector with all coordinates equal to 1. So finding the point of maximum of the above problem is equivalent to the program

$$\max_{\boldsymbol{\gamma}\in S_L} \boldsymbol{e}_1^T(\Pi^T)^{-1}\boldsymbol{\gamma} \;\; s.t. \;\; \boldsymbol{\nu}((\Pi^T)^{-1}\boldsymbol{\gamma})\in\Delta_{L-1}$$

The above objective function has the form $\boldsymbol{a}^T\boldsymbol{\gamma}$, where $\boldsymbol{a}$ is the first column of Π^{-1} which, by assumption, has its first coordinate positive and the others nonpositive. Therefore, the unconstrainted maximum over $\boldsymbol{\gamma}\in S_M$ is attained uniquely for $\boldsymbol{\gamma}=\boldsymbol{e}_1$. We now check that this value also satisfies the required constraint. Observe that $(\Pi^T)^{-1}\boldsymbol{e}_1$ is the (transpose of) the first row of Π^{-1}, denote this vector as $\boldsymbol{b}=(b_1,\ldots,b_L)$. We want to ensure that $\boldsymbol{\nu}(\boldsymbol{b})=b_1^{-1}(-b_2,\ldots,-b_L)\in\Delta_{L-1}$. By assumption, $\boldsymbol{b}$ has its first coordinate positive and the others nonpositive, ensuring all components of $\boldsymbol{\nu}(\boldsymbol{b})$ are nonnegative. Furthermore, the sum of the components of $\boldsymbol{\nu}(\boldsymbol{b})$ is

$$\sum_{i=2}^L -\frac{b_i}{b_1} = 1-\frac{\sum_{i=1}^L b_i}{b_1} = 1-\frac{1}{b_1}\leq 1;$$

the last equality is because the rows of Π^{-1} sum to 1 (since Π is a stochastic matrix, so is its inverse). It follows that $\boldsymbol{\nu}((\Pi^T)^{-1}\boldsymbol{e}_1)\in\Delta_{L-1}$. Thus, the unique maximum of the optimization problem is attained for $\boldsymbol{\gamma}=\boldsymbol{e}_1$, establishing (a).

B.3 Proof of Proposition 3

We start with the following Lemma:

Lemma B.1. *If Π is recoverable, then $\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_L$ are linearly independent. If $P_1,\ldots,P_L$ are jointly irreducible, then they are linearly independent. If $\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_L$ are linearly independent and $P_1,\ldots,P_L$ are linearly independent, then $\tilde{P}_1,\ldots,\tilde{P}_L$ are linearly independent.*

Proof of the lemma: The first statement follows from characterization (c) of Lemma 2: if Π is recoverable, it is invertible and thus has full rank.

For the second statement, suppose $\sum_i \beta_i P_i = 0$ is a nontrivial linear relation. Let j be any index such that $\beta_j\geq 0$. Then $\sum_i \gamma_i P_i = P_j$, where $\gamma_i=\beta_i$ if $i\neq j$, and $\gamma_j=\beta_j+1$. Since at least one $\beta_i<0$, $i\neq j$, joint irreducibility is violated.

For the third part, suppose $\sum_i \alpha_i \tilde{P}_i = 0$. Since $\tilde{P}_i = \boldsymbol{\pi}_i^T \boldsymbol{P}$, this implies $\sum_i \alpha_i \boldsymbol{\pi}_i^T \boldsymbol{P} = 0$, which implies $\sum_i \alpha_i \boldsymbol{\pi}_i = \mathbf{0}$, which implies $\alpha_i = 0$.

Proof of Proposition 3: Consider $\ell = 1$, the other cases being similar. Suppose G is such that

$$\tilde{P}_1 = (1 - \sum_{j \geq 2} \nu_j) G + \sum_{j \geq 2} \nu_j \tilde{P}_j. \tag{S.6}$$

Note that $\tilde{P}_1, \ldots, \tilde{P}_L$ are linearly independent by Lemma B.1. This implies $\sum_{j \geq 2} \nu_j < 1$, because otherwise $\tilde{P}_1 = \sum_{j \geq 2} \nu_j \tilde{P}_j$.

