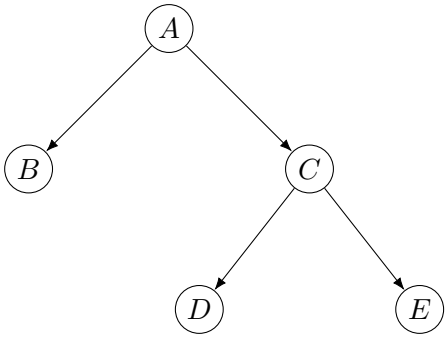

Figure 1: The allowed 5-node hierarchy. Leaves: B, D, E. Internal nodes: A, C.

1 Method: Stop-at-Node Posterior, Partial-Label Likelihood, and Bayes-Risk Decoding

1.1 Ontology, nodes, and the "stop" semantics

We use a rooted cell-type tree. For exposition we use a 5-node toy tree as shows above (1.1). Let the node set be

$$\mathcal{V} = \{A, B, C, D, E\}.$$

Stop variable. For an input cell patch x, we define a latent *stop-at-node* variable

$$S \in \mathcal{V},$$

where $S = v$ means *the model chooses to stop at node v and report v*. Importantly, predicting an internal node (e.g., C) does *not* claim the underlying biological subtype equals C; it means the model only commits to that resolution.

Posterior over nodes (softmax). The posterior tells us how willing the model is to stop at each node, given the input. The model produces logits $z_\theta(x) \in \mathbb{R}^{|\mathcal{V}|}$ and defines a valid distribution over *all nodes*:

$$p_\theta(S = v \mid x) \;=\; \frac{\exp(z_{\theta,v}(x))}{\sum_{u \in \mathcal{V}} \exp(z_{\theta,u}(x))} \qquad \forall v \in \mathcal{V}. \tag{1}$$

By construction,

$$\sum_{v \in \mathcal{V}} p_\theta(S = v \mid x) = 1, \qquad p_\theta(S = v \mid x) \geq 0.$$

Thus every node is a *possible output label* under a unified probabilistic object.

1.2 Ancestor matrix and subtree events

Ancestor indicator matrix. Define the ancestor matrix $A \in \{0,1\}^{|\mathcal{V}| \times |\mathcal{V}|}$ by

$$A_{u,v} = 1 \iff u \text{ is an ancestor of } v \text{ (including } u = v). \tag{2}$$

For the toy tree with node order (A, B, C, D, E), the matrix is

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Row u selects the subtree of u:

$$\text{subtree}(u) = \{v \in \mathcal{V} : A_{u,v} = 1\}.$$

Subtree event probability. Given the posterior distribution over stop nodes $p_\theta(S = \cdot \mid x)$, the probability that the model *stops anywhere within the subtree rooted at* u is

$$p_\theta\big(S \in \text{subtree}(u) \mid x\big) = \sum_{v \in \mathcal{V}} A_{u,v}\, p_\theta(S = v \mid x). \tag{3}$$

Crucially, this quantity should be interpreted as the probability of a *subtree event*, rather than the probability of an exact class. In particular, for an internal node u, the term $p_\theta(S = u \mid x)$ does *not* mean that the true cell type equals u. Instead, it represents the model's decision to *stop at node* u, indicating that the available evidence is insufficient to resolve any of its descendants.

To make this concrete, consider a node C with children D and E. Then

$$p_\theta\big(S \in \text{subtree}(C) \mid x\big) = p_\theta(S = C \mid x) + p_\theta(S = D \mid x) + p_\theta(S = E \mid x).$$

Here, the three terms correspond to mutually exclusive stop decisions: stopping at C itself (unresolved or potentially unknown subtype), stopping at D (resolved to a known subtype), or stopping at E (resolved to another known subtype). All three outcomes are compatible with the coarse observation that the cell belongs to the subtree of C.

This formulation naturally accommodates resolution-limited annotations and open-world structure. Cells that belong to node C but do not match any explicitly modeled descendants are captured by $p_\theta(S = C \mid x)$, without requiring the introduction of ad-hoc "other" classes.

