Part III

Stochastic Algorithms

8 Simulated Annealing

Global optimisation (say, minimisation) of an objective function $H(\sigma)$, framed as a Hamiltonian of a statistical mechanics system, via a sequence of Metropolis samplers for the Gibbs distributions determined by $H(\sigma)$ at decreasing values of the temperature parameter $T \to 0$.

Let $H : S \to \mathbb{R}$ be a function to be minimised over a finite (but typically very large) state space S. Assume that S has some neighbourhood structure $S = (S, N)$ (cf. page 24).

In any specific application of the method, the algorithm designer typically has a lot of freedom in the choice of the most appropriate N. This choice can have a significant effect on the efficiency of the algorithm: one would like to have N such that $N(\sigma)$ is small for each $\sigma \in S$, yet the resulting Metropolis chains converge rapidly.

The Gibbs distribution determined by H at temperature T is (recall page 58):

$$\pi_\sigma^{(T)} = \Pr_T(\sigma) = \frac{1}{Z_T} e^{-H(\sigma)/kT} = \frac{1}{Z_T} e^{-\beta H(\sigma)},$$

where $\beta = 1/kT$.

A relevant observation is that as $T \to 0$ (or $\beta \to \infty$), the distribution $\Pr_T(\sigma)$ gets more peaked according to H. Denoting by $S^* = \{\sigma^* \in S \mid H(\sigma^*) = \min\}$ the set of global optima of H, one observes that:

$$\frac{\Pr_T(\sigma)}{\Pr_T(\sigma^*)} = e^{-\beta(H(\sigma)-H(\sigma^*))} \xrightarrow[(\beta\to\infty)]{T\to 0} \begin{cases} 0, & \sigma \notin S^* \\ 1, & \sigma \in S^* \end{cases}$$

Thus, in the limit one obtains:

$$\pi^*_\sigma = \lim_{T\to 0} \Pr_T(\sigma) = \begin{cases} 0, & \sigma \notin S^* \\ 1/|S^*|, & \sigma \in S^* \end{cases}$$

Of course, one cannot directly sample according to π^*, but the idea is that by starting at a high value of T and then slowly (but how slowly?) decreasing it, one obtains a nonhomogenous Metropolis chain that converges reasonably fast (?) to π^*.

As regards the convergence of the chains at each fixed $T > 0$, we can appeal to the general results concerning Metropolis samplers from page 24 onwards.

Let us just check the form of the acceptance probabilities: a proposed move $\sigma \to \tau$, where $\tau \in N(\sigma)$, is accepted with probability:

$$\begin{aligned} p_{\sigma\tau} &= \min\left\{\frac{\pi_\tau d_\sigma}{\pi_\sigma d_\tau}, 1\right\} \\ &= \min\left\{\frac{e^{-\beta H(\tau)}}{e^{-\beta H(\sigma)}} \cdot \frac{d_\sigma}{d_\tau}, 1\right\} \\ &= \min\left\{e^{-\beta(H(\tau)-H(\sigma))} \cdot \frac{d_\sigma}{d_\tau}, 1\right\} \\ &= \min\left\{e^{-\beta(H(\tau)-H(\sigma))}, 1\right\}, \end{aligned}$$

if (S,N) is regular i.e. $|N(\sigma)| = |N(\tau)|$ for all σ, τ.

Thus, for a regular neighbourhood structure, and denoting $\Delta H = H(\tau) - H(\sigma)$, a proposed transition $\sigma \to \tau$ is accepted always if $\Delta H \leq 0$, and with probability $e^{-\beta \Delta H}$, if $\Delta H > 0$.

In summary, one obtains the following general method for minimising a function H over a state space S with neighbourhood structure N:

Algorithm SA(H, S, N):

$T \leftarrow T_{init}$;
$\sigma \leftarrow \sigma_{init}$;
while $T > T_{final}$ do

    $L \leftarrow$ sweep(T);
    for L times do

        choose $\tau \in N(\sigma)$ uniformly at random;
        $\Delta H \leftarrow H(\tau) - H(\sigma)$;

if $\Delta H \leq 0$ then $\sigma \leftarrow \tau$;
else choose $r \in [0,1)$ uniformly at random;
if $r \leq \exp(-\Delta H/T)$
then $\sigma \leftarrow \tau$;

end for;
$T \leftarrow \text{lower}(T)$;

end while;
result $\leftarrow \sigma$;

(For simplicity, the neighbourhood graph is assumed to be regular.)

