Supplementary Material for "Necessary and Sufficient Conditions for High-Dimensional Salient Subset Recovery"

Vincent Y. F. Tan, Matthew Johnson and Alan S. Willsky

April 29, 2010

The proofs of the paper "Necessary and Sufficient Conditions for High-Dimensional Salient Subset Recovery" presented in Austin, TX for ISIT 2010 are provided. Note that all equation numbers refer to the corresponding equation in the main paper (e.g., (1) refers to equation (1) in the main paper). The same holds for theorems, propositions and lemmas.

Contents

1 Proof of Proposition 1

Proof We will prove that (S3) $\Leftrightarrow$ (S1) $\Leftrightarrow$ (S2). Assuming (S1) holds, $D(P^{(d)} \,||\, Q^{(d)}) = D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d})$ implies that the conditional KL-divergence is identically zero, i.e,

$$D(P^{(d)}_{S^c_d|S_d} \,||\, Q^{(d)}_{S^c_d|S_d}) = 0. \tag{A-1}$$

Expanding the above expression yields the following:

$$\sum_{\mathbf{x}_{S_d}} P^{(d)}(\mathbf{x}_{S_d}) \sum_{\mathbf{x}_{S^c_d}} P^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d}) \log \frac{P^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d})}{Q^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d})} = 0. \tag{A-2}$$

From the positivity of the distributions and non-negativity of the KL-divergence, we have that

$$D(P^{(d)}_{S^c_d|S_d}(\cdot|\mathbf{x}_{S_d}) \,||\, Q^{(d)}_{S^c_d|S_d}(\cdot|\mathbf{x}_{S_d})) := \sum_{\mathbf{x}_{S^c_d}} P^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d}) \log \frac{P^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d})}{Q^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d})} = 0, \quad \forall\, \mathbf{x}_{S_d} \in \mathcal{X}^k. \tag{A-3}$$

We conclude that

$$P^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d}) = Q^{(d)}_{S^c_d|S_d}(\mathbf{x}_{S^c_d}|\mathbf{x}_{S_d}), \quad \forall\, \mathbf{x}_{S_d} \in \mathcal{X}^k,\, \mathbf{x}_{S^c_d} \in \mathcal{X}^{d-k}, \tag{A-4}$$

which implies that the conditional distributions are identical. This proves (S3). The reverse implication is obvious.

Assume that S_d is KL-divergence salient (S1). Then from the above, we have (6). The Chernoff information is then given by

$$D^*(P^{(d)}, Q^{(d)}) = -\min_{t\in[0,1]} \log \left(\sum_{\mathbf{z}} (P^{(d)}(\mathbf{z}))^t (Q^{(d)}(\mathbf{z}))^{1-t} \right), \tag{A-5}$$

$$= -\min_{t\in[0,1]} \log \left(\sum_{\mathbf{z}} (P^{(d)}_{S_d}(\mathbf{z}_{S_d}) W_{S^c_d|S_d}(\mathbf{z}_{S^c_d}|\mathbf{z}_{S_d}))^t (Q^{(d)}_{S_d}(\mathbf{z}_{S_d}) W_{S^c_d|S_d}(\mathbf{z}_{S^c_d}|\mathbf{z}_{S_d}))^{1-t} \right), \tag{A-6}$$

$$= -\min_{t\in[0,1]} \log \left(\sum_{\mathbf{z}} (P^{(d)}_{S_d}(\mathbf{z}_{S_d}))^t (Q^{(d)}_{S_d}(\mathbf{z}_{S_d}))^{1-t} W_{S^c_d|S_d}(\mathbf{z}_{S^c_d}|\mathbf{z}_{S_d}) \right), \tag{A-7}$$

$$= -\min_{t\in[0,1]} \log \left(\sum_{\mathbf{z}_{S_d}} (P^{(d)}_{S_d}(\mathbf{z}_{S_d}))^t (Q^{(d)}_{S_d}(\mathbf{z}_{S_d}))^{1-t} \sum_{\mathbf{z}_{S^c_d}} W_{S^c_d|S_d}(\mathbf{z}_{S^c_d}|\mathbf{z}_{S_d}) \right), \tag{A-8}$$

$$= -\min_{t\in[0,1]} \log \left(\sum_{\mathbf{z}_{S_d}} (P^{(d)}_{S_d}(\mathbf{z}_{S_d}))^t (Q^{(d)}_{S_d}(\mathbf{z}_{S_d}))^{1-t} \right) = D^*(P^{(d)}_{S_d}, Q^{(d)}_{S_d}), \tag{A-9}$$

which proves that S_d is Chernoff information salient (S2). Now for the reverse implication, we claim the following lemma:

Lemma A-1 (Monotonicity of Chernoff information) *For every set $A \subset V_d$, the Chernoff information satisfies*

$$D^*(P^{(d)}, Q^{(d)}) \geq D^*(P^{(d)}_A, Q^{(d)}_A), \tag{A-10}$$

with equality if and only if (6) *holds, i.e., the conditionals $P^{(d)}_{A^c|A}$ and $Q^{(d)}_{A^c|A}$ are identical.*

Assuming Lemma A-1 and assuming that S_d is Chernoff information-salient, we have that $P^{(d)}$ and $Q^{(d)}$ satisfy (S3). Since (S3) $\Leftrightarrow$ (S2), this completes the proof of Lemma 1. It remains to prove Lemma A-1. $\square$

Proof of Lemma A-1
We drop the superscript (d) for notational simplicity. Then we have the following chain

$$D^*(P,Q) = -\min_{t\in[0,1]} \log\left(\sum_{\mathbf{z}} P(\mathbf{z})^t Q(\mathbf{z})^{1-t}\right), \tag{A-11}$$

$$= -\min_{t\in[0,1]} \log\left(\sum_{\mathbf{z}} P_A(\mathbf{z}_A)^t Q_A(\mathbf{z}_A)^{1-t} P_{A^c|A}(\mathbf{z}_{A^c}|\mathbf{z}_A)^t Q_{A^c|A}(\mathbf{z}_{A^c}|\mathbf{z}_A)^{1-t}\right), \tag{A-12}$$

$$= -\min_{t\in[0,1]} \log\left(\sum_{\mathbf{z}_A} P_A(\mathbf{z}_A)^t Q_A(\mathbf{z}_A)^{1-t} \sum_{\mathbf{z}_{A^c}} P_{A^c|A}(\mathbf{z}_{A^c}|\mathbf{z}_A)^t Q_{A^c|A}(\mathbf{z}_{A^c}|\mathbf{z}_A)^{1-t}\right), \tag{A-13}$$

