Homework 10

孙锴

June 5, 2012

练习(7.3).
初始时读入a_1，记$a = a_1$，$b = a_1^2$。
接下来每次读入一个数a_i，均将$b + a_i^2$赋值给b，并以$\frac{a_i^2}{b}$的概率将a_i赋值给a。
则对于任何时刻，a都是满足条件的被选择的数。

练习(7.4).
记单词流为$w_1, w_2, \ldots, w_n$。
初始时读入w_1，记$w = w_1$。
接下来每次读入一个单词w_i，均以$\frac{1}{i}$的概率将w_i赋值给w。
则对于任何时刻，w都是满足条件的被选择的单词。

练习(7.5). 记数据流为$c_1, c_2, \ldots, c_n$。
初始时读入c_1，记$a_j = c_1$ $(1 \leq j \leq s)$，$b = c_1$。
接下来每次读入一个数据c_i，将$b + c_i$赋给b，然后对于每个a_j $(1 \leq j \leq s)$，均以$\frac{c_i}{b}$的概率将其赋值改为c_i。
则对于任何时刻，$a_1, a_2, \ldots, a_s$都是已读入数据中抽取出的s个独立的样本。
下面用归纳法证明上述算法的正确性：
显然读完*1*个数据时$a_1, a_2, \ldots, a_s$是s个独立的样本。
下面假设读完i个数据时$a_1, a_2, \ldots, a_s$是s个独立的样本，读第$i+1$个数据后，有$b = c_1 + c_2 + \ldots + c_{i+1}$，则$a_1$有$\frac{c_{i+1}}{b} = \frac{c_{i+1}}{c_1+c_2+\ldots+c_{i+1}}$概率赋值为$c_{i+1}$，有$\frac{c_j}{c_1+c_2+\ldots+c_i}(1 - \frac{c_{i+1}}{c_1+c_2+\ldots+c_i}) = \frac{c_j}{c_1+c_2+\ldots+c_{i+1}}$概率赋值为$(c_j\ 1 \leq j \leq i)$，即$a_1$是$c_1, c_2, \ldots, c_{i+1}$中以数值大小为权重随机抽取的数，同理$a_2, a_3, \ldots, a_s$也是$c_1, c_2, \ldots, c_{i+1}$中以数值大小为权重随机抽取的数。下面只需要说明$a_1, a_2, \ldots, a_s$相互独立。首先由假设，在读完$i$个数据时$a_1, a_2, \ldots, a_s$相互独立，而在读入$c_{i+1}$并按照概率依次对$a_1, a_2, \ldots, a_s$做可能的赋值时，显然任意两个赋值过程是彼此独立无关。因此$a_1, a_2, \ldots, a_s$依旧保持相互独立。

综上，由归纳法证得上述算法是正确的。