

05-12

二叉树

Huffman编码树：正确性

我生来就不像我所见过的任何一个人；我敢断言，我与世上的任何一个人都迥然不同；虽说我不比别人好，但至少我与他们完全两样。

邓俊辉

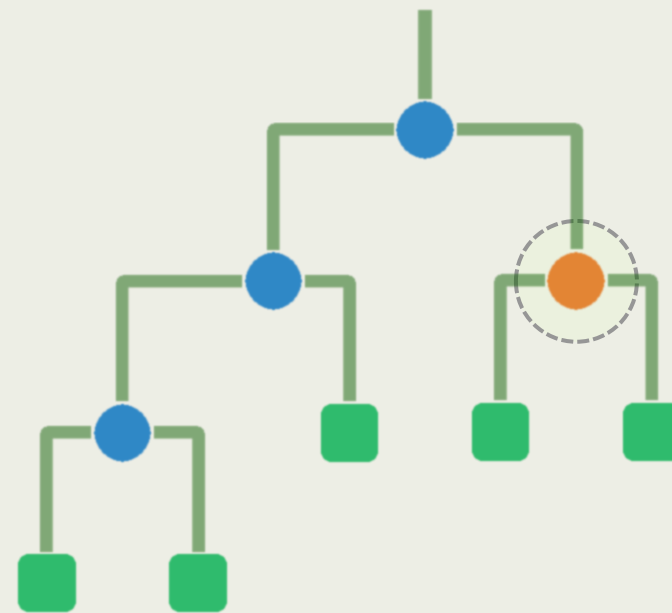
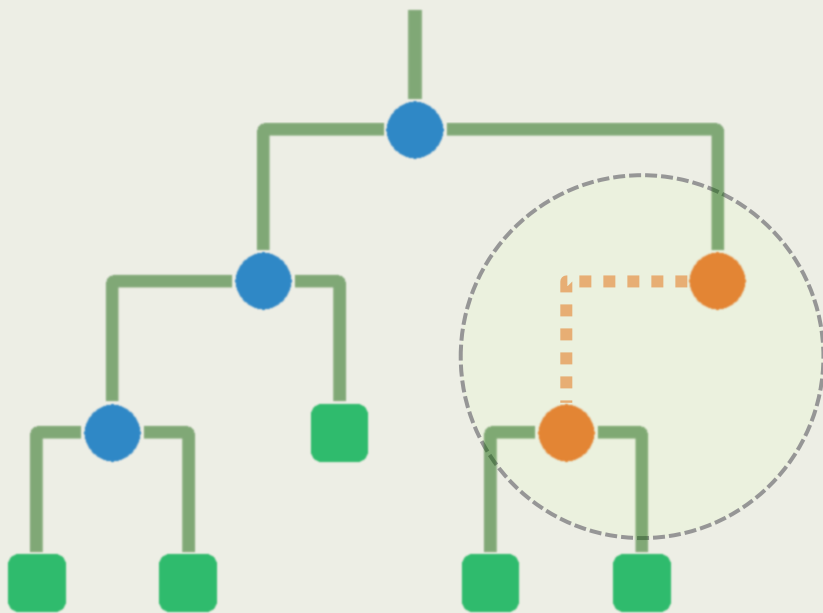
deng@tsinghua.edu.cn