Therefore, any G satisfying (S.6) has the form $\sum_{i=1}^{L} \gamma_i P_i$. The weights γ_i clearly sum to one, and by joint irreducibility, they are nonnegative. That is, $\boldsymbol{\gamma} := [\gamma_1, \ldots, \gamma_L]^T$ is a discrete distribution. Thus, Eqn. (S.6) is equivalent to

$$\boldsymbol{\pi}_1^T \boldsymbol{P} = (1 - \sum_{j \geq 2} \nu_j) \boldsymbol{\gamma}^T \boldsymbol{P} + \sum_{j \geq 2} \nu_j \boldsymbol{\pi}_j^T \boldsymbol{P}.$$

By linear independence of $P_1, \ldots, P_L$ (see Lemma B.1) and taking the transpose, this gives

$$\boldsymbol{\pi}_1 = (1 - \sum_{j \geq 2} \nu_j) \boldsymbol{\gamma} + \sum_{j \geq 2} \nu_j \boldsymbol{\pi}_j.$$

Therefore $\kappa^*(\tilde{P}_1 | \{\tilde{P}_j, j \neq 1\}) = \kappa^*(\boldsymbol{\pi}_1 | \{\boldsymbol{\pi}_j, j \neq 1\}) < 1$, and there is a one-to-one correspondence between feasible G in the definition of $\kappa^*(\tilde{P}_1 | \{\tilde{P}_j, j \neq 1\})$ and feasible $\boldsymbol{\gamma}$ in the definition of $\kappa^*(\boldsymbol{\pi}_1 | \{\boldsymbol{\pi}_j, j \neq 1\})$. Since Π is recoverable, the residue of $\boldsymbol{\pi}_1$ w.r.t. $\{\boldsymbol{\pi}_j, j \neq 1\}$ is $\boldsymbol{\gamma} = \boldsymbol{e}_1$, and so the residue of $\tilde{P}_1$ w.r.t. $\{\tilde{P}_j, j \neq 1\}$ is $G = \boldsymbol{e}_1^T \boldsymbol{P} = P_1$.

To see uniqueness of the maximizing ν_j, suppose

$$\tilde{P}_1 = (1 - \kappa^*) G + \sum_{j \geq 2} \nu_j \tilde{P}_j = (1 - \kappa^*) G + \sum_{j \geq 2} \nu'_j \tilde{P}_j.$$

Lemma B.1 implies $\nu_j = \nu'_j$.

B.4 Proof of Proposition 4

For brevity we at times omit the dependence of the errors and their estimates on f. For any f,