The obvious question is now how to choose appropriate functions lower(T) and sweep(T), i.e. what is a good "cooling schedule" $\langle T_0, L_0\rangle, \langle T_1, L_1\rangle, \ldots$

In practice, it is customary to just start from some "high" temperature T_0, and after each "sufficiently long" sweep L decrease the temperature by some "cooling factor" $\alpha \approx 0.8 \ldots 0.99$:

$$T_{k+1} = \alpha T_k.$$

Theoretically this is much too fast, as we shall see, but often seems to work well enough.

Consider an inhomogenous Markov chain with transition matrices $P^{(0)}, P^{(1)}, P^{(2)}, \ldots$ Denote

$$P(m,k) = P^{(m)} P^{(m+1)} \cdots P^{(m+k-1)}$$

i.e. $P_{ij}(m,k) = \Pr(X_{m+k} = j \mid X_m = i)$.

The chain $\mathcal{M}$ is *weakly ergodic* if for all $m \geq 0$:

$$\lim_{k\to\infty} \sup_{\mu,\nu} d_V\left(\mu^T P(m,k), \nu^T P(m,k)\right) = 0$$

and *strongly ergodic* if there is some distribution π such that for all $m \geq 0$:

$$\lim_{k\to\infty} \sup_{\mu} d_V(\mu^T P(m,k), \pi) = 0$$

Let Q be an $n \times m$ stochastic matrix. The *(Dobrushin) ergodic coefficient* of Q is defined as:

$$\begin{aligned} \rho = \rho(Q) &= \max_{i,j} d_V(q_i, q_j) \qquad \begin{matrix} q_i = (q_{i1}, \ldots, q_{im}) \\ q_j = (q_{j1}, \ldots, q_{jm}) \end{matrix} \\ &= \frac{1}{2} \max_{i,j} \sum_{k=1}^{m} |q_{ik} - q_{jk}| \end{aligned}$$

The following key technical lemmas will possibly be proved later. The proofs are not exceedingly difficult.

Lemma 8.1 ("Dobrushin's inequality")
Given the stochastic matrices $Q_1 \in [0,1]^{n\times m}, Q_2 \in [0,1]^{n\times l}$:

$$\rho(Q_1 Q_2) \leq \rho(Q_1)\rho(Q_2).$$

Lemma 8.2 ("Dobrushin convergence rate bound")
Given the stochastic matrix P and the distributions μ, ν:

$$d_V(\mu^T P^n, \nu^T P^n) \leq d_V(\mu,\nu)\rho(P)^n.$$

Lemma 8.3
An inhomogeneous Markov chain $\mathcal{M}$ with transition probability matrices $P^{(0)}$, $P^{(1)}, \ldots$ is weakly ergodic if and only if either (and hence both) of the following conditions hold:

(i) *for any $m \geq 0$: $\lim_{k\to\infty} \rho(P(m,k)) = 0$;*

(ii) *for some increasing sequence $0 \leq m_0 < m_1 < \cdots$*

$$\sum_{i=0}^{\infty} (1 - \rho(P(m_i, m_{i+1}))) = \infty.$$

Lemma 8.4
Let $\mathcal{M}$ be a weakly ergodic Markov chain with transition probability matrices $P^{(0)}, P^{(1)}, \ldots$ Suppose that there exists a sequence of distributions $\pi^{(0)}, \pi^{(1)}, \ldots$ such that

(i) *$\pi^{(m)} P^{(m)} = \pi^{(m)}$, for each $m \geq 0$;*

(ii) $\sum_{m=0}^{\infty} ||\pi^{(m)} - \pi^{(m+1)}||_1 < \infty.$

Then $\mathcal{M}$ is also strongly ergodic, with limit distribution

$$\pi_i^* = \lim_{m\to\infty} \pi_i^{(m)}.$$

Theorem 8.5
Consider a simulated annealing computation on input (H,S,N). *Assume the neighbourhood graph* (S,N) *is connected and regular of degree r. Denote:*

$$\Delta = \max\{H(\tau) - H(\sigma) \mid \sigma \in S, \tau \in N(\sigma)\}.$$

Suppose the cooling schedule used is of the form $\langle T_0, L\rangle, \langle T_1, L\rangle, \langle T_2, L\rangle, \ldots$, *where*

$$L \geq \min_{\sigma^* \in S^*} \max_{\sigma \notin S^*} \mathrm{dist}(\sigma, \sigma^*), \tag{1}$$

where $\mathrm{dist}(\sigma,\sigma^*)$ *is the distance in graph* (S,N) *from* σ *to* σ^*, *and for each cooling stage* $l \geq 2$*:*

$$T_l \geq \frac{L\Delta}{\ln l} \quad (\textit{but } T_l \xrightarrow[l\to\infty]{} 0). \tag{2}$$