双子性

❖ 最优编码树有何特征？

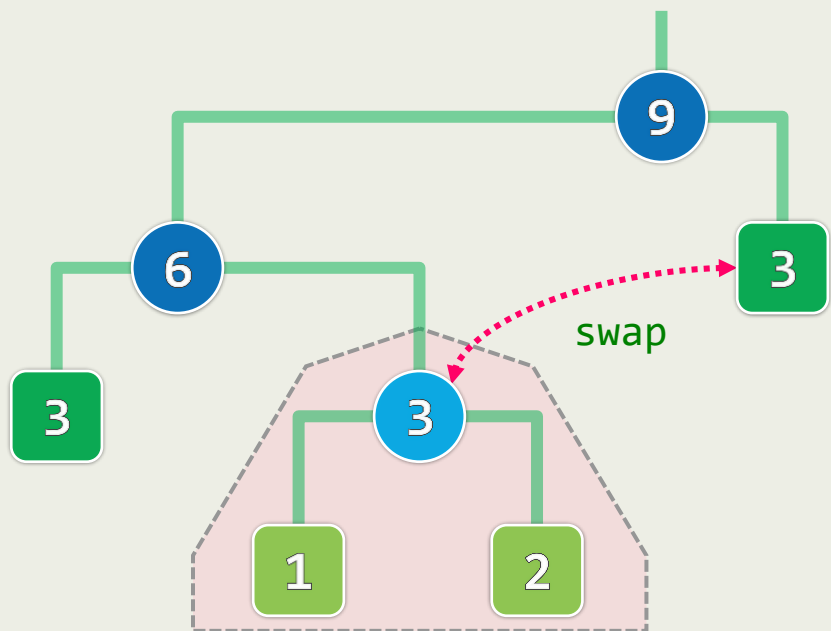
❖ 首先，每一内部节点都有**两个孩子**——节点度数均为偶数（0或2），即**真二叉树**

❖ 否则，将1度节点**替换**为其唯一的**孩子**，则新树的wald将**更小**

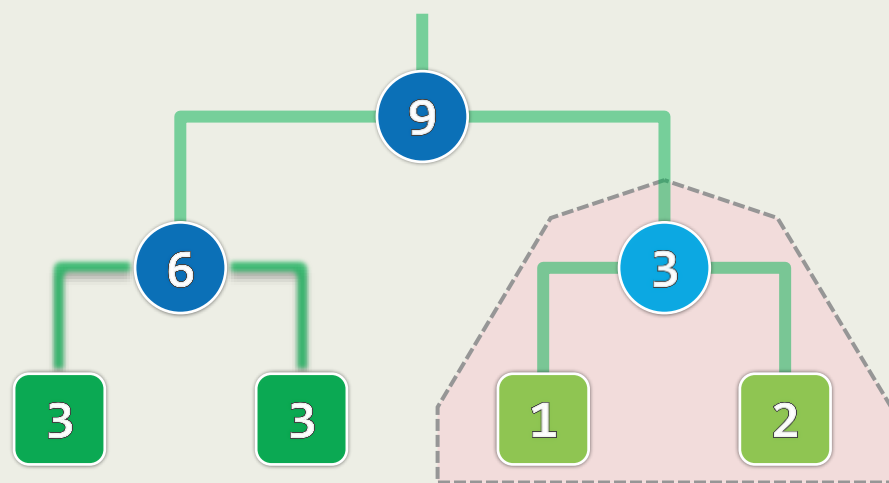


不唯一性

- ❖ 对任一内部节点而言
左、右子树**互换**之后wald不变
- ❖ 上述算法中，**兄弟**子树的次序系**随机**选取
故有可能...