1.3 Observed labels: fine vs. coarse (resolution-limited)

We distinguish two observation types for a training example $(x, \tilde{y})$.

Fine (exact) observation. If the label $\tilde{y} = v$ is *fine* (exact), it asserts

$$S = v.$$

Coarse (partial) observation. If the label $\tilde{y} = u$ is *coarse*, it asserts only the constraint

$$S \in \text{subtree}(u),$$

i.e., the true resolution lies somewhere under u but is not specified. This is precisely the scenario where an internal-node label appears due to annotation/measurement limits.

1.4 Likelihood and loss (why coarse and fine losses differ)

Unified likelihood. Using the event definitions above, the per-sample likelihood is

$$\mathcal{L}_\theta(x, \tilde{y}) = \begin{cases} p_\theta(S = v \mid x), & \text{if } \tilde{y} = v \text{ is fine (exact)}, \\ p_\theta\big(S \in \text{subtree}(u) \mid x\big), & \text{if } \tilde{y} = u \text{ is coarse (partial)}. \end{cases}$$

Taking the negative log gives the training loss:

$$\ell_\theta(x, \tilde{y}) = \begin{cases} -\log p_\theta(S = v \mid x), & \text{if } \tilde{y} = v \text{ is fine}, \\ -\log \sum_{w \in \mathcal{V}} A_{u,w}\, p_\theta(S = w \mid x), & \text{if } \tilde{y} = u \text{ is coarse}. \end{cases} \tag{4}$$

Why we introduce subtree events rather than using $p_\theta(S = u \mid x)$. A coarse observed label $\tilde{y} = u$ does *not* assert that the model should stop exactly at node u. Instead, it asserts only that the true resolution lies somewhere within the subtree rooted at u. In terms of the stop variable, the observation corresponds to the set-valued event

$$S \in \text{subtree}(u),$$

rather than the point event $S = u$.

Consequently, the likelihood of a coarse observation must be defined over this event. Using only $p_\theta(S = u \mid x)$ would incorrectly treat all descendants of u as negative outcomes, forcing probability mass away from valid fine-grained nodes. The subtree event probability in Eq. (4) is therefore not a modeling choice but a direct implication of the semantics of coarse labels.

Why the coarse loss is *not* the fine loss. If u is an internal node (e.g., C), then subtree(u) contains multiple nodes, and the coarse likelihood is a *sum* of probabilities over mutually exclusive stop decisions. In contrast, if v is a leaf node, then subtree$(v) = \{v\}$, and the coarse formula reduces to the fine one:

$$-\log \sum_w A_{v,w}\, p_\theta(S = w \mid x) = -\log p_\theta(S = v \mid x).$$

Thus Eq. (4) is a strict generalization: fine labels are a special case of coarse labels.

Concrete numeric illustration. Suppose the posterior over nodes (A, B, C, D, E) is

$$p_\theta(S = C \mid x) = 0.2, \quad p_\theta(S = D \mid x) = 0.4, \quad p_\theta(S = E \mid x) = 0.4, \quad p_\theta(S = A \mid x) = p_\theta(S = B \mid x) = 0.$$

If the observed label is coarse $\tilde{y} = C$, then the likelihood evaluates the subtree event:

$$p_\theta(S \in \text{subtree}(C) \mid x) = p_\theta(S = C) + p_\theta(S = D) + p_\theta(S = E) = 1,$$

and therefore

$$\ell_\theta(x, C) = -\log(1) = 0.$$

This reflects the fact that the prediction is fully compatible with the coarse observation, regardless of whether the model stops at C itself or commits to one of its descendants.

If the observed label is fine $\tilde{y} = D$, then only the point event $S = D$ is compatible, yielding

$$\ell_\theta(x, D) = -\log p_\theta(S = D) = -\log(0.4) \approx 0.916.$$

This illustrates the fundamental difference in supervision: coarse labels constrain only the subtree, whereas fine labels require commitment to a specific node.

