# Chain rule for squared Hellinger distance (following TS Jayram)

Scribe: Max Lovig

Lecturer: Yihong Wu

Feb 13, 2024

Throughout the note let $X^n \equiv (X_1, \ldots, X_n)$. Let $P_{X^n}$ and $Q_{X^n}$ be $n$-variant joint distributions factorized as

$$P_{X^n} = P_{X_1} P_{X_2|X_1} \cdots P_{X_n|X^{n-1}} \tag{1}$$
$$Q_{X^n} = Q_{X_1} Q_{X_2|X_1} \cdots Q_{X_n|X^{n-1}}. \tag{2}$$

Suppose we have some "distance" (e.g. $f$-divergences or Wasserstein distances) that computes the dissimilarity score between two distributions. A *chain rule* (also known as tensorization) aims to compute or bound the dissimilarity score between the two joint distributions by the scores between each conditionals. The simplest instance is the KL divergence:

$$D(P_{X^n} \| Q_{X^n}) = \sum_{t=1}^{n} \mathbb{E}_P[D(P_{X_t|X^{t-1}} \| Q_{X_t|X^{t-1}})] \tag{3}$$

which follows from the telescoping sum $\log \frac{P_{X^n}}{Q_{X^n}} = \sum_{t=1}^{n} \log \frac{P_{X_t|X^{t-1}}}{Q_{X_t|X^{t-1}}}$. Here $\mathbb{E}_P$ is with respect to the $P$-law, i.e., $X^{t-1} \sim P_{X^{t-1}}$ for each $t$.

How would we extend this to other $f$-divergences? Let us start with a simple chain rule for Hellinger distance (*not* squared):

$$H(P_{X^n}, Q_{X^n}) \le \sum_{t=1}^{n} \mathbb{E}_P[H(P_{X_t|X^{t-1}}, Q_{X_t|X^{t-1}})] \tag{4}$$

To see this, note that $H$ is a metric. Interpolating $P_{X^n}$ and $Q_{X^n}$, we can start with (1) and successively swap out $P$'s by $Q$'s. Define $P^{(t)} \triangleq P_{X_1} P_{X_2|X_1} \cdots P_{X_t|X^{t-1}} Q_{X_{t+1}|X^t} \cdots Q_{X_n|X^{n-1}}$, with $P^{(0)} = Q_{X^n}$ and $P^{(n)} = P_{X^n}$. Then (4) is precisely $H(P^{(0)}, P^{(n)}) \le \sum_{t=0}^{n-1} H(P^{(t)}, P^{(t+1)})$. Since the proof only applies triangle inequality, any distance would work. In particular, we have

$$\mathrm{TV}(P_{X^n}, Q_{X^n}) \le \sum_{t=1}^{n} \mathbb{E}_P[\mathrm{TV}(P_{X_t|X^{t-1}}, Q_{X_t|X^{t-1}})] \tag{5}$$

Compared with (4), the following result[1] due to TS Jayram [Jay09] is considerably deeper and stronger. It says that the desired chain rule holds also for the *squared Hellinger* within a constant factor.

[1] We are grateful to Yanjun Han for telling us this result.

**Theorem 1** (Jayram)**.**

$$H^2(P_{X^n}, Q_{X^n}) \le 4 \cdot \sum_{t=1}^{n} \mathbb{E}_P[H^2(P_{X_t|X^{t-1}}, Q_{X_t|X^{t-1}})] \tag{6}$$

To compare (4) and (6), suppose that each pair of conditionals differ by a Hellinger distance of $\epsilon$. Then (4) says the Hellinger distance between the full joint is $O(n\epsilon)$ while (6) shows it is actually $O(\sqrt{n}\epsilon)$.

# 1 Proof of Theorem 1

We will prove Theorem 1 through the use of three lemmas.

First, in order to have a "smarter" interpolation between $P_{X^n}$ and $Q_{X^n}$, let us define the following intermediate distribution: For $A \subset [n]$, define $P^A$ by substituting the $t$-th conditional $P_{X_t|X^{t-1}}$ in the factorization (1) of $P_{X^n}$ by $Q_{X_t|X^{t-1}}$ for all $t \in A$. Formally,

$$P^A \triangleq \prod_{t=1}^{n} (P_{X_t|X^{t-1}} \mathbf{1}_{\{t \notin A\}} + Q_{X_t|X^{t-1}} \mathbf{1}_{\{t \in A\}})$$

So $P^{\varnothing} = P_{X^n}$ and $P^{[n]} = Q_{X^n}$. Then

$$P^A = P^{\varnothing} \prod_{t=1}^{n} \left( \frac{Q_{X_t|X^{t-1}}}{P_{X_t|X^{t-1}}} \right)^{\mathbf{1}_{\{t \in A\}}}$$

The following lemma is known as the "cut-and-paste" lemma:

**Lemma 1.** *Let $A, B, C, D \subset [n]$. Denote their indicator vectors by $a, b, c, d \in \{0,1\}^n$. If $a + b = c + d$, then $H^2(P^A, P^B) = H^2(P^C, P^D)$.*

Note: later in our application we only need the following special case. Let $A, B$ be disjoint. Then $H^2(P^A, P^B) = H^2(P^{A \cup B}, P^{\varnothing})$.