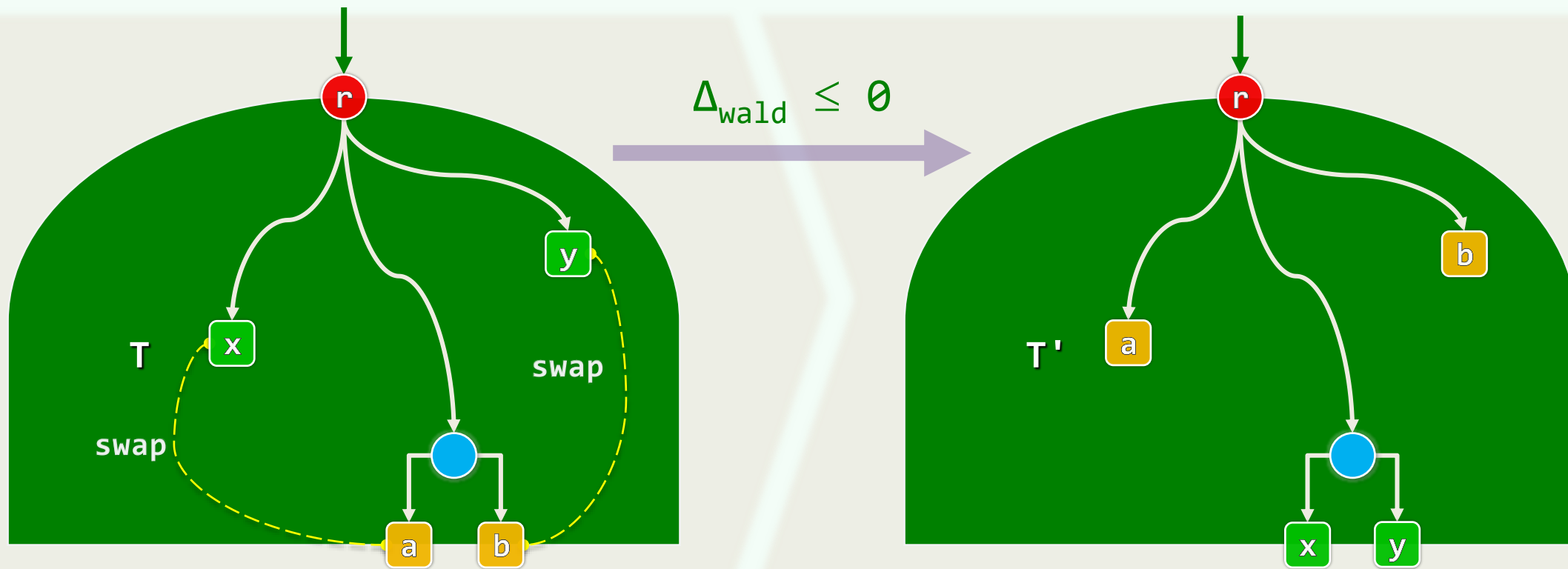


- ❖ 为消除这种歧义，可以（比如）
明确要求**左**子树的频率更**低**
- ❖ 不过，倘若
