5.二叉树

Huffman编码树

正确性

邓俊辉

deng@tsinghua.edu.cn

正确性？

❖ 贪婪策略？

在多数场合并不适用

不见得能得到最优解

甚至反而得到 最差解 //比如，最短路径

❖ Huffman树的构造采用了贪婪策略，它是最优编码树？总是？

❖ 易见：任一指定频率的字符集，都存在对应的最优编码树

❖ 然而，最优编码树可能 不止 一棵

❖ 断言：Huffman树必是其中 之一 //为什么？

❖ 不妨，先来考察最优编码树的特性...

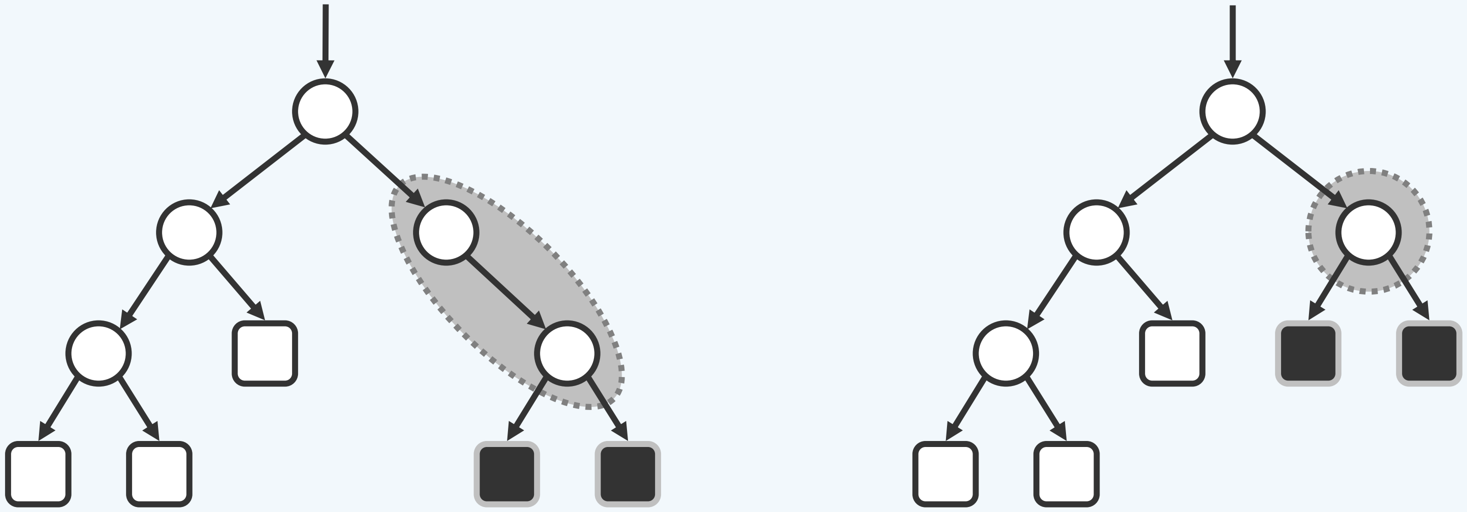
双子性

- ❖ 只要$|\Sigma| > 1$，最优编码树中每一内部节点都有两个孩子，亦即

  节点度数均为偶数（$\boxed{0}$或$\boxed{2}$）

  Huffman树必为$\boxed{\text{真}}$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



