Conditional Independence on the Offset Graph: When Skip-Bigram Conditionals Can Be Combined

Claude and MJC

February 12, 2026

Abstract

We formalize the *shared-offset catastrophe*: naïve Bayesian combination of 8 skip-bigram conditionals $P(o \mid x_{t-d_1}, x_{t-d_2})$ gives 27 bpc (worse than the uniform 8 bpc) when all pairs share offset $d = 1$. We define the *offset graph* $G = (V, E)$ where vertices are temporal offsets and edges are offset pairs used in conditionals. We prove that two conditionals can be combined via Bayes if and only if their edges are vertex-disjoint in G. We give the correct combination formula for overlapping offsets (the *absorption rule*) and show that the optimal offset graph for K pairs is a perfect matching on $2K$ vertices, achieving $O(K)$ bits of information with $O(1)$ bits of redundancy. The path to <2 bpc requires 8 disjoint pairs using 16 distinct offsets, each contributing independent conditional information.

1 The Bayesian Combination Problem

Given a sequence $x_0, x_1, \ldots$ over alphabet Σ (with $|\Sigma| = 256$ for bytes), a skip-bigram conditional at offset pair (d_1, d_2) is:

$$P_k(o \mid x_{t-d_{k,1}}, x_{t-d_{k,2}}) = \frac{c(x_{t-d_{k,1}}, x_{t-d_{k,2}}, o)}{c(x_{t-d_{k,1}}, x_{t-d_{k,2}})}, \tag{1}$$

where $c(\cdot)$ counts joint occurrences in the training data and $o = x_t$ is the byte to predict.

The naïve Bayesian combination of K such conditionals is:

$$P_{\mathrm{NB}}(o) \propto P(o)^{1-K} \prod_{k=1}^{K} P_k(o \mid \mathrm{pair}_k). \tag{2}$$

This assumes the K conditionals are *independent* given o: the evidence from pair k is independent of the evidence from pair k'.

Definition 1 (Shared offset)**.** *Pairs* (d_1, d_2) *and* (d_3, d_4) share an offset *if* $\{d_1, d_2\} \cap \{d_3, d_4\} \neq \emptyset$.

2 The Shared-Offset Catastrophe

Theorem 2 (Bayesian combination with shared offsets)**.** *Let K offset pairs all share a common offset d^*, i.e., pair $k = (d^*, d'_k)$ for $k = 1, \ldots, K$. Then the naïve Bayesian combination* (**??**) *over-counts the evidence from offset d^* by a factor of K:*

$$P_{\mathrm{NB}}(o) \propto P(o)^{1-K} \cdot P(o \mid x_{t-d^*})^K \cdot \prod_{k=1}^{K} \frac{P_k(o \mid x_{t-d^*}, x_{t-d'_k})}{P(o \mid x_{t-d^*})}. \tag{3}$$

The first factor ($P(o)^{1-K}$) *cancels* $K-1$ *copies of the prior. The second factor* ($P(o \mid x_{t-d^*})^K$) *raises the single-offset conditional to the* K*-th power, creating extreme overconfidence.*

Proof. Each conditional decomposes as:

$$P_k(o \mid x_{t-d^*}, x_{t-d'_k}) = P(o \mid x_{t-d^*}) \cdot \frac{P_k(o \mid x_{t-d^*}, x_{t-d'_k})}{P(o \mid x_{t-d^*})}.$$

The product over k gives $P(o \mid x_{t-d^*})^K$ times the product of correction ratios. In the naïve formula, the prior cancellation removes $K-1$ copies of $P(o)$ but NOT the redundant copies of $P(o \mid x_{t-d^*})$. □

Example 3 (The empirical catastrophe)**.** *For the 1024-byte model, the top 8 pairs by MI are:* $(1,7)$, $(1,8)$, $(1,3)$, $(1,20)$, $(1,12)$, $(1,2)$, $(1,6)$, $(1,4)$. *All share* $d^* = 1$. *The naïve combination gives:*

K	*Naive Bayes (bpc)*	*Independent (bpc)*
1	*3.31*	*3.31*
2	*8.17*	*2.80*
4	*14.3*	*2.15*
8	*27.2*	*1.40*

The "Independent" column is the hypothetical result if the same MI were contributed by disjoint pairs (each with unique offsets).

