How Much Randomness is Needed to Sample from a Discrete Distribution?

Eklavya Sharma

There is a discrete distribution $\mathcal{D}$ of support $S \subseteq \mathbb{Z}$. We want to sample from $\mathcal{D}$, and our only source of randomness is a (potentially unfair) k-sided die ($k \geq 2$), where the probability of each face is known.

For $i \in \{1, \ldots, k\}$, let q_i be the probability that the die shows i. Without loss of generality assume $0 < q_1 \leq q_2 \leq \ldots \leq q_k$. For $i \in \{0, \ldots, k\}$, let $Q_i := \sum_{j=1}^{i} q_j$.

1 Algorithm and Notation

Let Z be a random variable from $\mathcal{D}$. For any $i \in S$, let $p_i := \Pr(Z = i)$ and let I_i be the half-open interval $(\Pr(Z < i), \Pr(Z \leq i)]$.

Lemma 1. $(0, 1) \subseteq \bigcup_{i \in S} I_i \subseteq [0, 1]$.

Proof. (TODO) □

Algorithm 1 `sampleFromDie`: Sample from $\mathcal{D}$, where `throwDie` $\sim \text{unif}(\{1, 2, \ldots, k\})$.

```
Set ℓ = 0 and r = 1.
for t ∈ ℤ_{≥0} do
    if (ℓ, r] ⊆ I_i for some i ∈ S then
        return (i, t)
    end if
    v = throwDie()
    ℓ, r = ℓ + Q_{v−1}(r − ℓ), ℓ + Q_v(r − ℓ)
end for
```

Let (Y, T) `sampleFromDie`'s output. Here Y is the random sample from $\mathcal{D}$ we are looking for and T is the number of times `throwDie` is called.

For $t \in \mathbb{Z}_{\geq 1}$, let v_t be the value of the t^{th} die roll. For $t \in \mathbb{Z}_{\geq 0}$, let ℓ_t and r_t be the value of ℓ and r in `sampleFromDie` after t iterations have completed. Let $J_t := (\ell_t, r_t]$. (The algorithm may terminate before the t^{th} iteration completes. However, we look at the hypothetical scenario where we continue running the algorithm instead of terminating at return statements. Hence, v_t, ℓ_t, and r_t are well-defined for all t.)

Let $\mathcal{J}_t$ be the support of J_t.

2 Correctness

For any interval I, let $\mathrm{len}(I)$ be its length.

Lemma 2. *For all $t \in \mathbb{Z}_{\geq 0}$ and $W_t \in \mathcal{J}_t$, we have $\Pr(J_t = W_t) = \mathrm{len}(W_t)$.*

Proof. (TODO) □

Theorem 3. *Let Y be* `sampleFromDie`*'s output. Then $\Pr(Y = i) = p_i$ for all $i \in S$.*

Proof. (TODO) □

3 Running Time for Bounded $|S|$

Without loss of generality, assume $S = \{1, 2, \ldots, n\}$. For $i \in \{0, 1, \ldots, n\}$, let $s_i := \Pr(Z \leq i)$. For any proposition P, let $:= (P)$ be 1 if P is true and 0 if P is false.

Lemma 4. *For all $t \in \mathbb{Z}_{\geq 0}$ and $W_t \in \mathcal{J}_t$, we have $\mathrm{len}(W_t) \in [q_1^t, q_k^t]$.*

Proof. (TODO) □

Lemma 5. $\Pr(T > t) \leq \min(q_k^t(n-1), 1)$.

Proof. Let $C := \{s_1, \ldots, s_{n-1}\}$. Then $T > t \iff C \cap J_t \neq \emptyset$. For any $W \in \mathcal{J}_t$, let $n_W := |C \cap W|$. Then

$$\begin{aligned}
\Pr(T > t) &= \Pr(C \cap J_t \neq \emptyset) \\
&= \sum_{W \in \mathcal{J}_t} \Pr(J_t = W) := (n_W \geq 1) \\
&\leq \sum_{W \in \mathcal{J}_t} q_k^t n_W \\
&= q_k^t \sum_{W \in \mathcal{J}_t} |C \cap W| \\
&= q_k^t (n-1).
\end{aligned}$$

□

Corollary 5.1. *Let $\delta \in (0,1)$ and $t := \left\lceil \dfrac{\log(n-1) + \log(1/\delta)}{\log(1/q_k)} \right\rceil$. Then $\Pr(T > t) \leq \delta$.*

Theorem 6. *Let $h := \left\lceil \dfrac{\log(n-1)}{\log(1/q_k)} \right\rceil$. Then $\mathrm{E}(T) \leq h + \dfrac{(n-1)q_k^h}{1-q_k} \leq h + \dfrac{1}{1-q_k}$.*

Proof. For any $t \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned}
t \geq h &\iff t \geq \frac{\log(n-1)}{\log(1/q_k)} \iff t \log(1/q_k) \geq \log(n-1) \\
&\iff q_k^{-t} \geq n-1 \iff (n-1)q_k^t \leq 1.
\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{E}(T) &= \sum_{t=0}^{\infty} \Pr(T > t) \le \sum_{t=0}^{\infty} \min(1, q_k^t(n-1)) \\ &\le h + \sum_{t=h}^{\infty} q_k^t(n-1) = h + (n-1)q_k^h \sum_{t=0}^{\infty} q_k^t \\ &= h + \frac{(n-1)q_k^h}{1-q_k} \le h + \frac{1}{1-q_k}.\end{aligned}$$

□

If we have a fair n-sided die, then we get $\mathrm{E}(T) \le 2$.