$$\begin{aligned}
|R_\ell(f) - \widehat{R}_\ell(f)| &= \left| \frac{\tilde{R}_{\ell\ell} - \sum_{j \neq \ell} \nu_{\ell j} \tilde{R}_{j\ell}}{1 - \kappa_\ell} - \frac{\widehat{\tilde{R}}_{\ell\ell} - \sum_{j \neq \ell} \widehat{\nu}_{\ell j} \widehat{\tilde{R}}_{j\ell}}{1 - \widehat{\kappa}_\ell} \right| \\
&\leq \left| \frac{\tilde{R}_{\ell\ell} - \sum_{j \neq \ell} \nu_{\ell j} \tilde{R}_{j\ell}}{1 - \kappa_\ell} - \frac{\widehat{\tilde{R}}_{\ell\ell} - \sum_{j \neq \ell} \widehat{\nu}_{\ell j} \widehat{\tilde{R}}_{j\ell}}{1 - \kappa_\ell} \right| \\
&\quad + \left| \frac{\widehat{\tilde{R}}_{\ell\ell} - \sum_{j \neq \ell} \widehat{\nu}_{\ell j} \widehat{\tilde{R}}_{j\ell}}{1 - \kappa_\ell} - \frac{\widehat{\tilde{R}}_{\ell\ell} - \sum_{j \neq \ell} \widehat{\nu}_{\ell j} \widehat{\tilde{R}}_{j\ell}}{1 - \widehat{\kappa}_\ell} \right|
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{|\tilde{R}_{\ell\ell} - \widehat{\tilde{R}}_{\ell\ell}| + \sum_{j\neq\ell} |\nu_{\ell j}\tilde{R}_{j\ell} - \widehat{\nu}_{\ell j}\widehat{\tilde{R}}_{j\ell}|}{1-\kappa_\ell} + \left|\frac{1}{1-\kappa_\ell} - \frac{1}{1-\widehat{\kappa}_\ell}\right| \\
&= \frac{|\tilde{R}_{\ell\ell} - \widehat{\tilde{R}}_{\ell\ell}| + \sum_{j\neq\ell} \left(|\nu_{\ell j}\tilde{R}_{j\ell} - \widehat{\nu}_{\ell j}\tilde{R}_{j\ell} + \widehat{\nu}_{\ell j}\tilde{R}_{j\ell} - \widehat{\nu}_{\ell j}\widehat{\tilde{R}}_{j\ell}|\right)}{1-\kappa_\ell} \\
&\qquad + \left|\frac{1}{1-\kappa_\ell} - \frac{1}{1-\widehat{\kappa}_\ell}\right| \\
&\le \frac{|\tilde{R}_{\ell\ell} - \widehat{\tilde{R}}_{\ell\ell}| + \sum_{j\neq\ell} \left(|\nu_{\ell j} - \widehat{\nu}_{\ell j}| + |\tilde{R}_{j\ell} - \widehat{\tilde{R}}_{j\ell}|\right)}{1-\kappa_\ell} + \left|\frac{1}{1-\kappa_\ell} - \frac{1}{1-\widehat{\kappa}_\ell}\right|.
\end{aligned}
$$

The VC inequality [1] implies that for any $\epsilon > 0$, $\sup_{f\in\mathcal{F}_{k(\boldsymbol{n})}} |R_{i\ell}(f) - \widehat{R}_{i\ell}(f)| \le \epsilon$ with probability tending to 1, since (12) holds, and by our convention for multiclass VC dimension. Noting that $\kappa_\ell < 1$ by Proposition 3, the other terms tend to zero in probability by consistency of $\widehat{\kappa}_\ell$ and the $\widehat{\nu}_{\ell j}$. This completes the proof.

B.5 Proof of Theorem 1

Consider the decomposition into estimation and approximation errors,

$$
R(\widehat{f}) - R^* = R(\widehat{f}) - \inf_{f\in\mathcal{F}_{k(\boldsymbol{n})}} R(f) + \inf_{f\in\mathcal{F}_{k(\boldsymbol{n})}} R(f) - R^*.
$$

The approximation error converges to zero by **P3** and since $k(\boldsymbol{n}) \to \infty$. To analyze the estimation error, let $\epsilon > 0$. For each positive integer k, let $f_k^* \in \mathcal{F}_k$ such that $R(f_k^*) \le \inf_{f\in\mathcal{F}_k} R(f) + \frac{\epsilon}{4}$. Then

$$
\begin{aligned}
R(\widehat{f}) - \inf_{f\in\mathcal{F}_{k(\boldsymbol{n})}} R(f) &= R(\widehat{f}_{k(\boldsymbol{n})}) - \inf_{f\in\mathcal{F}_{k(\boldsymbol{n})}} R(f) \\
&\le R(\widehat{f}_{k(\boldsymbol{n})}) - R(f^*_{k(\boldsymbol{n})}) + \frac{\epsilon}{4} \\
&\le \widehat{R}(\widehat{f}_{k(\boldsymbol{n})}) - \widehat{R}(f^*_{k(\boldsymbol{n})}) + \frac{\epsilon}{2} \\
&\qquad \text{(with probability tending to 0, by Proposition 4)} \\
&\le \tau_{k(\boldsymbol{n})} + \frac{\epsilon}{2} \\
&\le \epsilon
\end{aligned}
$$

where the last step holds for $\boldsymbol{n}$ sufficiently large. The result now follows.

References

[1] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.