Then the distribution of states visited by the computation converges in the limit to π^*, *where*

$$\pi^*_\sigma = \begin{cases} 0, & \text{if } \sigma \notin S^* \\ 1/|S^*|, & \text{if } \sigma \in S^* \end{cases}$$

Proof: Denote by $P^{(0)}, P^{(1)}, \ldots$ the sequence of transition matrices for the Markov chain on S determined by the SA algorithm with the given parameters. We shall show, based on Lemma 8.4, that this chain is strongly ergodic with the given limit distribution.

Let us first verify weak ergodicity using Lemma 8.3 (ii). Let $\sigma^* \in S^*$ be some ground state achieving the lower bound in condition (1). We shall show that for any $\sigma \in S$ and $k \geq k_0$, where k_0 is sufficiently large:

$$P_{\sigma\sigma^*}(k, k+L) \geq \left(\frac{1}{r} e^{-\Delta/t_k}\right)^L, \tag{3}$$

where $t_k = T_{\lfloor k/L \rfloor}$ = cooling temperature at step k.

It then follows from condition (3) and from the fact $|p-q| = p+q-2\min\{p.q\}$ that

$$\begin{aligned}
& 1 - \rho(P(k,k+L)) \\
&= 1 - \frac{1}{2} \max_{\sigma,\tau} \sum_{\nu \in S} |P_{\sigma\nu}(k,k+L) - P_{\tau\nu}(k,k+L)| \\
&= \min_{\sigma,\tau} \sum_{\nu \in S} \min\{P_{\sigma\nu}(k,k+L), P_{\tau\nu}(k,k+L)\} \\
&\geq \min_{\sigma \in S} P_{\sigma\sigma^*}(k,k+L) \\
&\geq r^{-L} e^{-L\Delta/t_k},
\end{aligned}$$

and so (choosing $m_l = l \cdot L$):

$$\sum_{l=0}^{\infty}(1-\rho(P(m_l,m_{l+1}))) \geq \sum_{l=k_0}^{\infty}(1-\rho(P(lL,lL+L)))$$
$$\geq \sum_{l=k_0}^{\infty} r^{-L}e^{-L\Delta/t_k} \geq r^{-L}\sum_{l=k_0}^{\infty}\frac{1}{l} = \infty.$$

Thus, let us check that condition (3) holds for some sufficiently large k_0. Observe first that for any $\sigma \in S$ and $\tau \in N(\sigma)$:

$$P_{\sigma\tau}(k) = \frac{1}{r}\min\{e^{-(H(\tau)-H(\sigma))/t_k}, 1\} \geq \frac{1}{r}e^{-\Delta/t_k}.$$

Similarly, for any $\sigma^* \in S^*$ there is some k_0 such that for all $k \geq k_0$:

$$P_{\sigma^*\sigma^*}(k) \geq \frac{1}{r}e^{-\Delta/t_k}.$$

Namely, let $\delta = \min\{H(\tau)-H(\sigma^*) \mid \sigma^* \in S^*, \tau \in N(\sigma^*) \setminus S^*\}$. Now $\delta > 0$, unless H is a constant function. Thus for all $k \geq k_0$, where k_0 is sufficiently large:

$$1 - e^{-\delta/t_k} \geq e^{-\Delta/t_k},$$

and so

$$\begin{aligned}
P_{\sigma^*\sigma^*} &= 1 - \sum_{\tau\in N(\sigma^*)} P_{\sigma^*\tau}(k) \\
&= 1 - \sum_{\tau\in N(\sigma^*)} \frac{1}{r}e^{-(H(\tau)-H(\sigma))/t_k} \\
&\geq 1 - \frac{1}{r}(r-1+e^{-\delta/t_k}) \\
&= \frac{1}{r}(1-e^{-\delta/t_k}) \\
&\geq \frac{1}{r}e^{-\Delta/t_k}.
\end{aligned}$$