*Proof.* Let $r_{X_t|X^{t-1}} \triangleq \frac{Q_{X_t|X^{t-1}}}{P_{X_t|X^{t-1}}}$, we have

$$\begin{aligned}
1 - \frac{1}{2} H^2(P^A, P^B) &= \int \sqrt{P^A P^B} \\
&= \mathbb{E}_{P^{\varnothing}} \left[ \sqrt{\prod_{t=1}^{n} r_{X_t|X^{t-1}}^{a_t + b_t}} \right] \\
&= \mathbb{E}_{P^{\varnothing}} \left[ \sqrt{\prod_{t=1}^{n} r_{X_t|X^{t-1}}^{c_t + d_t}} \right] \\
&= 1 - \frac{1}{2} H^2(P^C, P^D).
\end{aligned}$$

□

The second lemma is an "$\ell_2$" fact:

**Lemma 2.** *Let $P^0, P^1, \ldots, P^m$ be arbitrary probability distributions. Then*[2]

$$\frac{1}{m} \sum_{1 \le s < t \le m} H^2(P^t, P^s) \le \sum_{t=1}^{m} H^2(P^t, P^0) \tag{7}$$

*Proof.*

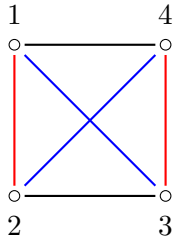
$$\begin{aligned}
2 \cdot \mathrm{LHS} &= \frac{1}{m} \sum_{s,t=1}^{m} H^2(P^s, P^t) \\
&= \frac{1}{m} \sum_{s,t=1}^{m} \int (\sqrt{P^s} - \sqrt{P^t})^2 \\
&= \frac{1}{m} \sum_{s,t=1}^{m} \int (\sqrt{P^s} - \sqrt{P^0} + \sqrt{P^0} - \sqrt{P^t})^2 \\
&= \frac{1}{m} \sum_{s,t=1}^{m} \int (\sqrt{P^s} - \sqrt{P^0})^2 + (\sqrt{P^t} - \sqrt{P^0})^2 - 2(\sqrt{P^s} - \sqrt{P^0})(\sqrt{P^t} - \sqrt{P^0}) \\
&= 2 \cdot \mathrm{RHS} - \frac{2}{m} \int \left( \sum_{t=1}^{m} (\sqrt{P^s} - \sqrt{P^0}) \right)^2 .
\end{aligned}$$

□

The third lemma is a well-known fact in graph factorization.

**Lemma 3.** *Denote by $K_n$ the complete graph with $n$ vertices. For even $n$, $K_n$ can be decomposed into $n-1$ edge-disjoint perfect matchings.*

As a sanity check, the number of edges $\binom{n}{2} = (n-1) \cdot \frac{n}{2}$ match up. For example, the factorization for $n = 4$ is shown below:




For general $n$ there are many constructions. For a geometric one, see https://en.wikipedia.org/wiki/Graph_factorization#2-factorization.

We are now in a position to prove Theorem 1.

*Proof of Theorem 1.* Without loss of generality we can assume that $n = 2^K$ by padding $P_{X^n}$ and $Q_{X^n}$ with additional random variables which are 0.

[2] It is crucial to have no extra constant factors on the RHS. For example, if we apply $(a+b)^2 \le 2a^2 + 2b^2$ in the proof, we get a factor of 2 which would ruin the induction later.

We aim to show the following statement: For any $0 \le k \le K$, any partition $A_1, \dots, A_{2^k}$ of $[n]$ (each subset of size $2^{K-k}$),

$$\sum_{t=1}^{2^k} H^2(P^{A_t}, P^{\varnothing}) \ge c_k \cdot H^2(P^{[n]}, P^{\varnothing}), \quad c_k \triangleq \prod_{i=1}^{k}(1-2^{-i}) \tag{8}$$

and $c_0 \triangleq 1$. Note that $c_k \ge c_\infty \approx 0.289$. Applying this for $k = K$ gives us the desired theorem.

We prove (8) by induction on $k$.

**Base Case** $k = 0$: This is the coarsest partition and (8) trivially holds with equality.

**Induction, from $k-1$ to $k$:** Fix any partition $A_1, \dots, A_{2^k}$ of $[n]$ where $|A_t| = 2^{K-k}$ for all $t \in [2^k]$. Assume that (8) holds for $k-1$. Applying Lemma 3 with $2^k$ vertices yields an edge-disjoint partition of $K_{2^k}$ as $\{E_a : a = 1, \dots, 2^k - 1\}$, where each $E_a$ is a perfect matching between $2^{k-1}$ pairs of vertices. Then

$$\sum_{t=1}^{2^k} H^2(P^{A_t}, P^{\varnothing}) \overset{\text{Lemma 2}}{\ge} \frac{1}{2^k} \sum_{1 \le s < t \le 2^k} H^2(P^{A_s}, P^{A_t}) \tag{9}$$

$$\overset{\text{Lemma 1}}{=} \frac{1}{2^k} \sum_{1 \le s < t \le 2^k} H^2(P^{A_s \cup A_t}, P^{\varnothing}) \tag{10}$$

$$\overset{\text{Lemma 3}}{=} \frac{1}{2^k} \sum_{a=1}^{2^k-1} \sum_{\{s,t\} \in E_a} H^2(P^{A_s \cup A_t}, P^{\varnothing}). \tag{11}$$

Note that for each $a$, because $E_a$ is a perfect matching, $\{A_s \cup A_t : \{s,t\} \in E_a\}$ is a (coarser) partition of $[n]$ consisting of $2^{k-1}$ subsets each of cardinality $2^{K-k+1}$. Applying the induction hypothesis, we conclude

$$\sum_{\{s,t\} \in E_a} H^2(P^{A_s \cup A_t}, P^{\varnothing}) \ge c_{k-1} H^2(P^{[n]}, P^{\varnothing}).$$

Combining the above two displays yields

$$\sum_{t=1}^{2^k} H^2(P^{A_t}, P^{\varnothing}) \ge \underbrace{(1-2^{-k})c_{k-1}}_{c_k} H^2(P^{[n]}, P^{\varnothing}), \tag{12}$$

completing the proof. □

## References


[Jay09] TS Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 562–573. Springer, 2009.