$$\geq -\min_{t\in[0,1]} \log\left(\sum_{\mathbf{z}_A} P_S(\mathbf{z}_A)^t Q_A(\mathbf{z}_A)^{1-t}\right) = D^*(P_A, Q_A), \tag{A-14}$$

where (A-14) results from Hölder's inequality: For non-negative vectors $\mathbf{v} = [v_k]$ and $\mathbf{w} = [w_k]$ that sum to 1, $\sum_k v_k^t w_k^{1-t} \leq (\sum_k v_k)^t (\sum_k w_k)^{1-t} = 1$ for every $t \in [0,1]$. The inequality in (A-10) is tight iff Hölder's inequality holds with equality. This occurs iff $\mathbf{v} = \mathbf{w}$ (since both vectors need to sum to unity). Thus, for equality to hold in (A-10), we need the conditionals $P_{A^c|A}$ and $Q_{A^c|A}$ to be identical, i.e., (6). This completes the proof. □

2 Proof of Proposition 2

Proof Consider the following collection of events $E_{S'_d} := \{\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) = S'_d\}$ for all $S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}$. Alternatively,

$$E_{S'_d} := \left\{ S'_d = \operatorname*{argmax}_{\tilde{S}_d \in \mathfrak{S}_{k,d}} D(\widehat{P}^{(d)}_{\tilde{S}_d} \,||\, \widehat{Q}^{(d)}_{\tilde{S}_d}) \right\}, \tag{A-15}$$

where the quantities in hats are the empirical distributions. That is $E_{S'_d}$ is the event that the output of the exhaustive search decoder is the non-salient set S'_d.

We now bound the probability of each $E_{S'_d}$ (wrt the probability measure $\mathbb{P}^n$). By Sanov's theorem [1, Ch. 11] applied to the product distribution $P^{(d)}_{S_d\cup S'_d} \times Q^{(d)}_{S_d\cup S'_d}$, we have the upper bound

$$\mathbb{P}^n(E_{S'_d}) \leq (n+1)^{|\mathcal{X}|^{|S_d\cup S'_d|}} \exp(-nJ_{S'_d|S_d}) \leq (n+1)^{|\mathcal{X}|^{2k}} \exp(-nJ_{S'_d|S_d}), \tag{A-16}$$

where the error rate is given as the information projection:

$$J_{S'_d|S_d} = \min_{\nu\in\Gamma_{S'_d|S_d}} D(\nu \,||\, P^{(d)}_{S_d\cup S'_d} \times Q^{(d)}_{S_d\cup S'_d}). \tag{A-17}$$

Note that in the above, we have implicitly applied the contraction principle [2, Ch. 4] to the *continuous* function $f : \mathcal{P}(\mathcal{X}^{2|S_d\cup S'_d|}) \to \mathbb{R}$ given by the recipe

$$f\left((P^{(d)}_{S_d\cup S'_d}, Q^{(d)}_{S_d\cup S'_d})\right) := D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}) - D(P^{(d)}_{S'_d} \,||\, Q^{(d)}_{S'_d}). \tag{A-18}$$

The constraint set $\Gamma_{S'_d|S_d}$ was defined in (13). Note also that the minimum in (A-17) is achieved because the objective function is continuous and the constraint set is compact. Also, the minimizer in (A-17) is achieved at the boundary of the constraint set

$$\Lambda_{S'_d|S_d} := \{\nu = (P,Q) \in \mathcal{P}(\mathcal{X}^{|S_d\cup S'_d|}) : D(P_{S_d} \,||\, Q_{S_d}) \leq D(P_{S'_d} \,||\, Q_{S'_d})\} \tag{A-19}$$

as can be readily checked, i.e., $\nu^* \in \mathrm{Bd}(\Lambda_{S'_d|S_d}) = \Gamma_{S'_d|S_d}$. This follows from the convexity of the KL-divergence objective in (A-17) (see [3, Theorem 2] for the details). Next, we complete the proof by applying the union bound and "largest-exponent-wins" principle [2, 4].

$$\mathbb{P}^n(\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) = \mathbb{P}^n\left(\bigcup_{S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}} E_{S'_d}\right) \tag{A-20}$$

$$\leq \sum_{S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}} \mathbb{P}^n(E_{S'_d}) \tag{A-21}$$

$$\leq \sum_{S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}} (n+1)^{|\mathcal{X}|^{2k}} \exp(-n J_{S'_d|S_d}) \tag{A-22}$$

$$\doteq \exp(-n C(P^{(d)}, Q^{(d)})), \tag{A-23}$$

where the notation $\doteq$ denotes equality in the first order of the exponent[1] $C(P^{(d)}, Q^{(d)})$ was given in (14). Note that the constancy of k and d is crucial in (A-23) (this restrictive assumption is relaxed in Theorem 3). The conclusion in (A-23) means that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}^n(\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) \leq -C(P^{(d)}, Q^{(d)}). \tag{A-24}$$

Together with the trivial lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}^n(\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) \geq -C(P^{(d)}, Q^{(d)}), \tag{A-25}$$

we conclude that the limit exists and equals the error exponent, i.e,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}^n(\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) = C(P^{(d)}, Q^{(d)}). \tag{A-26}$$

This completes the proof. □

3 Proof of Theorem 3

We first state four basic lemmas.

Lemma A-2 *For a continuously differentiable real-valued function $f : A \subset \mathbb{R}^n \to \mathbb{R}$, define the Lipschitz constant*

$$L := \sup_{\mathbf{x}\in A} \|\nabla f(\mathbf{x})\|_\infty = \sup_{\mathbf{x}\in A} \left(\max_{1\leq i\leq n} \left| \frac{\partial f}{\partial x_i}(x_i) \right| \right), \tag{A-27}$$

and assume $L < \infty$. Then, we have the Lipschitz *condition*

$$\forall\, \mathbf{x}, \mathbf{y} \in A, \qquad |f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|_1. \tag{A-28}$$

Remark In fact this claim holds for any pair of conjugate exponents[2] $p, q \in [1, \infty]$, i.e., if the ∞ norm in (A-27) is replaced by p norm and the 1 norm in (A-28) is replaced by q norm.

Lemma A-3 *The following bound for the binomial coefficient holds:*

$$\binom{d}{k} \leq \exp\left(d H_b(\frac{k}{d}) \right) \leq \exp\left[k \left(\log(\frac{d}{k}) + 1 \right) \right], \tag{A-29}$$

where H_b is the binary entropy function.