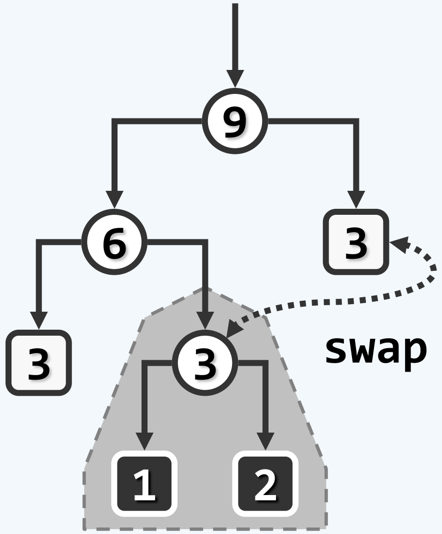
不唯一性

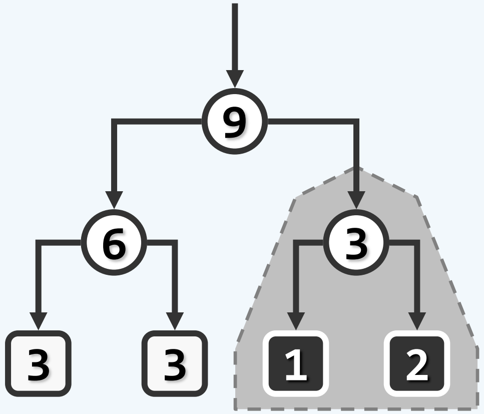
❖ 任一内部节点的左、右子树相互交换之后，wald不变

//上述算法中，左右子树的次序可以随机选取，故此...

❖ 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...



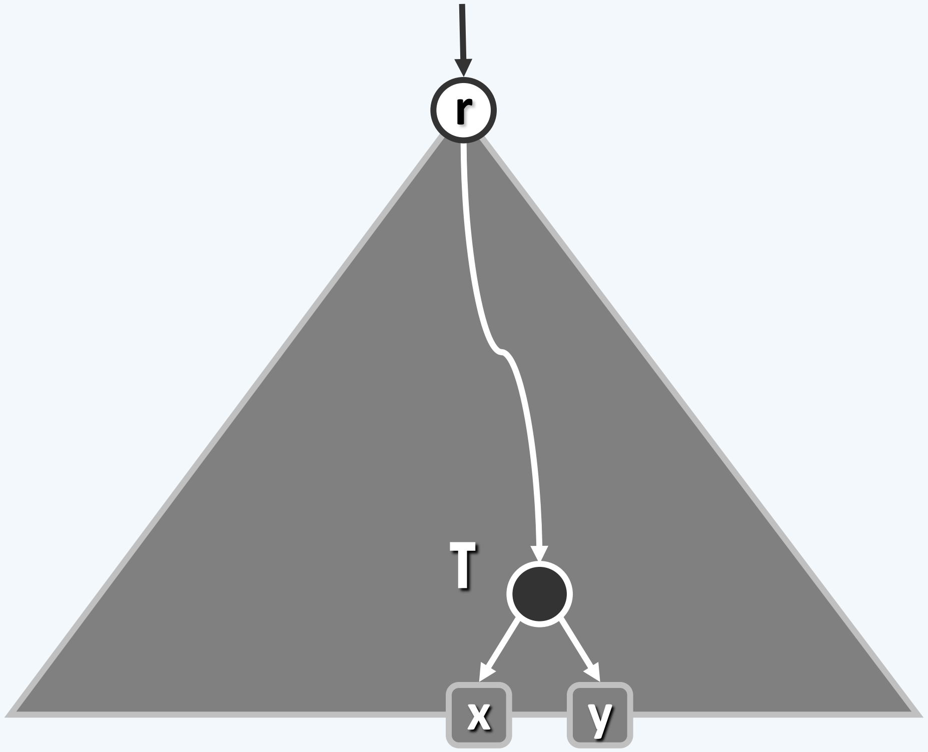


层次性

❖ 若：在字符表中，$\boxed{x}$和$\boxed{y}$是出现频率最低的两个字符

则：存在$\boxed{\text{某棵}}$最优编码树，$\boxed{x}$和$\boxed{y}$在其中处于最$\boxed{\text{底层}}$，且互为$\boxed{\text{兄弟}}$