它们（甚至更多节点）的频率恰好相等...



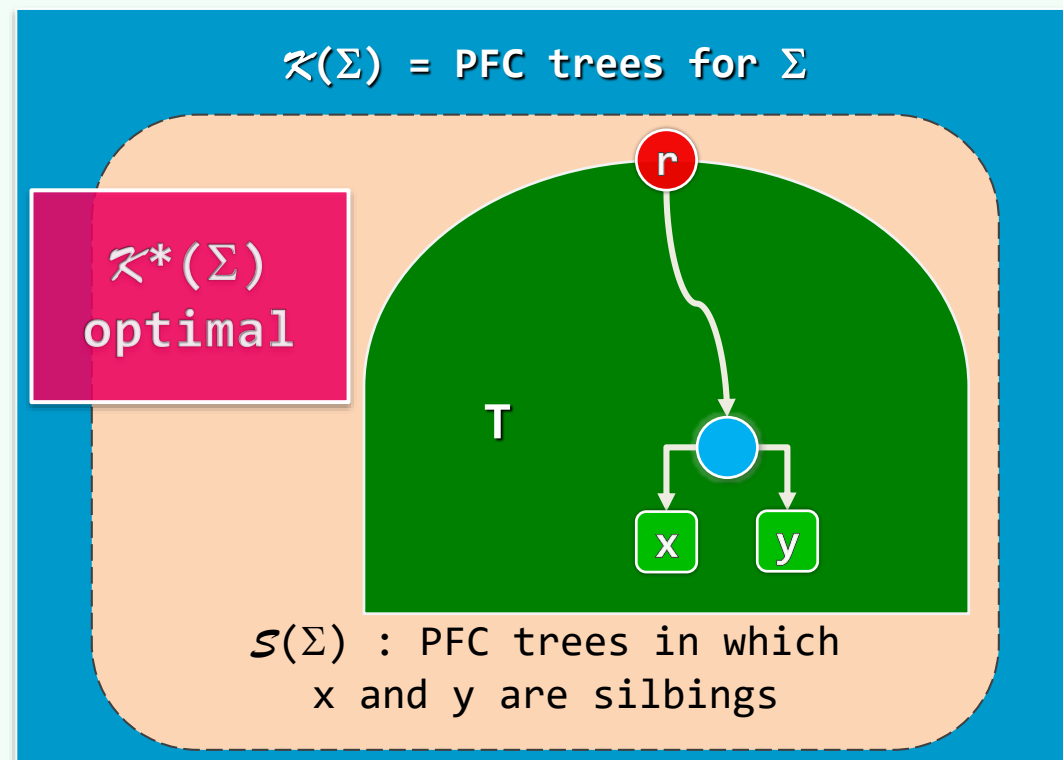
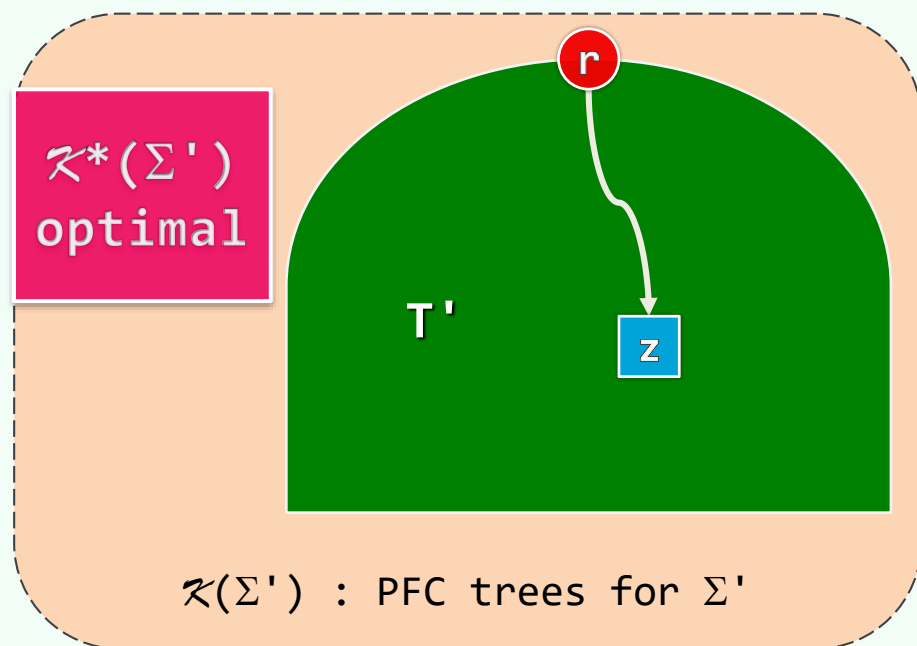
层次性