[1] We say that two positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are equal to first order in the exponent (written $a_n \doteq b_n$) if $\lim_{n\to\infty} n^{-1} \log(a_n/b_n) = 0$.

[2] p and q are called conjugate exponents if $1/p + 1/q = 1$.

Lemma A-4 *Let n be a positive integer and $\epsilon \in (0,1)$. Then the following relation*[3] *holds*

$$\binom{n+n^{1-\epsilon}}{n} \in e^{o(n)}, \tag{A-30}$$

where the binomial coefficient defined in terms of Gamma functions, namely

$$\binom{n+n^{1-\epsilon}}{n} := \frac{\Gamma(n+n^{1-\epsilon}+1)}{\Gamma(n^{1-\epsilon}+1)\Gamma(n+1)}. \tag{A-31}$$

Lemma A-5 *For two distributions Q_1, Q_2 with the same support Ω (a finite set), we have*

$$\frac{\partial D(Q_1 \,||\, Q_2)}{\partial Q_1(a)} = 1 + \log \frac{Q_1(a)}{Q_2(a)}, \qquad \frac{\partial D(Q_1 \,||\, Q_2)}{\partial Q_2(a)} = -\frac{Q_1(a)}{Q_2(a)}, \qquad \forall\, a \in \Omega. \tag{A-32}$$

We defer the proofs of the first three lemmas to after the proof of the theorem. The fourth follows by simple calculus and is thus omitted. We now prove the theorem assuming Lemmas A-2 – A-5.

Proof of Theorem 3
Step 1: We first prove that the family of differentiable functions (indexed by d) $h_d : \mathcal{P}(\mathcal{X}^{2|S_d \cup S_d'|}) \to \mathbb{R}$ given by the recipe

$$h_d\left((P^{(d)}_{S_d \cup S_d'}, Q^{(d)}_{S_d \cup S_d'})\right) := D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}) - D(P^{(d)}_{S_d'} \,||\, Q^{(d)}_{S_d'}), \tag{A-33}$$

is *equi-Lipschitz continuous in the l_1 norm*, i.e., there exists a $L' < \infty$ (independent of d), such that

$$\forall\, d \in \mathbb{N}, \quad \forall\, \nu = (P^{(d)}_{S_d \cup S_d'}, Q^{(d)}_{S_d \cup S_d'}),\, \tilde{\nu} = (\widetilde{P}^{(d)}_{S_d \cup S_d'}, \widetilde{Q}^{(d)}_{S_d \cup S_d'}), \qquad |h_d(\nu) - h_d(\tilde{\nu})| \le L' \|\nu - \tilde{\nu}\|_1, \tag{A-34}$$

To prove this first claim, we first argue that $\nu, \tilde{\nu}$ defined in (A-34) satisfy condition A3, i.e., the log-likelihood ratio between the distributions $P^{(d)}_{S_d \cup S_d'}$ and $Q^{(d)}_{S_d \cup S_d'}$ is uniformly bounded (by L). By using A1 and A3 (which says that the log-likelihood ratio of $P^{(d)}_{S_d}$ and $Q^{(d)}_{S_d}$ is uniformly bounded by L), we conclude that

$$\forall\, \mathbf{x}_{S_d \cup S_d'} \in \mathcal{X}^{|S_d \cup S_d'|}, \qquad \log \frac{P^{(d)}_{S_d \cup S_d'}(\mathbf{x}_{S_d \cup S_d'})}{Q^{(d)}_{S_d \cup S_d'}(\mathbf{x}_{S_d \cup S_d'})} \in [-L, L], \tag{A-35}$$

because the union of a non-salient set to the salient set S_d does not change the log-likelihood ratio (cf. the argument after Proposition 1). Thus, the L-boundedness condition also holds for $P^{(d)}_{S_d \cup S_d'}$ and $Q^{(d)}_{S_d \cup S_d'}$. Denote the set of such distributions (where the log-likelihood ratio is bounded by L) as $\mathcal{D}_L$. By evaluating the partial derivative of the KL-divergences in (A-33) with respect to each of its components and applying Lemma A-5 repeatedly, we conclude that the l_∞ norm of the gradient vector of each function h_d in (A-33) is uniformly bounded, i.e., there exists a $L' < \infty$ such that

$$\sup_{(P^{(d)}_{S_d \cup S_d'}, Q^{(d)}_{S_d \cup S_d'}) \in \mathcal{D}_L} \left\| \nabla h_d \left((P^{(d)}_{S_d \cup S_d'}, Q^{(d)}_{S_d \cup S_d'}) \right) \right\|_\infty = L'. \tag{A-36}$$

In fact, we can verify directly from Lemma A-5 that $L' = \max\{2e^L, 2L+2\} < \infty$. Now since the right-hand side of (A-36) is independent of d, we can take the supremum over all d on the left-hand side, i.e.,

$$\sup_{d \in \mathbb{N}} \left\{ \sup_{(P^{(d)}_{S_d \cup S_d'}, Q^{(d)}_{S_d \cup S_d'}) \in \mathcal{D}_L} \left\| \nabla h_d \left((P^{(d)}_{S_d \cup S_d'}, Q^{(d)}_{S_d \cup S_d'}) \right) \right\|_\infty \right\} = L'. \tag{A-37}$$

[3]The asymptotic notation $h(n) \in e^{o(n)}$ means that $\log h(n)$ is a sublinear function, i.e., to every ϵ, there is a $N \in \mathbb{N}$ such that $\log h(n) < \epsilon n$ for all $n > N$.

Finally apply Lemma A-2 to every $d \in \mathbb{N}$ to conclude that the equi-Lipschitz continuity condition (A-34) for the family of functions $\{h_d\}_{d\in\mathbb{N}}$ in (A-33) holds with equi-Lipschitz constant L'.