❖ 为什么？

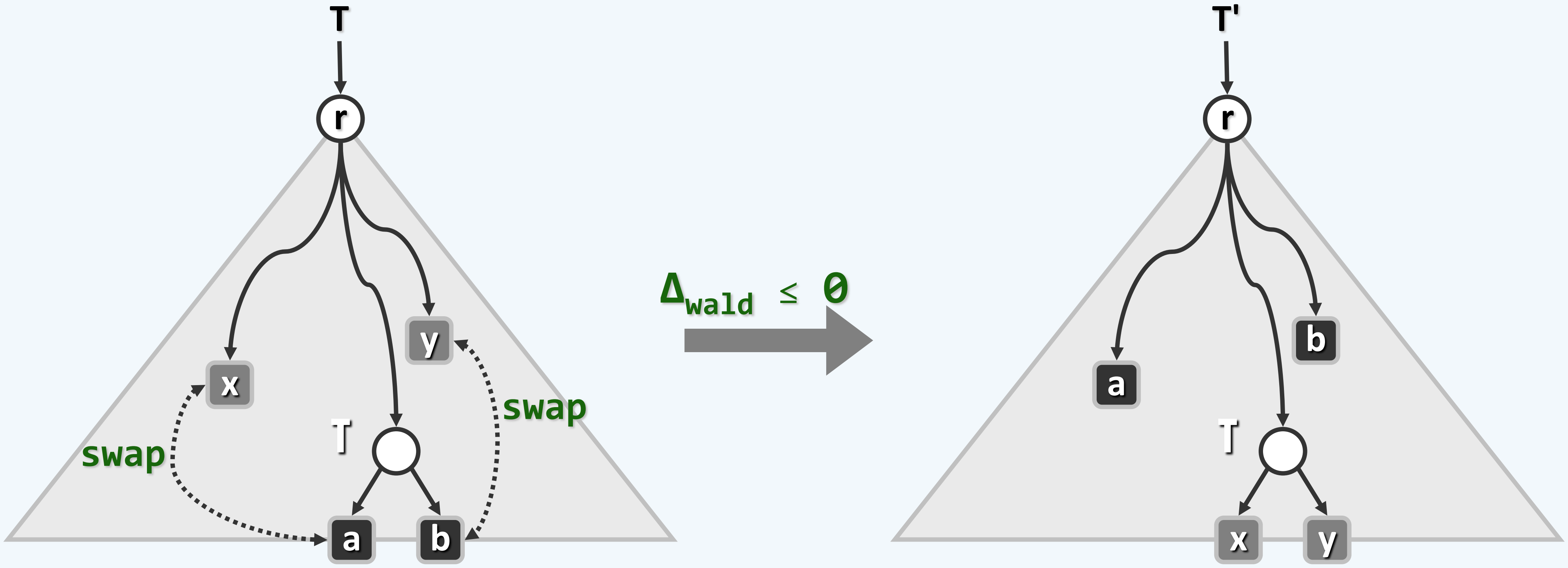


层次性

❖ 任取一棵最优编码树 //注意T的存在性

在其最底层，任取一对兄弟a和b //同样，注意其存在性

交换a和x，交换b和y之后，wald绝不会增加 //正如此前已看到的