Training versus inference. Subtree constraints appear only in the training loss because they are required to construct a valid likelihood from the observed labels. In particular, coarse labels specify which *subtree* the sample belongs to, and therefore the loss must be defined in terms of the probability of that subtree event. At inference time, however, the model has already predicted a posterior distribution over *individual stop nodes*. The remaining problem is no longer to validate a label, but to choose a single node to output. Risk is introduced at this stage to encode that not all stop decisions are equally desirable: predicting a sibling is worse than predicting a parent, and predicting a distant node is worse than predicting a nearby one.

1.5 Why prediction should not be $\arg\max p_\theta(S = v \mid x)$

Not all mistakes are equally costly. In a hierarchy, predicting a sibling leaf is typically worse than predicting a parent. Therefore we introduce a *risk (cost) matrix* that encodes severity.

Risk matrix from tree distance. Let $R \in \mathbb{R}_+^{|\mathcal{V}|\times|\mathcal{V}|}$ be defined by

$$R_{a,v} = \mathrm{dist}(a, v), \tag{5}$$

where dist is the tree edge distance between nodes. For the toy tree (order (A, B, C, D, E)), the distance-based risk matrix is:

$$R = \begin{pmatrix} 0 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 3 & 3 \\ 1 & 2 & 0 & 1 & 1 \\ 2 & 3 & 1 & 0 & 2 \\ 2 & 3 & 1 & 2 & 0 \end{pmatrix}.$$

Bayes-risk decision rule. Given a posterior $p_\theta(S = \cdot \mid x)$, we output the node that minimizes expected risk:

$$\hat{a}(x) = \arg\min_{a\in\mathcal{V}} \sum_{v\in\mathcal{V}} p_\theta(S = v \mid x)\, R_{a,v}. \tag{6}$$

Why this is the correct decision rule. For any fixed x, choosing output a incurs conditional expected cost

$$\mathbb{E}[R_{a,S} \mid x] = \sum_v p_\theta(S = v \mid x)\, R_{a,v}.$$

Since this quantity depends on a only through the sum above, the minimizer in Eq. (6) is the Bayes-optimal decision for that x (minimizes conditional expected cost). This is standard decision theory: likelihood models uncertainty, risk specifies utility.

Counterexample showing $\arg\max$ **can be suboptimal.** Let

$$p_\theta(S = C \mid x) = 0.2, p_\theta(S = D \mid x) = 0.4, \quad p_\theta(S = E \mid x) = 0.4, \quad \text{others } 0.$$

Then $\arg\max$ predicts D (or E). Compare expected risks:

$$\mathbb{E}[R_{C,S} \mid x] = 0.2 \cdot R_{C,C} + 0.4 \cdot R_{C,D} + 0.4 \cdot R_{C,E} = 0.2 \cdot 0 + 0.4 \cdot 1 + 0.4 \cdot 1 = 0.8.$$

$$\mathbb{E}[R_{D,S} \mid x] = 0.2 \cdot R_{D,C} + 0.4 \cdot R_{D,D} + 0.4 \cdot R_{D,E} + = 0.2 \cdot 1 + 0.4 \cdot 0 + 0.4 \cdot 2 = 1.0,$$

Thus Bayes-risk decoding predicts C (a safer coarse node) even though $p(D)$ is maximal. This formalizes why we introduce risk: it encodes that coarse predictions can be preferable under uncertainty.

1.6 Summary

- Eq. (1) defines a valid posterior distribution over *all nodes* in the ontology. Each node, coarse or fine, is a possible stop decision, and the probabilities are mutually exclusive and sum to one.
- Coarse observed labels do not supervise a specific stop node. Instead, they specify a *subtree membership constraint*, formalized by the ancestor matrix in Eq. (2) and the subtree event probability in Eq. (3). This makes explicit that $p_\theta(S = u \mid x)$ represents stopping at node u, whereas $p_\theta(S \in \mathrm{subtree}(u) \mid x)$ represents belonging to the semantic scope of u.
- Prediction is a separate decision problem. Since different stop decisions incur different semantic costs on the ontology, selecting $\arg\max_v p_\theta(S = v \mid x)$ is generally suboptimal. Eq. (6) gives the Bayes-optimal decision rule under a tree-structured cost matrix, allowing conservative fallback to coarser nodes when uncertainty is high.