Step 2: Now, most importantly, we prove that $B > 0$, where B is defined in (16). Assume, to the contrary, that $B = 0$ (since B cannot be negative). For a set of distributions Γ, let $D(\Gamma \,||\, \mu) := \min_{\nu\in\Gamma} D(\nu \,||\, \mu)$. By the definition of B and the infimum, there exists a $d \in \mathbb{N}$ (and a minimizing non-salient set S'_d) such that the divergence satisfies

$$D(\Gamma_{S'_d|S_d} \,||\, P^{(d)}_{S'_d\cup S_d} \times Q^{(d)}_{S'_d\cup S_d}) < \left(\frac{\eta}{2L'\sqrt{2\log 2}}\right)^2. \tag{A-38}$$

The quantity η was defined in (10) and represents how distinguishable the salient set S_d is from the non-salient sets $S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}$. The quantity $L' < \infty$ is the equi-Lipschitz constant in (A-37). Let ν be the product distribution $P^{(d)}_{S'_d\cup S_d} \times Q^{(d)}_{S'_d\cup S_d}$ and ν^* be the minimizer of the optimization problem in the information projection (15) or equivalently (A-17), i.e.,

$$\nu^* := \operatorname{argmin} \left\{ D(\nu \,||\, P^{(d)}_{S'_d\cup S_d} \times Q^{(d)}_{S'_d\cup S_d}) : \nu \in \Gamma_{S'_d|S_d} \right\}. \tag{A-39}$$

Now referring back to (A-34) and applying Pinsker's inequality [1, Ch. 11], we have the chain of inequalities

$$|h_d(\nu) - h_d(\nu^*)| \le L'\|\nu - \nu^*\|_1 \le L'\sqrt{2\log 2}\sqrt{D(\Gamma_{S'_d|S_d} \,||\, P^{(d)}_{S'_d\cup S_d} \times Q^{(d)}_{S'_d\cup S_d})} < \frac{\eta}{2}, \tag{A-40}$$

where the final inequality is because of (A-38). Notice how the finiteness and uniformity (independence from d) of L' are crucial in (A-38) and (A-40). Consequently, $h_d(\nu) \ge \eta$ (by assumption A2 on η-distinguishability) and $h_d(\nu^*) = 0$ (because $\nu^* \in \Gamma_{S'_d|S_d}$ by compactness of the constraint set $\Gamma_{S'_d|S_d}$). Thus,

$$|h_d(\nu) - h_d(\nu^*)| = h_d(\nu) - h_d(\nu^*) \ge \eta \tag{A-41}$$

and from (A-40), we conclude that $\eta < \eta/2$, which is a contradiction. Hence $B > 0$.

Step 3: Now we simply put together the pieces in the proof by upper bounding the error probability p_n, defined in (9). Indeed, we have

$$\begin{aligned}
\mathbb{P}^n(\psi_n(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) &\le \sum_{S'_d\in\mathfrak{S}_{k,d}\setminus\{S_d\}} \mathbb{P}^n(E_{S'_d}), && \text{(A-42)}\\
&\le \sum_{l=0}^{k-1} \binom{k}{l}\binom{d-k}{k-l} \max_{S'_d\in\mathfrak{S}_{k,d}\setminus\{S_d\}} \mathbb{P}^n(E_{S'_d}), && \text{(A-43)}\\
&\le \sum_{l=0}^{k-1} \binom{k}{l}\binom{d-k}{k-l} \max_{S'_d\in\mathfrak{S}_{k,d}\setminus\{S_d\}} \binom{n+|\mathcal{X}|^{|S_d\cup S'_d|}-1}{n} \exp(-nJ_{S'_d|S_d}), && \text{(A-44)}\\
&\le \sum_{l=0}^{k-1} \binom{k}{l}\binom{d-k}{k-l} \binom{n+|\mathcal{X}|^{2k}-1}{n} \exp(-nB), && \text{(A-45)}\\
&\le \sum_{l=0}^{k-1} \exp(k)\exp\left[k\left(\log(\frac{d-k}{k})+1\right)\right] \binom{n+|\mathcal{X}|^{2k}}{n} \exp(-nB), && \text{(A-46)}\\
&< k\exp\left[k\left(\log(\frac{d-k}{k})+2\right)\right] \binom{n+n^{1-\epsilon}}{n} \exp(-nB), && \text{(A-47)}\\
&\le \exp\left[k\log(\frac{d-k}{k})\right]\exp(2k+\log k)\binom{n+n^{1-\epsilon}}{n}\exp(-nB), && \text{(A-48)}\\
&\le \exp\left[k\log(\frac{d-k}{k})\right]\exp(o(n))\exp(-nB), && \text{(A-49)}
\end{aligned}$$

where

- (A-42) follows from the union bound and definition of the event $E_{S'_d}$ given in the proof of Proposition 2 (cf. (A-15)).
- (A-43) follows by a simple counting argument that the number of non-salient sets S'_d that overlap with S_d in l indices is exactly $\binom{k}{l}\binom{d-k}{k-l}$. We also upper bound the probability $\mathbb{P}^n(E_{S'_d})$ by the largest possible probability.
- (A-44) follows from Sanov's theorem and the fact that the number of types [5] with denominator n for a distributions with support $\mathcal{X}^{|S_d \cup S'_d|}$ is precisely $\binom{n+|\mathcal{X}|^{|S_d \cup S'_d|}-1}{n}$.
- (A-45) follows from the definition of $B > 0$ in (16) (infimum over all error rates over all d) and the fact that $|S_d \cup S'_d| \leq 2k$ (because $|S_d| = |S'_d| = k$). Notice how the positivity of B, proved in Step 2, is crucial here.
- (A-46) follows from two applications of Lemma A-3. In particular, we note that $\binom{k}{l} \leq \exp(kH_b(l/k)) \leq \exp(k)$ (for every $l = 0, 1, \ldots, k-1$) and also $\binom{d-k}{k-l}$ is maximized when $l = 0$. We also employ a trivial upper bound of the second binomial coefficient.
- (A-47) follows from the fact that there are only k terms in the sum and assumption that there exists a ϵ such that
$$k < \frac{(1-\epsilon)\log n}{2\log|\mathcal{X}|} \quad \Longleftrightarrow \quad \exp\left(\frac{2k\log|\mathcal{X}|}{1-\epsilon}\right) < n. \tag{A-50}$$
This is given by the function g_1 in (17).
- (A-48) follows by simple rearrangement. Note that $\exp(2k + \log k) \in \exp(o(n))$ by (A-50).
- Lastly (A-49) follows from Lemma A-4 and the absorption of all subexponential terms into $\exp(o(n))$.

Finally, from (A-49), we notice by a simple rearrangement that the exponent is given by $-n(B - o(1) - (k/n)\log((d-k)/k))$. In order to ensure that the error probability decays to zero, it is suffices to have

$$B - o(1) - \frac{k}{n}\log(\frac{d-k}{k}) > 0. \tag{A-51}$$

Condition (A-51) holds if for sufficiently large n

$$n > \frac{k}{B-\epsilon'}\log(\frac{d-k}{k}), \tag{A-52}$$

Take $\epsilon' \to 0$. We conclude from (A-50) and (A-52) that if $n > g_1(k,\epsilon) \vee g_2(d,k)$, then $\{(n,d,k)\}_{n\in\mathbb{N}}$ is achievable, where g_1 and g_2 were defined in (17). Now it is easy to see that the rate of decay $\limsup_{n\to\infty} n^{-1}\log p_n$ is simply given by $-c$ where c is the difference between B and the contribution from the binomial coefficient term $\binom{d-k}{k}$, i.e.,

$$c = B - \limsup_{n\to\infty} \frac{k}{n}\log(\frac{d-k}{k}), \tag{A-53}$$

which concludes the proof of Theorem 3. □

Now we prove the remaining lemmas.