3 The Offset Graph

Definition 4 (Offset graph)**.** *The* offset graph $G = (V, E)$ *has:*

- *Vertices* $V = \{d_1, d_2, \ldots\}$, *the temporal offsets used across all pairs.*
- *Edges* $E = \{(d_{k,1}, d_{k,2})\}_{k=1}^K$, *one edge per offset pair.*

G *is a simple graph (at most one edge per vertex pair). Each edge corresponds to one skip-bigram conditional.*

Definition 5 (Vertex-disjoint edges)**.** *Two edges* $e_1 = (a,b)$ *and* $e_2 = (c,d)$ *are* vertex-disjoint *if* $\{a,b\} \cap \{c,d\} = \emptyset$. *A set of pairwise vertex-disjoint edges is called a* matching *in* G.

Theorem 6 (Independence criterion)**.** *Two skip-bigram conditionals* $P_1(o \mid x_{t-a}, x_{t-b})$ *and* $P_2(o \mid x_{t-c}, x_{t-d})$ *are conditionally independent given* o *(the output byte) if and only if:*

1. *Their edges* (a,b) *and* (c,d) *are vertex-disjoint:* $\{a,b\} \cap \{c,d\} = \emptyset$.
2. *The data process* $\{x_t\}$ *satisfies:* $(x_{t-a}, x_{t-b}) \perp (x_{t-c}, x_{t-d}) \mid x_t$.

Condition 2 is a property of the data. For i.i.d. data, it holds automatically (all offsets are independent). For structured data like text, it is approximately satisfied when the offsets are "far enough apart" that the correlations decay. The key point: condition 1 is *necessary*—shared offsets guarantee dependence regardless of the data process.

Proof of necessity. If (a,b) and (c,d) share a vertex, say $a = c$, then both conditionals depend on x_{t-a}. For any data process where x_{t-a} carries information about x_t beyond what x_t itself provides (which is the case whenever $\mathrm{MI}(x_{t-a}; x_t) > 0$), the conditionals are dependent given x_t: knowing $P_1(o \mid x_{t-a}, x_{t-b})$ constrains x_{t-a}, which constrains $P_2(o \mid x_{t-a}, x_{t-d})$. □

4 The Absorption Rule

When two conditionals share an offset, naïve Bayesian combination is wrong. The correct rule is *absorption*: combine the two pairs into a single 3-offset conditional.

Definition 7 (Absorption)**.** *Given pairs* (a, b) *and* (a, c) *sharing offset* a*, their* absorption *is the 3-offset conditional:*

$$P_{\text{abs}}(o \mid x_{t-a}, x_{t-b}, x_{t-c}) = \frac{c(x_{t-a}, x_{t-b}, x_{t-c}, o)}{c(x_{t-a}, x_{t-b}, x_{t-c})}. \tag{4}$$

Proposition 8 (Absorption is correct combination)**.** *The absorbed conditional satisfies:*

$$P_{\text{abs}}(o \mid x_{t-a}, x_{t-b}, x_{t-c}) = P_1(o \mid x_{t-a}, x_{t-b}) \cdot \frac{P(o \mid x_{t-a}, x_{t-b}, x_{t-c})}{P_1(o \mid x_{t-a}, x_{t-b})}.$$

The correction factor accounts for the additional information from x_{t-c} *beyond what* (x_{t-a}, x_{t-b}) *already provides. This is NOT the naïve product; it is the* chain rule *applied correctly.*

Remark 9 (The cost of absorption)**.** *The absorbed 3-offset conditional requires a* $|\Sigma|^3 \times |\Sigma|$ *count table (*$256^3 \times 256 \approx 4 \times 10^9$ *entries for bytes), which is infeasible to populate from limited data. This is the fundamental tension: absorption is statistically correct but requires exponentially more data. The Bayesian combination is data-efficient but wrong when offsets overlap.*

5 The Optimal Offset Graph

Theorem 10 (Optimal structure for K conditionals)**.** *To combine* K *skip-bigram conditionals with valid Bayesian combination:*

1. *The offset graph* G *must be a matching (no shared vertices).*
2. *The matching uses* $2K$ *distinct offsets.*
3. *Each conditional contributes independent information.*
4. *The total information is additive:* $I_{\text{total}} = \sum_{k=1}^{K} I_k$.