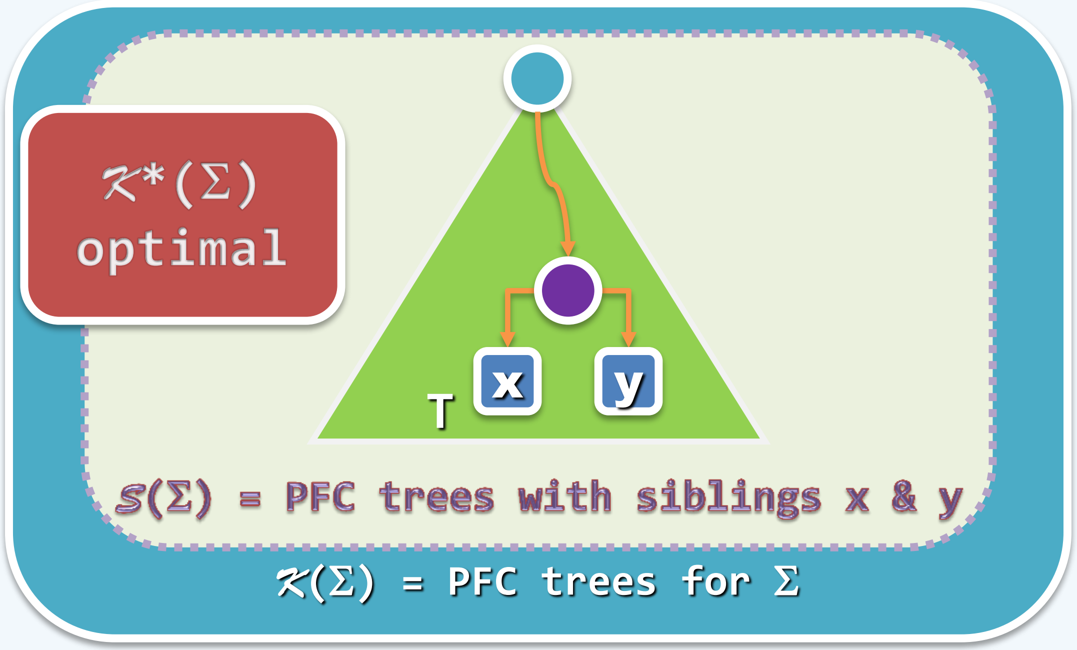


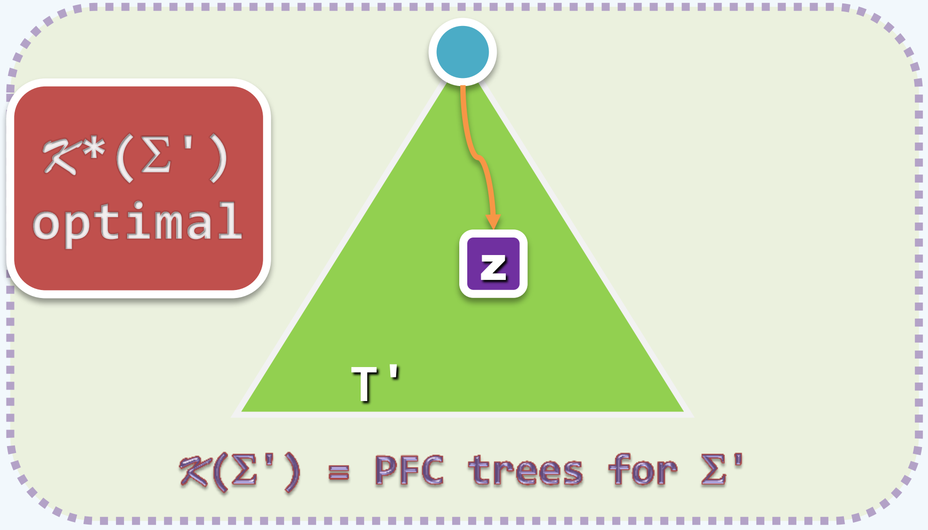
正确性

❖ Huffman（算法所生成的）编码树，的确最优！

❖ 对|Σ|做归纳：|Σ| = 1时显然

设|Σ| < n时Huffman算法都能最优编码，考虑|Σ| = n的情况...