Thus, for any $\sigma \in S$ and $k \geq k_0$:

$$\begin{aligned}
&P_{\sigma\sigma^*}(k,k+L) \\
&= \sum_{\tau_1}\sum_{\tau_2}\cdots\sum_{\tau_{L-1}} P_{\sigma\tau_1(k)}P_{\tau_1\tau_2}(k+1)\cdots P_{\tau_{L-1}\sigma^*}(k+L-1) \\
&\geq P_{\sigma\sigma_1}(k)P_{\sigma_1\sigma_2}(k+1)\cdots P_{\sigma_{L-1}\sigma^*}(k+L) \\
&\geq \left(\frac{1}{r}e^{-\Delta/t_k}\right)^L,
\end{aligned}$$

where $\sigma, \sigma_1, \sigma_2, \ldots, \sigma_{L-1}, \sigma^*$ is a shortest path from σ to σ^* in (S, N), with possibly state σ^* repeated several times if the length of the actual path is less than L.

Having now established the weak ergodicity of our chain, let us check conditions (i) and (ii) of Lemma 8.4 to complete the proof.

For condition (i) it suffices to observe that the stationary distribution at stage l of the algorithm:

$$\pi_\sigma^{(l)} = \frac{1}{Z_l} e^{-H(\sigma)/T_l}, \quad Z_l = \sum_{\sigma \in S} e^{-H(\sigma)/T_l},$$

satisfies the condition $\pi^{(l)} P^{(m)} = \pi^{(l)}$, for values of m from lL to $(l+1)L - 1$.

For condition (ii), one can show by a somewhat tedious calculation (cf. Aarts & Korst, "Simulated Annealing ... ", p. 22) that for each of the intermediate stationary distributions $\pi^{(l)}$:

$$\text{if } \sigma^* \in S^*, \text{ then } \frac{\partial}{\partial T} \pi_{\sigma^*}^{(l)} < 0;$$
$$\text{if } \sigma \notin S^*, \text{ then } \frac{\partial}{\partial T} \pi_\sigma^{(l)} > 0 \text{ for } l \geq l_1 \text{ sufficiently large.}$$

As $T_{l+1} \leq T_l$ at each stage l, it thus follows that:

$$\pi_{\sigma^*}^{(l+1)} \geq \pi_{\sigma^*}^{(l)} \text{ for } \sigma^* \in S^*$$
$$\pi_\sigma^{(l+1)} \leq \pi_\sigma^{(l)} \text{ for } \sigma \notin S^* \text{ and } l \geq l_1$$

Thus, for $l \geq l_1$:

$$\begin{aligned} \left|\left| \pi^{(l)} - \pi^{(l+1)} \right|\right|_1 &= \sum_{\sigma \in S} \left| \pi_\sigma^{(l)} - \pi_\sigma^{(l+1)} \right| \\ &= \sum_{\sigma^* \in S^*} \left| \pi_{\sigma^*}^{(l)} - \pi_{\sigma^*}^{(l+1)} \right| + \sum_{\sigma \notin S^*} \left| \pi_\sigma^{(l)} - \pi_\sigma^{(l+1)} \right| \\ &= 2 \left(\sum_{\sigma^* \in S^*} \pi_{\sigma^*}^{(l+1)} - \sum_{\sigma^* \in S^*} \pi_{\sigma^*}^{(l)} \right). \end{aligned}$$

Hence, denoting $\hat{\pi}^{(m)} = \pi^{(\lfloor m/L \rfloor)}$:

$$\begin{aligned}
\sum_{m=0}^{\infty} \left|\left| \hat{\pi}^{(m)} - \hat{\pi}^{(m+1)} \right|\right|_1 &= \sum_{l=0}^{\infty} \left|\left| \hat{\pi}^{(l)} - \hat{\pi}^{(l+1)} \right|\right|_1 \\
&= \sum_{l=0}^{l_1} \left|\left| \hat{\pi}^{(l)} - \hat{\pi}^{(l+1)} \right|\right|_1 + \sum_{l=l_1+1}^{\infty} \left|\left| \hat{\pi}^{(l)} - \hat{\pi}^{(l+1)} \right|\right|_1 \\
&\leq 2l_1 + 2 \left(\sum_{\sigma^* \in S^*} \pi^*_{\sigma^*} - \sum_{\sigma^* \in S^*} \pi^{(l_1+1)}_{\sigma^*} \right) \\
&\leq 2l_1 + 2 < \infty.
\end{aligned}$$