Proof of Lemma A-2
Consider $n = 2$. The general case is easily deducible by extending the argument below straightforwardly. Let $\mathbf{x} = (x_1, x_2), \mathbf{y} = (y_1, y_2) \in A \subset \mathbb{R}^2$ be any two points.

$$\begin{aligned}
|f(x_1, x_2) - f(y_1, y_2)| &= |f(x_1, x_2) - f(y_1, x_2) + f(y_1, x_2) - f(y_1, y_2)| && \text{(A-54)}\\
&\leq |f(x_1, x_2) - f(y_1, x_2)| + |f(y_1, x_2) - f(y_1, y_2)| && \text{(A-55)}\\
&= \left|\frac{\partial f}{\partial x_1}(\xi_1)\right| |x_1 - y_1| + \left|\frac{\partial f}{\partial x_2}(\xi_2)\right| |x_2 - y_2| && \text{(A-56)}\\
&\leq \sup_{\xi_1:(\xi_1,y_1)\in A} \left|\frac{\partial f}{\partial x_1}(\xi_1)\right| |x_1 - y_1| + \sup_{\xi_2:(x_2,\xi_2)\in A} \left|\frac{\partial f}{\partial x_2}(\xi_2)\right| |x_2 - y_2| && \text{(A-57)}\\
&\leq L(|x_1 - y_1| + |x_2 - y_2|) = L\|\mathbf{x} - \mathbf{y}\|_1, && \text{(A-58)}
\end{aligned}$$

where in (A-56) we have made use of the 1-dimensional mean-value theorem [6, Ch. 5] and $\xi_j \in (x_j, y_j)$ for $j = 1, 2$ and in (A-57) and (A-58) we made use of the hypothesis in the lemma (cf. (A-27)). The claim thus follows. □

Proof of Lemma A-3
From [1, Ch. 11], we have the straightforward upper bound

$$\binom{d}{k} \leq \exp\left(dH_b(\frac{k}{d})\right). \tag{A-59}$$

It remains to bound the binary entropy function $H_b(q)$ for $q \in [0, 1]$. Note that for all $0 \leq q \leq 3$,

$$-(1-q)\log(1-q) \leq -(1-q)(-q + \frac{q^2}{2}) = q - \frac{3}{2}q^2 + \frac{q^3}{2} \leq q, \tag{A-60}$$

where we have used the fact that $\log(1-t) \geq -t + t^2/2$. Thus, we have

$$H_b(q) = -q\log q - (1-q)\log(1-q) \leq -q\log q + q = q(-\log q + 1). \tag{A-61}$$

The proof is completed with the identification $q = k/d$ in (A-59). □

Proof of Lemma A-4
We make use of the following bound from [7, Corollary 2.3]:

$$\forall\, \alpha \in \mathbb{R}_+, n \in \mathbb{N}, \qquad \binom{\alpha n}{n} < \frac{1}{\sqrt{2\pi}} n^{-1/2} \frac{\alpha^{\alpha n + 1/2}}{(\alpha - 1)^{(\alpha-1)n + 1/2}}. \tag{A-62}$$

Note from close examination of the proof in [7] that this bound applies to the case where αn may not be an integer. In this case, the binomial coefficient is defined by the one involving Gamma functions (cf. (A-31)). Thus, taking $\alpha = 1 + n^{-\epsilon}$ in (A-62), we have

$$\binom{n + n^{1-\epsilon}}{n} = \binom{n(1 + n^{-\epsilon})}{n} < \text{poly}(n) \frac{(1 + n^{-\epsilon})^{n(1+n^{-\epsilon})}}{(n^{-\epsilon})^{n^{1-\epsilon}}} =: \text{poly}(n) M(n). \tag{A-63}$$

where $\text{poly}(n) \in e^{o(n)}$ is some polynomial function in n. It suffices to prove that $M(n) \in e^{o(n)}$. Indeed,

$$\begin{aligned}
\log M(n) &= n(1 + n^{-\epsilon})\log(1 + n^{-\epsilon}) - n^{1-\epsilon}\log n^{-\epsilon} && \text{(A-64)}\\
&\leq n(1 + n^{-\epsilon})n^{-\epsilon} + \epsilon n^{1-\epsilon}\log n \in o(n) && \text{(A-65)}
\end{aligned}$$

where (A-65) comes from the inequality $\log(1 + t) \leq t$. Thus $M(n) \in e^{o(n)}$ and this completes the proof. □

4 Proof of Corollary 4

Proof Assume that $k = k_0$ is constant. The claim follows by replacing the upper bound for $\binom{d}{k_0}$ in (A-46) with the trivial upper bound d^{k_0}. If $k_0 R < B$, the corresponding exponent (cf. (A-51)) is positive. □

5 Proof of Theorem 5

Proof Recall the Markov chain given in Section III-D of the main paper:

$$S_d \xrightarrow{\varphi_n} (\mathbf{x}^n, \mathbf{y}^n) \xrightarrow{\psi_n} \widehat{S}_d \tag{A-66}$$