Corollary 11 (The disjoint-pair bound)**.** *With* D *available offsets, at most* $K = \lfloor D/2 \rfloor$ *disjoint pairs can be formed. For the RNN with BPTT-50,* $D \leq 50$*, giving* $K \leq 25$ *independent conditionals.*

5.1 Greedy offset selection for matchings

The greedy algorithm from the carrier-signal paper selects offsets by MI. For matchings, we modify it:

1. Compute $\text{MI}(x_t; x_{t-d_1}, x_{t-d_2})$ for all pairs (d_1, d_2).
2. Select the pair with highest MI.
3. Remove both offsets from the available set.
4. Repeat until K pairs are selected or no offsets remain.

Example 12 (Greedy matching vs greedy star)**.** *For the 1024-byte model:*

Selection	*Pair 1*	*Pair 2*	*Pair 3*	*Pair 4*
Greedy star	$(1,7)$	$(1,8)$	$(1,3)$	$(1,20)$
Greedy matching	$(1,7)$	$(8,3)$	$(20,12)$	$(2,6)$

The greedy star has higher per-pair MI (all pairs include the most informative offset $d = 1$) but catastrophic combination. The greedy matching sacrifices per-pair MI for combinability.

6 Information-Theoretic Analysis

Definition 13 (Redundancy)**.** *The* redundancy *of K conditionals is:*

$$R = \sum_{k=1}^{K} I_k - I_{\text{combined}},$$

where $I_k = MI(o; pair_k)$ and $I_{\text{combined}} = MI(o; all\ pairs\ jointly)$.

Theorem 14 (Redundancy bounds)**.**

1. ***Matching (disjoint pairs):*** *$R = O(1)$ bits. The residual redundancy comes from higher-order correlations in the data process, not from shared offsets. For approximately Markov data, $R \to 0$ as offset separation increases.*

2. ***Star (shared center):*** *$R = (K-1) \cdot I_{center}$. The redundancy grows linearly in K, dominated by the over-counted center offset. For $K = 8$ and $I_{center} = 3.3$ bits (offset $d = 1$ for enwik9), $R \approx 23$ bits—explaining the catastrophic 27 bpc.*

3. ***General graph:*** *$R \leq \sum_{v \in V}(\deg(v) - 1) \cdot I_v$, where $\deg(v)$ is the degree of vertex v in the offset graph and $I_v = MI(o; x_{t-v})$ is the single-offset MI. This bounds redundancy by the sum of over-counting at each shared vertex.*

Proof of the star bound. In the star graph, the center vertex v^* has degree K. Each of the K conditionals includes $P(o \mid x_{t-v^*})$ as a factor. The naïve combination raises this to the K-th power, over-counting by $(K-1) \cdot \log_2 P(o \mid x_{t-v^*})$ on average. Taking expectations: $R = (K-1) \cdot I_{v^*}$. □

7 The Correction Formula

For non-matching offset graphs, the correct combination is:

Theorem 15 (Inclusion-exclusion combination)**.** *Given K conditionals with offset graph G, the correct joint conditional is:*

$$\log P(o \mid \textit{all offsets}) = \sum_{\textit{edges}} \log P(o \mid \textit{pair}) - \sum_{\textit{vertices}} (\deg(v)-1) \log P(o \mid x_{t-v}) + \textit{higher-order corrections}. \tag{5}$$

The first sum adds each pair's log-conditional. The second sum subtracts the over-counted single-offset terms (one subtraction per extra degree at each shared vertex). The higher-order corrections involve 3-way and higher interactions at vertices of degree ≥ 3.

For a star graph with center v^* and K pairs:

$$\log P(o \mid \text{all}) \approx \sum_{k=1}^{K} \log P_k(o \mid v^*, d'_k) - (K-1) \log P(o \mid x_{t-v^*}). \tag{6}$$

This subtracts $K - 1$ copies of the single-offset conditional that the naïve formula over-counts.