This completes the proof, because according to Lemma 8.4 the chain has the limit distribution π^*, where

$$\pi^*_\sigma = \lim_{l\to\infty} \pi^{(l)}_\sigma = \lim_{l\to\infty} \frac{1}{Z_l} e^{-H(\sigma)/T_l} = \begin{cases} 0, & \text{if } \sigma \notin S^* \\ 1/|S^*|, & \text{if } \sigma \in S^* \end{cases} \quad \square$$

9 Approximate counting

Let Σ be an alphabet (without loss of generality $\Sigma = \{0,1\}$) and $R \subseteq \Sigma^* \times \Sigma^*$ an NP relation over Σ^*, i.e.

- for some polynomial $p(n)$, $R(x,w) \;\Rightarrow\; |w| \leq p(|x|)$, where $|z|$ denotes the length of string z
- the condition $R(x,w)$ can be tested in polynomial time, for any given $\langle x,w \rangle$

Well-known examples of NP relations:

- SAT(ϕ,t), where ϕ is (an encoding of) a Boolean formula and $t : \mathrm{Var}_\phi \to \{T,F\}$ is a truth assignment to its variables; relation holds if ϕ evaluates to T under t.
- $\mathrm{COL}_q(G,\sigma)$, where $G = (V,E)$ is a graph and $\sigma : V \to \{1,\ldots,q\}$ is a candidate q-colouring of its nodes; relation holds if σ is valid for G, i.e. if $(u,v) \in E \;\Rightarrow\; \sigma(u) \neq \sigma(v)\ \forall\, u,v \in V$.

Denote $R(x) = \{w \in \Sigma^* | R(x,w) \text{ holds}\}$.

One may consider different computational problems related to R:

- *existence problem*: given x, determine if $R(x) \neq \varnothing$
- *counting problem*: given x, determine $N_R(x) = |R(x)|$
- *sampling problem*: given x, provide $w \in R(x)$ uniformly at random

A *randomised approximation scheme (ras)* for the counting problem associated to R is a randomised algorithm $A(x,\varepsilon)$ such that for any $x \in \Sigma^*, \varepsilon > 0$:

$$\Pr((1-\varepsilon)N_R(x) \leq A(x,\varepsilon) \leq (1+\varepsilon)N_R(x)) \geq \frac{3}{4},$$

where the probability is with respect to the random choices made by the algorithm. The ras is *fully polynomial (fpras)* if its running time is polynomial in $|x|$ and $1/\varepsilon$.

An *almost uniform sampler (aus)* for R is a randomised algorithm $S(x,\delta)$ such that for any $x \in \Sigma^*, S(x,\delta) \in R(x)$ and $d_V(S(x,\delta), U_R(x)) \leq \delta$, where $S(x,\delta)$ denotes (by slight abuse of notation) the distribution of the output of $S(x,\delta)$, and $U_R(x)$ denotes the uniform distribution over $R(x)$. An aus is *fully polynomial (fpaus)* is its running time is polynomial in $|x|$ and $\ln 1/\delta$.

It can be shown (Jerrum et al. 1986, Sinclair 1993) that if R is "self-reducible", then R has an fpras if and only if it has an fpaus.

Self-reducibility of R means roughly (the exact definition is somewhat more general) that there is a small collection of polynomial time functions $f_i, g_i, i = 1, \ldots, k$, such that for any $x \in \Sigma^*$, $|f_i(x)| < |x|$ and

$$R(x) = \bigcup_{i=1}^{k} g_i(x, R(f_i(x))).$$

E.g. for the SAT relation SAT(ϕ) = SAT(ϕ_T) $\cup$ SAT(ϕ_F), where ϕ_T (ϕ_F) is the formula obtained from ϕ by substituting T (F) for the first variable and simplifying. Almost all "natural" NP-complete relations are self-reducible.

Let us see concretely, in the case of low-degree graph colouring, how an efficient fpaus (pages 46-50) can be converted into an efficient fpras.