Applying Fano's inequality [1, Ch. 1], we have

$$\mathbb{P}^n(S_d \neq \widehat{S}_d) \geq \frac{H(S_d|\widehat{S}_d) - 1}{\log \binom{d}{k}} \tag{A-67}$$

$$= \frac{H(S_d) - I(S_d; \widehat{S}_d) - 1}{\log \binom{d}{k}} \tag{A-68}$$

$$= \frac{\log \binom{d}{k} - I(S_d; \widehat{S}_d) - 1}{\log \binom{d}{k}} \tag{A-69}$$

where (A-69) follows from the uniform distribution on S_d, which implies that $H(S_d) = \log |\mathfrak{S}_{k,d}|$. Now we upper bound the mutual information term:

$$I(S_d; \widehat{S}_d) \overset{(a)}{\leq} I(S_d; \mathbf{x}^n, \mathbf{y}^n) \overset{(b)}{\leq} H(\mathbf{x}^n, \mathbf{y}^n) \overset{(c)}{\leq} n(H(\mathbf{x}) + H(\mathbf{y})) \overset{(d)}{=} n(\mathcal{H}(P^{(d)}) + \mathcal{H}(Q^{(d)})), \tag{A-70}$$

where (a) follows from the data processing inequality (cf. (A-66)), (b) follows from non-negativity of conditional entropy, (c) follows from conditioning reduces entropy and (d) follows from equivalence of $H(\mathbf{x})$ and $\mathcal{H}(P^{(d)})$. Inserting (A-70) into (A-69), we have

$$\mathbb{P}^n(S_d \neq \widehat{S}_d) \geq 1 - \frac{n(\mathcal{H}(P^{(d)}) + \mathcal{H}(Q^{(d)}))}{\log \binom{d}{k}} - o(1) \overset{(a)}{\geq} 1 - \frac{n(\mathcal{H}(P^{(d)}) + \mathcal{H}(Q^{(d)}))}{k \log \frac{d}{k}} - o(1), \tag{A-71}$$

where (a) follows from the fact that $\binom{d}{k} \geq (d/k)^k$. The claim in part (i) thus follows. Note the independence of the proof on the decoder ψ_n. □

6 Proof of Corollary 6

Proof With the added assumption that the conditional entropies are bounded by a linear function in k, i.e., $\max\{\mathcal{H}(P^{(d)}_{S^c_d|S_d}), \mathcal{H}(Q^{(d)}_{S^c_d|S_d})\} \leq Mk$, the entropy decomposes as follows:

$$\mathcal{H}(P^{(d)}) = \mathcal{H}(P^{(d)}_{S_d}) + \mathcal{H}(P^{(d)}_{S^c_d|S_d}) \overset{(a)}{\leq} \log |\mathcal{X}|^k + \mathcal{H}(P^{(d)}_{S^c_d|S_d}) \leq k \log |\mathcal{X}| + Mk = (\log |\mathcal{X}| + M)k, \tag{A-72}$$

where (a) is due to the fact that $P^{(d)}_{S_d} \in \mathcal{P}(\mathcal{X}^k)$ and hence $\mathcal{H}(P^{(d)}_{S_d}) \leq \log |\mathcal{X}|^k$. Substituting this and the corresponding upper bound for $\mathcal{H}(Q^{(d)})$ into (A-71) completes the proof of the claim in part (ii). □

7 Proof of Corollary 7

Proof Take $\lambda = 1$ in (25). Then the claim follows by replacing d in (25) with Ce^{nR} (for some $C > 0$) and further noticing that the inequality is satisfied if and only if

$$R > 2(M + \log|\mathcal{X}|) + \frac{\log k}{n} = 2(M + \log|\mathcal{X}|) + o(1). \tag{A-73}$$

This completes the proof. □

8 Proof of Proposition 8

Proof Recall that k and d are kept constant. We first demonstrate that the Chow-Liu algorithm for learning the common tree is consistent, as for a single tree [3, 8]. Let $\mathcal{T}^d$ be set of trees (or edge sets) with d nodes. Also let $\mathcal{D}(\mathcal{X}^d; \mathcal{T}^d) \subset \mathcal{P}(\mathcal{X}^d)$ be the set of distributions Markov on some tree in $\mathcal{T}^d$. The consistency claim follows from the equivalence of the following optimizations:

$$\min_{\widetilde{P},\widetilde{Q} \in \mathcal{D}(\mathcal{X}^d;\mathcal{T}^d): T_{\widetilde{P}} = T_{\widetilde{Q}}} \quad D((\widehat{P}, \widehat{Q}) \,||\, (\widetilde{P}, \widetilde{Q})), \tag{A-74}$$

$$\min_{\widetilde{P},\widetilde{Q} \in \mathcal{D}(\mathcal{X}^d;\mathcal{T}^d): T_{\widetilde{P}} = T_{\widetilde{Q}}} \quad D(\widehat{P} \,||\, \widetilde{P}) + D(\widehat{Q} \,||\, \widetilde{Q}), \tag{A-75}$$

$$\min_{E_{\widetilde{P}}, E_{\widetilde{Q}} \in \mathcal{T}^d: E_{\widetilde{P}} = E_{\widetilde{Q}}} \quad \sum_{(i,j) \in E_{\widetilde{P}}} I(\widehat{P}_{i,j}) + \sum_{(i,j) \in E_{\widetilde{P}}} I(\widehat{Q}_{i,j}), \tag{A-76}$$

$$\min_{E \in \mathcal{T}^d} \quad \sum_{(i,j) \in E} I(\widehat{P}_{i,j}) + \sum_{(i,j) \in E} I(\widehat{Q}_{i,j}), \tag{A-77}$$

$$\min_{E \in \mathcal{T}^d} \quad \sum_{(i,j) \in E} I(\widehat{P}_{i,j}) + I(\widehat{Q}_{i,j}), \tag{A-78}$$

where (A-76) follows from Chow-Liu [9] and (A-77) follows from enforcing the equality constraint $E_{\widetilde{P}} = E_{\widetilde{Q}}$ into the objective. Thus, the edge weights are indeed given by the sum of the empirical mutual informations.

Furthermore, the KL-divergence is continuous in its arguments. To be more explicit, as $n \to \infty$, $\widehat{D}_i \to D_i$ and $\widehat{W}_{i,j} \to W_{i,j}$ in probability. Thus, the node and edge weights in (28) converge to their true values and the overall algorithm is consistent. The second claim follows from the fact that the complexity of Chow-Liu is $O(nd^2|\mathcal{X}|^2)$ and the complexity of the k-CARD TREE procedure is $O(dk^2)$ [10, 11]. □

9 Relationship between the Exhaustive Search and Maximum Likelihood Decoders

In this section, we will define the Exhaustive Search Decoder (ESD) in terms of the *symmetrized* KL divergence, which we will denote by the symbol D_{sym}:

$$D_{\text{sym}}(P||Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} + Q(x) \log \frac{Q(x)}{P(x)}$$

where $(P, Q) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ for some alphabet $\mathcal{X}$. The analysis in the rest of the paper still holds with this definition of the ESD, but while it is clumsier to use in the other parts, it provides the clearest connection with the Maximum Likelihood Decoder (MLD).