4 Entropy-Based Bound

All intervals in $\mathcal{J}_t$ are said to have *depth* t.

Definition 1. *Let* $\mathcal{J} := \bigcup_{t=0}^{\infty} J_t$. *An interval* $W \in \mathcal{J}$ *is said to be* maximal in I_i *if* $W \subseteq I_i$ *and for all* $W' \in \mathcal{J}$, *we have* $W' \supsetneq W \implies W' \nsubseteq I_i$.

A die-decomposition *of* I_i *is the set of all sets in* $\mathcal{J}$ *that are maximal in* I_i.

Lemma 7. *For any* $i \in S$, *let* $\mathcal{K}_i$ *be a die-decomposition of* I_i. *Then* $\bigcup_{W \in \mathcal{K}_i} W = I_i$.

Proof. (TODO) □

Lemma 8. *Let* (Y, T) *be the output of* `sampleFromDie`. *Then* $(\ell_T, r_T) \in \mathcal{K}_Y$.

Proof. (TODO) □

Lemma 9. *For any* $i \in S$ *and* $t \in \mathbb{Z}_{\ge 0}$, *we get*

$$\sum_{W \in \mathcal{K}_i \cap \mathcal{J}_t} \mathrm{len}(W) \le (2 - q_1 - q_k) q_k^{t-1}.$$

Proof. (TODO) □

Lemma 10.

$$\mathrm{E}(T) \le \frac{2 - q_1 - q_k}{1 - q_k} \sum_{i \in S} p_i^{\frac{\log(1/q_k)}{\log(1/q_1)}} \left(\left\lceil \frac{\log(1/p_i)}{\log(1/q_1)} \right\rceil + \frac{q_k}{1 - q_k} \right).$$

Proof. Let $\beta_{i,t} := \sum_{W \in \mathcal{K}_i \cap \mathcal{J}_t} \mathrm{len}(W)$. Then $\beta_{i,t} \le \min(p_i, (2 - q_1 - q_k) q_k^{t-1})$. $\beta_{i,t} > 0 \iff |\mathcal{K}_i \cap \mathcal{J}_t| > 0 \implies q_1^t \le p_i$. Let $h_i := \left\lceil \frac{\log(1/p_i)}{\log(1/q_1)} \right\rceil$. Then

$$q_1^t \le p_i \iff t \log(1/q_1) \ge \log(1/p_i) \iff t \ge h_i.$$

Hence, $t < h_i \implies \beta_{i,t} = 0$.

$$
\begin{aligned}
\mathrm{E}(T) &= \sum_{i\in S}\sum_{t=0}^{\infty} \mathrm{E}(T|J_T \in \mathcal{K}_i \cap \mathcal{J}_t)\Pr(J_T \in \mathcal{K}_i \cap \mathcal{J}_t) \\
&= \sum_{i\in S}\sum_{t=0}^{\infty} t\Pr(J_T \in \mathcal{K}_i \cap \mathcal{J}_t) \\
&= \sum_{i\in S}\sum_{t=0}^{\infty} t\beta_{i,t} = \sum_{i\in S}\sum_{t=h_i}^{\infty} t\beta_{i,t} \\
&\le (2-q_1-q_k)\sum_{i\in S}\sum_{t=h_i}^{\infty} t q_k^{t-1} \\
&= \frac{2-q_1-q_k}{1-q_k}\sum_{i\in S} q_k^{h_i-1}\left(h_i + \frac{q_k}{1-q_k}\right) \\
&= \frac{2-q_1-q_k}{1-q_k}\sum_{i\in S} p_i^{\frac{\log(1/q_k)}{\log(1/q_1)}}\left(h_i + \frac{q_k}{1-q_k}\right)
\end{aligned}
$$

□

Suppose $q_1 = q_k = 1/k$. Then we get

$$
\begin{aligned}
\mathrm{E}(T) &\le \frac{2(1-1/k)}{1-1/k}\sum_{i\in S} p_i\left(\lceil \log_k(1/p_i)\rceil + \frac{k}{k-1}\right) \\
&= 2\sum_{i\in S} p_i \lceil \log_k(1/p_i)\rceil + \frac{2k}{k-1}.
\end{aligned}
$$