Given a graph $G = (V,E)$ with maximum node degree $\Delta < q$, denote for brevity $\Omega(G) = \mathrm{COL}_q(G)$, and assume the existence of a fpaus $S(G,\delta)$ for q-colourings. (Actually, the fpaus-construction on pages 46-50 requires more strongly that $\Delta < q/2$.)

One possible self-reduction for graph colouring is

$$\Omega(G) = g(G, \Omega(G')),$$

where $G' \sim G$ with one edge (e.g. highest-numbered one) removed, and

$$g(G,\sigma) = \begin{cases} \sigma & \text{if } \sigma \text{ is valid for } G \\ \perp & \text{otherwise} \end{cases}$$

where $\perp$ is a "null-value" ($S \cup \{\perp\} = S$ for any S).

Assuming $|E| = m$, denote $G = G_m$, $G' = G_{m-1}, \ldots, G^{(m)} = G_0 = (V, \varnothing)$. Now clearly $|\Omega(G_0)| = q^n$, where $n = |V|$. Then the quantity we are interested in can be expressed as:

$$\begin{aligned} N(G) \;&=\; |\Omega(G)| = \frac{|\Omega(G)_m|}{|\Omega(G)_{m-1}|} \cdot \frac{|\Omega(G)_{m-1}|}{|\Omega(G)_{m-2}|} \cdots \frac{|\Omega(G)_1|}{|\Omega(G)_0|} \cdot |\Omega(G)_0| \\ &= \rho_m \cdot \rho_{m-1} \cdots \rho_1 \cdot q^n, \end{aligned} \tag{4}$$

where

$$\rho_k = \frac{|\Omega(G)_k|}{|\Omega(G)_{k-1}|}.$$

Now each of the ratios in ρ_k and hence the product (4) can be estimated using our presumed fpaus to generate a "sufficiently large" number of samples form each $\Omega(G_{k-1})$ and seeing how many of those fall also in $\Omega(G_k)$. Some analysis is needed to determine the appropriate numbers.

Before going into the analysis, let us note that the same approach, combined with more complicated samplers, has been used to provide fpras for such important problems as:

- approximating the volume of a convex body (Dyer, Frieze, Kannan 1991)
- approximating the partition function of a ferromagnetic Ising model (Jerrum & Sinclair 1993)
- approximating the permanent of a positive matrix (Jerrum, Sinclair & Vigoda 2001)

Let us then complete the analysis of the graph colouring fpras. Recall that

$$|\Omega(G)| = \rho_m \cdot \rho_{m-1} \cdots \rho_1 \cdot q^n,$$

where each

$$\rho_k = \frac{|\Omega(G)_k|}{|\Omega(G)_{k-1}|}.$$

Now clearly each $\Omega(G_k) \subseteq \Omega(G_{k-1})$, so that $\rho_k \leq 1$. On the other hand, each colouring $\sigma \in \Omega(G_{k-1}) \setminus \Omega(G_k)$ must be such that it assigns the same colour to both endpoints u, v of the edge e removed from G_k to obtain G_{k-1}. Let u be the lower-numbered of the nodes. Then σ can be transformed to a valid colouring of G_k by recolouring u with one of the $\geq q - \Delta \geq 1$ colours free for it. On the other hand, each colouring in $\Omega(G_k)$ is generated by this process in at most one way. Thus

$$|\Omega(G_{k-1}) \setminus \Omega(G_k)| \leq |\Omega(G_k)|,$$

and so $\rho_k \geq \frac{1}{2}$.

Assume then without loss of generality that $m \geq 1$ and $0 < \varepsilon \leq 1$. (Recall $\varepsilon \sim$ error tolerance for the fpras to be constructed).

Let $Z_k \in \{0,1\}$ be a random variable obtained by running the presumed fpaus for G_{k-1} and testing whether the resulting colouring is also valid for G_k ($\rightarrow Z_k = 1$) or not ($\rightarrow Z_k = 0$). Denote $\mu_k = E[Z_k]$.

By setting $\delta = \frac{\varepsilon}{6m}$ in the fpaus one may ensure that

$$\rho_k - \frac{\varepsilon}{6m} \leq \mu_k \leq \rho_k + \frac{\varepsilon}{6m}, \tag{5}$$

and noting the bounds on ρ_k, that

$$\left(1 - \frac{\varepsilon}{3m}\right)\rho_k \leq \mu_k \leq \left(1 + \frac{\varepsilon}{3m}\right)\rho_k. \tag{6}$$

Note also that by (5), $\mu_k \geq \frac{1}{3}$.