First, we write the Maximum Likelihood decoder in terms of minimizing an asymmetric KL divergence:

$$S^*_{ML} := \arg \min_{S \in \mathfrak{S}_{k,d}} \min_{P_S, Q_S, W_{S^c|S}} D(\widehat{P} \times \widehat{Q} || P_S W_{S^c|S} \times Q_S W_{S^c|S}) \tag{A-79}$$

Recall that $\mathfrak{S}_{k,d}$ denotes the set of cardinality-k subsets in $\{1,\ldots,d\}$. In the inner optimization, we seek the distributions $P := P_S W_{S^c|S}$ and $Q := Q_S W_{S^c|S}$ that form the closest match to the empirical product distribution $\widehat{P} \times \widehat{Q}$ subject to the constraint that they factor according to the fixed salient set S, c.f. Lemma 1. The outer optimization ranges over all possible sets S of the given size k. This expression encodes the standard maximum likelihood optimization: minimizing the KL divergence from the empirical distribution to the feasible set of models, which in this case takes the form of pairs of distributions with a salient set of size k.

We can expand and transform the above objective for the MLD as follows:

$$D(\widehat{P} \times \widehat{Q} || P_S W_{S^c|S} \times Q_S W_{S^c|S}) = \sum_{(x_1,x_2)\in\mathcal{X}\times\mathcal{X}} \widehat{P}(x_1)\widehat{Q}(x_2) \log \frac{\widehat{P}(x_1)\widehat{Q}(x_2)}{(P_S W_{S^c|S})(x_1)(Q_S W_{S^c|S})(x_2)} \tag{A-80}$$

$$= \sum_{x_1} \widehat{P}(x_1) \log \frac{\widehat{P}(x_1)}{(P_S W_{S^c|S})(x_1)} + \sum_{x_2} \widehat{Q}(x_2) \log \frac{\widehat{Q}(x_2)}{(Q_S W_{S^c|S})(x_2)} \tag{A-81}$$

For each of the above two terms, we can split the summations over the salient and non-salient variable subsets. For the first term, we have

$$\sum_{x_1} \widehat{P}(x_1) \log \frac{\widehat{P}(x_1)}{(P_S W_{S^c|S})(x_1)} = \sum_{x_{1,S}} \widehat{P}_S(x_{1,S}) \log \frac{\widehat{P}_S(x_{1,S})}{P_S(x_{1,S})} + \sum_{x_{1,S}} \widehat{P}_S(x_{1,S}) \sum_{x_{1,S^c}} \widehat{P}^c_S(x_{1,S}) \log \frac{\widehat{P}_S(x_{1,S})}{W_{S^c|S}(x_{1,S})} \tag{A-82}$$

We can rewrite the second term similarly to arrive at the following expression for the objective

$$\sum_{x_S} \widehat{P}_S(x_S) \log \frac{\widehat{P}_S(x_S)}{P_S(x_S)} + \sum_{x_S} \widehat{Q}_S(x_S) \log \frac{\widehat{Q}_S(x_S)}{Q_S(x_S)} + \tag{A-83}$$

$$\left(\sum_{x_S} \widehat{P}_S(x_S) \sum_{x_{S^c}} \widehat{P}_{S^c|S}(x_{S^c}|x_S) \log \frac{\widehat{P}_{S^c|S}(x_{S^c}|x_S)}{W_{S^c|S}(x_{S^c}|x_S)}\right) + \left(\sum_{x_S} \widehat{Q}_S(x_S) \sum_{x_{S^c}} \widehat{Q}_{S^c|S}(x_{S^c}|x_S) \log \frac{\widehat{Q}_{S^c|S}(x_{S^c}|x_S)}{W_{S^c|S}(x_{S^c}|x_S)}\right) \tag{A-84}$$

We can recognize the first two summations as divergences between the empirical distributions restricted to the salient set and the free optimization parameters P_S and Q_S, and therefore these terms can be set to zero. To minimize the terms that remain, the task is to choose a $W_{S^c|S}$ that simultaneously minimizes divergence-like objectives from $\widehat{P}_{S^c|S}$ and $\widehat{Q}_{S^c|S}$.

We can regard the choice of a $W_{S^c|S}$ as choosing a set of conditional distributions, $W_{S^c|S}(\cdot|\bar{x}_S)$, for each fixed $\bar{x}_S \in \mathcal{X}^k$, and the objective decouples so that we are left with a set of optimization problems of the form

$$\min_{\tilde{W}} a \cdot D(\widehat{P}_{S^c|S} || \tilde{W}) + b \cdot D(\widehat{Q}_{S^c|S} || \tilde{W}) \tag{A-85}$$

for the appropriate constants $a := \widehat{P}_S(\bar{x}_S)$ and $b := \widehat{Q}_S(\bar{x}_S)$. Note that there are a set of constants a and b, one pair for each $\bar{x}_S$ and hence for each additive term in the objective, but since we will generally examine these terms individually, we suppress the notational dependence on $\bar{x}_S$.

We can apply calculus to show that the optimizing choice for $\tilde{W}$ is given by

$$\tilde{W}^* = \frac{a \cdot \widehat{P}_{S^c|S} + b \cdot \widehat{Q}_{S^c|S}}{a+b} \tag{A-86}$$

and thus each term can be expressed as

$$a \cdot D(\widehat{P}_{S^c|S} || \frac{a \cdot \widehat{P}_{S^c|S} + b \cdot \widehat{Q}_{S^c|S}}{a+b}) + b \cdot D(\widehat{Q}_{S^c|S} || \frac{a \cdot \widehat{P}_{S^c|S} + b \cdot \widehat{Q}_{S^c|S}}{a+b}) \tag{A-87}$$

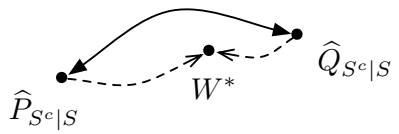

Figure 1: An illustration of the relationship between the terms summed in the Maximum Likelihood Decoder and Exhaustive Search Decoder objectives. The solid line represents the re-weighted symmetrized KL divergence which is used in the ESD objective (c.f. (A-91)). The corresponding term in the MLD objective is represented by the dashed lines, in which KL divergence is measured from each empirical distribution to a convex interpolation W^* (c.f. (A-87)). Lines are curved to emphasize the general non-Euclidean nature of the divergences.

We emphasize that there is one such value for each $\bar{x}_S \in \mathcal{X}^k$, and the sum of all the terms is used as the score for a given subset S.