Example 16 (Corrected star combination at $K = 8$)**.** *For the 1024B model with 8 pairs sharing $d = 1$:*

Method	*bpc*
Naïve Bayes (8 pairs)	*27.2*
Corrected star (subtract 7 copies of $d = 1$)	*3.1*
Greedy matching (4 disjoint pairs)	*2.8*
Absorption (single 9-offset table)	*1.40*

The corrected star recovers reasonable performance but is still worse than the matching (3.1 vs 2.8 bpc) because the correction is an approximation that ignores higher-order interactions.

8 Connection to the RNN

The trained RNN combines information from multiple offsets via the recurrent dynamics $h(t) = \tanh(W_h h(t-1) + W_x x_t + b_h)$. The weight matrix W_h determines HOW offset information combines.

Proposition 17 (The RNN performs implicit absorption)**.** *The recurrent update combines all offset information through matrix multiplication, which is the correct multivariate conditional (not naïve Bayes). The RNN never faces the shared-offset catastrophe because it never factorizes the prediction into independent pair-wise conditionals.*

However, the RNN's capacity is limited:

- 128 hidden units can represent at most 128 bits of context.
- Each neuron computes a 2-offset conjunction ($R^2 = 0.83$).
- With 45 neurons sharing pair $(1, 8)$ and 30 sharing $(1, 7)$, 75/128 neurons are "redundant" in the sense that they encode overlapping offset information.

The analytic construction must solve this same problem differently: it has access to exact data statistics but must combine them correctly. The matching strategy (disjoint pairs) is the analytic analog of the RNN's implicit absorption—both avoid the shared-offset catastrophe, but via different mechanisms.

9 The Path to <2 bpc

Combining all results:

1. **8 disjoint pairs on 16 offsets**: e.g., $(1, 2)$, $(3, 4)$, $(5, 6)$, $(7, 8)$, $(9, 10)$, $(11, 12)$, $(13, 14)$, $(15, 16)$. Each pair contributes ∼4 bits of MI (for the enwik9 data at 10M bytes, where byte MI at offsets 1–16 is 1.5–4.5 bits). Total: ∼20 bits with $O(1)$ bits of redundancy.
2. **KN smoothing per pair**: Each pair uses a KN-smoothed $256^2 \times 256$ table, requiring 16M entries. At 10M bytes, this is feasible with order 2 (the pair itself provides order 2).
3. **Bayesian combination**: With disjoint pairs, the naïve formula (**??**) is correct (condition 1 of Theorem **??** holds). Condition 2 is approximate for text data but the error is small for offsets > 3 apart.

4. **Target**: 8 pairs × 2.5 bits effective MI each = 20 bits total. At 256 outputs, $H(o) = 8$ bits. With 20 bits of context, the conditional entropy $H(o \mid \text{context})$ should be < 2 bits.

Remark 18 (The factor map predicts the matching)**.** *The factor map shows that the RNN's neurons cluster on pairs* $(1, 7)$*,* $(1, 8)$*,* $(1, 3)$*, etc.—all star edges from offset 1. A matching-based analytic model would instead use pairs* $(1, 7)$*,* $(2, 8)$*,* $(3, 9)$*, . . . , each using fresh offsets. This is a* different *and potentially better factorization of the same data: one that sacrifices per-pair MI for combinability.*

The RNN converges to the star because gradient descent optimizes each neuron independently—each neuron finds the locally best pair, which always includes offset 1. The global optimum (the matching) requires joint *optimization across neurons, which gradient descent does not naturally achieve.*

10 Conclusions

1. The shared-offset catastrophe is a violation of the independence assumption in naïve Bayes, caused by graph-theoretic structure (vertex sharing) in the offset graph.

2. The correct criterion for independent combination is vertex-disjointness (matching) in the offset graph.

3. The optimal offset graph for K conditionals is a perfect matching on $2K$ vertices, using $2K$ distinct offsets.

4. The corrected formula for overlapping offsets subtracts $(\deg(v) - 1)$ copies of each shared offset's contribution, an inclusion-exclusion principle on the offset graph.

5. The path to <2 bpc requires replacing the RNN's star structure with a matching structure, trading per-pair MI for combinability.