To decrease the variance of our ρ_k-estimate, let $Z_k^{(1)}, \ldots, Z_k^{(s)}$ be $s = \lceil 74\varepsilon^{-2}m \rceil \leq 75\varepsilon^{-2}m$ independent copies of variable Z_k, and let

$$\bar{Z}_k = \frac{1}{s}\sum_{i=1}^{s} Z_k^{(i)}$$

be their mean. Then $E[\bar{Z}_k] = E[Z_k] = \mu_k$ and

$$\frac{\mathrm{Var}(\bar{Z}_k)}{\mu_k^2} = \frac{s^{-2} \cdot s \cdot \mathrm{Var}(Z_k)}{\mu_k^2} = \frac{s^{-1}(\mu_k - \mu_k^2)}{\mu_k^2} = s^{-1}(\mu_k^{-1} - 1) \leq 2s^{-1}$$

We shall take as our estimator for $|\Omega(G)|$ the random variable $Y = q^n \mu_1 \cdots \mu_m$.

The variance of Y can be bounded as:

$$\begin{aligned}
\frac{\mathrm{Var}(Y)}{E(Y)^2} &= \frac{\mathrm{Var}(\bar{Z}_1 \cdots \bar{Z}_m)}{(\mu_1 \cdots \mu_m)^2} \\
&= \prod_{k=1}^{m} \left(1 + \frac{\mathrm{Var}(\bar{Z}_k)}{\mu_k^2}\right) - 1 \\
&\leq \left(1 + \frac{2}{s}\right)^m - 1 \qquad\qquad s = \lceil 74\tfrac{m}{\varepsilon^2} \rceil \;\Rightarrow\; \tfrac{2}{s} \leq \tfrac{2\varepsilon^2}{74m} = \tfrac{\varepsilon^2}{37m} \\
&\leq \left(1 + \frac{\varepsilon^2}{37m}\right)^m - 1 \\
&\leq e^{\varepsilon^2/37} - 1 \qquad\qquad e^x - 1 = x + \underbrace{\frac{x^2}{2!} + \frac{x^3}{3!} + \cdots}_{\text{small!}} \\
&\leq \frac{\varepsilon^2}{36}.
\end{aligned}$$

Since by Chebyshev's inequality:

$$\Pr(|Y - E(Y)| \geq \lambda E(Y)) \leq \frac{1}{\lambda^2} \frac{\mathrm{Var}(Y)}{E(Y)^2}$$

i.e.

$$\Pr\left(\left|\frac{Y}{q^n} - \mu_1 \cdots \mu_m\right| \geq \lambda \mu_1 \cdots \mu_m\right) \leq \frac{1}{\lambda^2} \frac{\varepsilon^2}{36}$$

we obtain, by choosing $\lambda = \varepsilon/3$, the bound

$$\Pr\left(\left(1 - \frac{\varepsilon}{3}\right)\mu_1 \cdots \mu_m \leq q^{-n} Y \leq \left(1 + \frac{\varepsilon}{3}\right)\mu_1 \cdots \mu_m\right) \geq \frac{3}{4}.$$

But from inequality (6) we obtain the bound

$$\begin{aligned}
&\left(1 - \frac{\varepsilon}{3m}\right)^m \rho_1 \cdots \rho_m \leq \mu_1 \cdots \mu_m \leq \left(1 + \frac{\varepsilon}{3m}\right)^m \rho_1 \cdots \rho_m \\
\Rightarrow\; &\left(1 - \frac{\varepsilon}{2}\right)\rho_1 \cdots \rho_m \leq \mu_1 \cdots \mu_m \leq \left(1 + \frac{\varepsilon}{2}\right)\rho_1 \cdots \rho_m
\end{aligned}$$

Putting these two bounds together yields the desired fpras condition:

$$\Pr\left((1-\varepsilon)\underbrace{q^n \rho_1 \cdots \rho_m}_{|\Omega(G)|} \leq Y \leq (1+\varepsilon)\underbrace{q^n \rho_1 \cdots \rho_m}_{|\Omega(G)|}\right) \geq \frac{3}{4}.$$