We can examine the ESD from a similar perspective. The ESD can be written as the optimization

$$S^*_{ESD} := \arg \max_{S \in \mathfrak{S}_{k,d}} D_{\text{sym}}(\widehat{P}_S || \widehat{Q}_S) \tag{A-88}$$

where we have employed the symmetrized definition of KL divergence. The following identity always holds:

$$D_{\text{sym}}(\widehat{P}||\widehat{Q}) = D_{\text{sym}}(\widehat{P}_S||\widehat{Q}_S) + \left(\sum_{x_S} \widehat{P}_S \sum_{x_{S^c}} \widehat{P}_{S^c|S} \log \frac{\widehat{P}_{S^c|S}}{\widehat{Q}_{S^c|S}}\right) + \left(\sum_{x_S} \widehat{Q}_S \sum_{x_{S^c}} \widehat{Q}_{S^c|S} \log \frac{\widehat{Q}_{S^c|S}}{\widehat{P}_{S^c|S}}\right) \tag{A-89}$$

and since for any empirical distribution the left-hand side is a constant, we can rewrite the ESD optimization expression equivalently as

$$S^*_{ESD} := \arg \min_{S \in \mathfrak{S}_{k,d}} \left(\sum_{x_S} \widehat{P}_S \sum_{x_{S^c}} \widehat{P}_{S^c|S} \log \frac{\widehat{P}_{S^c|S}}{\widehat{Q}_{S^c|S}}\right) + \left(\sum_{x_S} \widehat{Q}_S \sum_{x_{S^c}} \widehat{Q}_{S^c|S} \log \frac{\widehat{Q}_{S^c|S}}{\widehat{P}_{S^c|S}}\right) \tag{A-90}$$

To compare this objective to the terms in the MLD objective given by (A-87), we can again examine terms for each fixed $\bar{x}_S \in \mathcal{X}^k$, each of which is given by

$$a \cdot D(\widehat{P}_{S^c|S}||\widehat{Q}_{S^c|S}) + b \cdot D(\widehat{P}_{S^c|S}||\widehat{Q}_{S^c|S}) \tag{A-91}$$

for the same constants a and b in (A-87). As before, these terms are summed to produce a final score for a given set S.

Finally, we can relate the expressions in (A-87) and (A-91) by interpreting their components as similar but distinct divergences between $\widehat{P}_{S^c|S}$ and $\widehat{Q}_{S^c|S}$. The divergence used by the ESD, given in (A-91), can be viewed as a re-weighting of the standard symmetrized KL divergence, while the divergence in the MLD is given by the asymmetric KL divergence to a convex combination of the two distributions. Indeed, numerical experiments show that the two are not equivalent in general. The relationship is summarized in Figure 1.

It is of interest to note that in the regime where $\widehat{P} \approx \widehat{Q}$, the ESD and the MLD become identical. This result can be seen by applying the Euclidean geometric approximations to the divergences in (A-87) and (A-91). More specifically, we can express $\widehat{P} \approx \widehat{Q}$ as:

$$||\widehat{P} - \widehat{Q}||_\infty < \epsilon \tag{A-92}$$

which implies $||\widehat{P}_S - \widehat{Q}_S|| < |\mathcal{X}|^{d-k}\epsilon$. Thus we have, in both (A-87) and (A-91), $a \approx b$. We can thus simplify

(A-87) as:

$$a \cdot D(\widehat{P}_{S^c|S}||\frac{a \cdot \widehat{P}_{S^c|S} + b \cdot \widehat{Q}_{S^c|S}}{a+b}) + b \cdot D(\widehat{Q}_{S^c|S}||\frac{a \cdot \widehat{P}_{S^c|S} + b \cdot \widehat{Q}_{S^c|S}}{a+b}) \tag{A-93}$$

$$\approx a||\widehat{P}_{S^c|S} - \tilde{W}^*||^2 + b||\widehat{Q}_{S^c|S} - \tilde{W}^*||^2 \tag{A-94}$$

$$\approx a\frac{b^2}{(a+b)^2}||\widehat{P}_{S^c|S} - \widehat{Q}_{S^c|S}||^2 + b\frac{a^2}{(a+b)^2}||\widehat{Q}_{S^c|S} - \widehat{P}_{S^c|S}||^2 \tag{A-95}$$

$$\approx \frac{(a+b)}{2}||\widehat{P}_{S^c|S} - \widehat{Q}_{S^c|S}||^2 \approx a||\widehat{P}_{S^c|S} - \widehat{Q}_{S^c|S}||^2 \approx b||\widehat{P}_{S^c|S} - \widehat{Q}_{S^c|S}||^2 \tag{A-96}$$

where we have used the identity $||P - W^*|| = \frac{b}{a+b}||P - Q||$ when $W^* := \frac{a}{a+b}P + \frac{b}{a+b}Q$, and for the last line we have used $a \approx b$. We can directly apply the Euclidean approximation to (A-91) to show the equivalence of the objectives:

$$a \cdot D(\widehat{P}_{S^c|S}||\widehat{Q}_{S^c|S}) + b \cdot D(\widehat{P}_{S^c|S}||\widehat{Q}_{S^c|S}) \approx (a+b)||\widehat{P}_{S^c|S} - \widehat{Q}_{S^c|S}||^2 \tag{A-97}$$

where the scalar factor is unimportant since we are comparing the two expressions as objectives.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[2] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, 1998.

[3] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "Large-deviations for learning tree structures," in *Intl. Symp. Info. Th.*, 2009.

[4] F. D. Hollander, *Large Deviations (Fields Institute Monographs, 14).* American Mathematical Society, Feb 2000.

[5] I. Csiszár, "The method of types," *IEEE Trans. on Info. Th.*, vol. 44, no. 6, pp. 2505–2523, Oct 1998.

[6] W. Rudin, *Principles of Mathematical Analysis.* McGraw-Hill Science/Engineering/Math, 1976.

[7] P. Stănică, "Good Upper and Lower Bounds on Binomial Coefficients," *Journal of Inequalities in Pure and Applied Mathematics*, vol. 2, no. 3, 2003.

[8] C. K. Chow and T. Wagner, "Consistency of an estimate of tree-dependent probability distributions ," *IEEE Transactions in Information Theory*, vol. 19, no. 3, pp. 369 – 371, May 1973.

[9] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees." *IEEE Trans. on Info. Th.*, vol. 14, no. 3, pp. 462–467, May 1968.

[10] M. Fischetti, W. Hamacher, K. Jornsten, and F. Maffioli, "Weighted k-cardinality trees: Complexity and polyhedral structure," *Networks*, vol. 24, pp. 11–21, 1994.

[11] C. Blum, "Revisiting dynamic programming for finding optimal subtrees in trees," *European J. of Ops. Research*, vol. 177, no. 1, 2007.