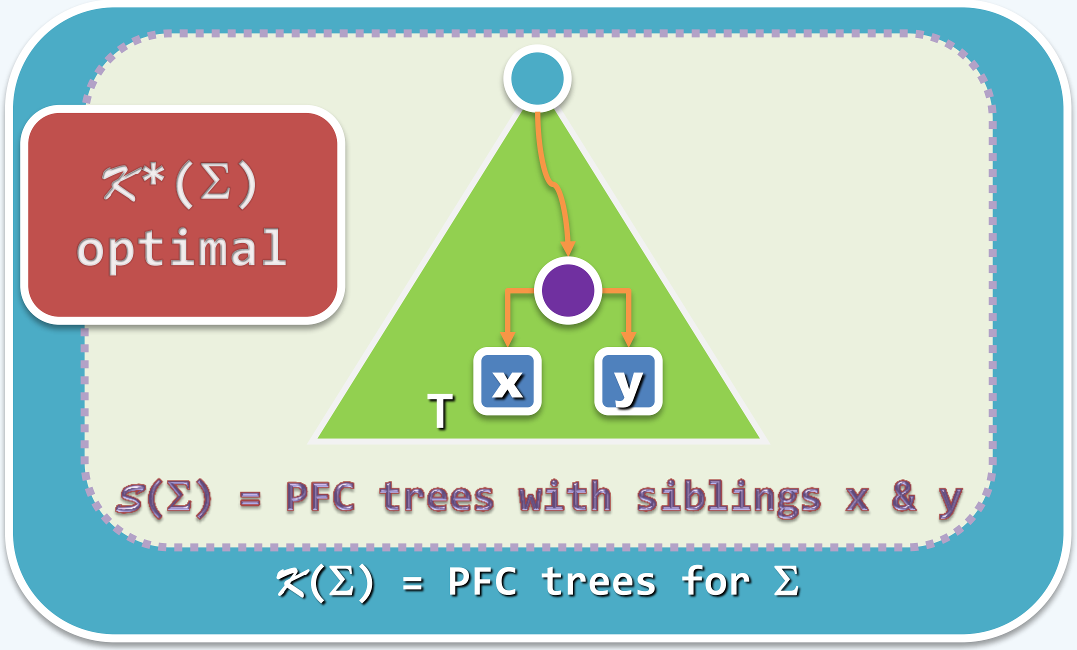


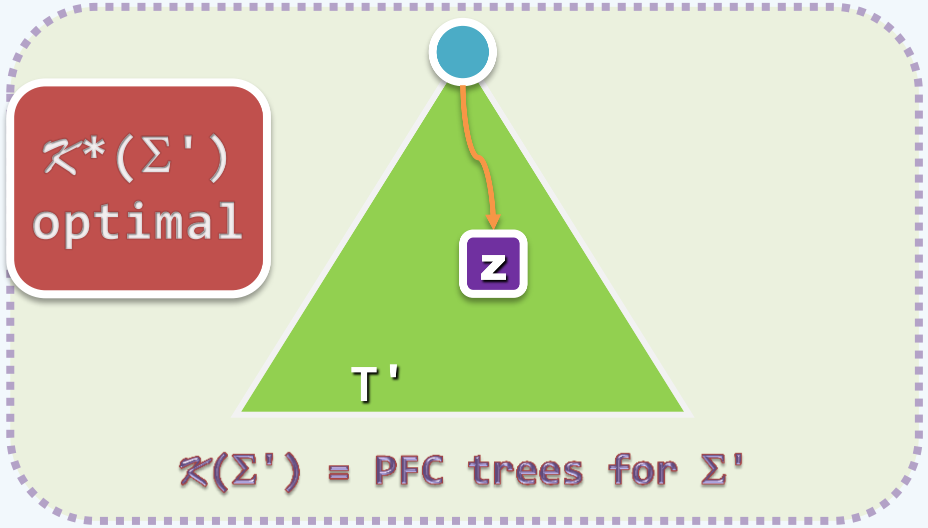


正确性

❖ 取Σ中频率最低的x和y //由层次性，仅考虑其互为兄弟的情形

❖ 令$\Sigma' = (\Sigma \backslash \{x, y\}) \cup \{z\}$，$w(z) = w(x) + w(y)$



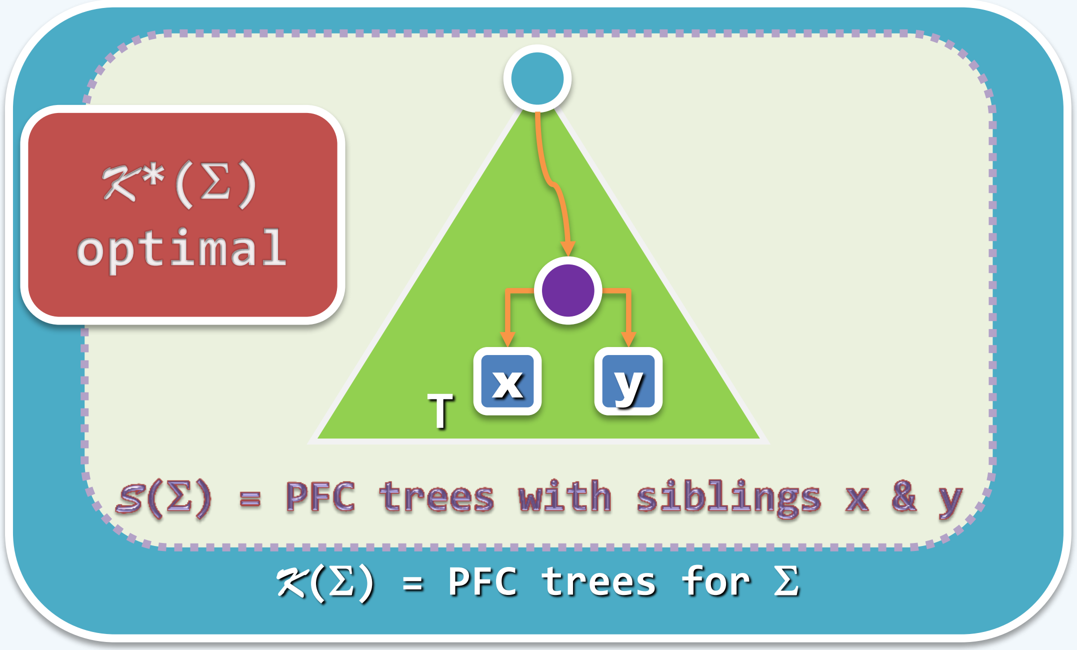


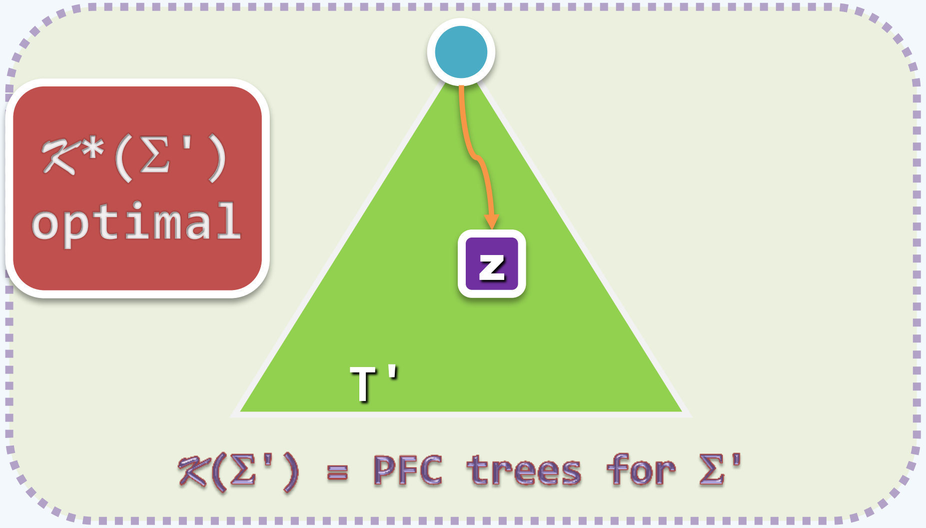
正确性

❖ 对于Σ'的 $\boxed{\text{任一}}$ 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 $\boxed{wd(T) - wd(T')} = \boxed{w(x) + w(y)} = \boxed{w(z)}$

❖ 亦即，如此对应的T和T'，wd之差与 $\boxed{\text{T的具体形态}}$ 无关





正确性

❖ 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