- ❖ 出现**频率最低**的字符 x 和 y ，必在**某棵**最优编码树中处于**最底层**，且互为**兄弟**
 - ❖ 否则，**任取**一棵最优编码树，并在其最底层**任取**一对兄弟 a 和 b
- 于是， a 和 x 、 b 和 y 交换之后， $wald$ 绝不会增加



数学归纳

- ❖ 对 $|\Sigma|$ 做归纳可证：Huffman算法所生成的，必是一棵**最优**编码树！ $|\Sigma| = 2$ 时显然
- ❖ 设算法在 $|\Sigma| < n$ 时均正确。现设 $|\Sigma| = n$ ，取 Σ 中频率最低的 x 、 y （不妨就设二者互为兄弟）
- ❖ 令： $\Sigma' = (\Sigma \setminus \{x, y\}) \cup \{z\}$ ， $w(z) = w(x) + w(y)$

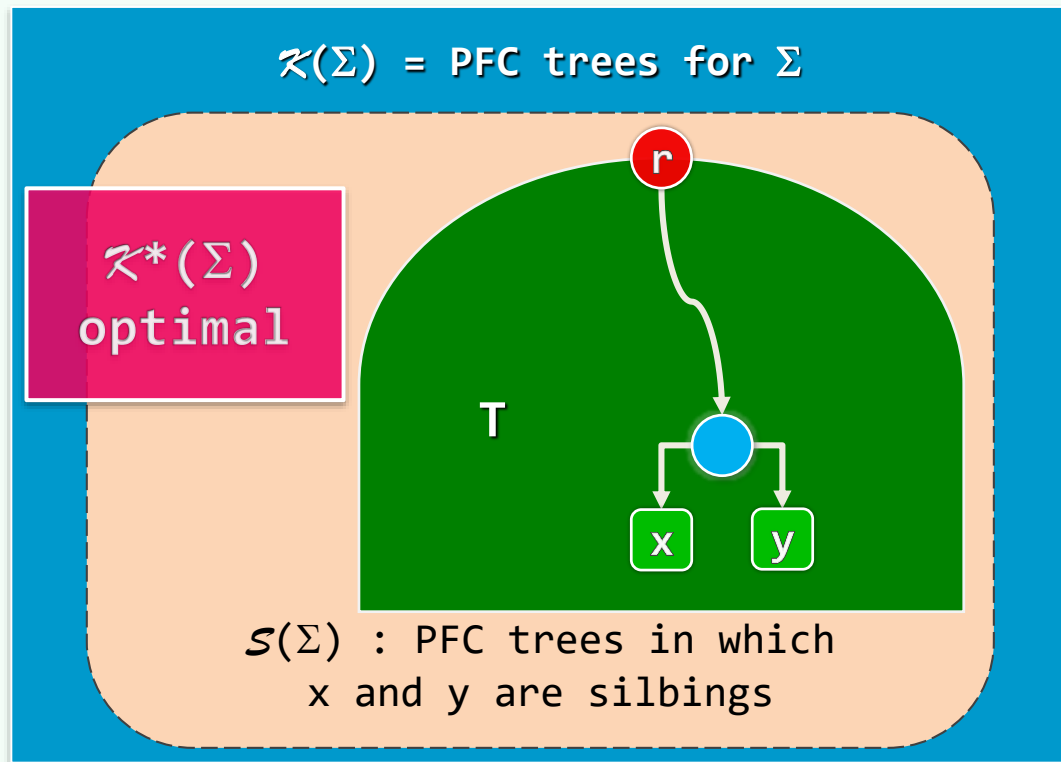
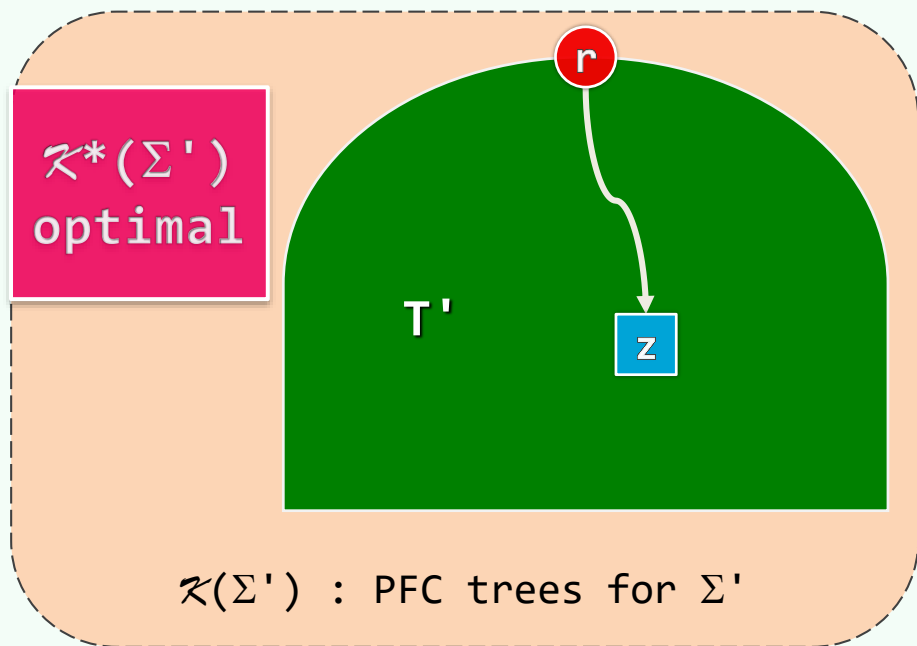


定差

❖ 对于 Σ' 的任一编码树 T' ，只要为 z 添加孩子 x 和 y ，即可得到 Σ 的一棵编码树 T ，且

$$wd(T) - wd(T') = w(x) + w(y) = w(z)$$

❖ 可见，如此对应的 T 和 T' ， wd 之差与 T 的具体形态无关



从最优，到最优

❖ 因此，只要 T' 是 Σ' 的最优编码树，则 T 也必是 Σ 的最优编码树（之一）

❖ 实际上，Huffman算法的过程，与上述归纳过程完全一致

——每一步迭代都可视作，从某棵 T 转入对应的 T'

