Stats 318: Lecture # 15

Agenda: Monte Carlo Methods

- Approximate Counting
- How many $q$-colorings?
Monte Carlo method

Collection of tools for estimating values through sampling & estimation

Example: How do we estimate the number $\pi = 3.1415\ldots$?

Consider a point $(x, y)$ chosen uniformly at random inside the square $D$: Disk centered at $(0, 0)$ of radius 1, has area $\pi$

$$
\therefore Z = \begin{cases} 1 & \text{if } (x, y) \in D \\ 0 & \text{otherwise} \end{cases}
$$

obeys $\mathbb{P}(Z = 1) = \frac{\pi}{4}$

Run this experiment $n$ times $\rightarrow Z_1, \ldots, Z_n$

$$
\mathbb{E} \overline{Z} = \frac{\pi}{4} \implies 4 \overline{Z} \text{ is a natural estimate of } \pi
$$

$$
\mathbb{P}( |4\overline{Z} - \pi| \geq \epsilon \pi) \leq \frac{1}{\epsilon^2 \pi^2} \text{var}(4\overline{Z}) = \frac{3(4 - \pi)}{n\epsilon^2 \pi}
$$
Better results are possible: e.g. Chernoff bound

\[ \mathbb{P}( |4\overline{Z} - \pi| \geq \epsilon \pi ) \leq 2e^{-n\pi \epsilon^2 / 12} \]

Sufficiently large # of samples \( \implies \) approx is as tight as we wish

**Definition**

A randomized algorithm gives an \((\epsilon, \delta)\) – approximation for the value \(V\) if output \(X\) obeys

\[ \mathbb{P}( |X - V| \geq \epsilon V ) \leq \delta \]

Example shows that \( n \geq \frac{12 \log(2/\delta)}{\pi \epsilon^2} \) gives an \((\epsilon, \delta)\) approx algorithm
Why settle for an approximation?

- For $\pi$, we cannot find an exact answer
- Many problems for which existence of an efficient algorithm would imply $P = NP$
  ~ Does not preclude the possibilities of efficient *approximate* algorithms
Counting

- Given some set $\mathcal{X}$, what is the number of elements in $\mathcal{X}$?
- How many feasible configurations in the model of hard spheres?
- How many $q$-colorings of a graph?
- How many assignments obey a Boolean formula in disjunctive normal form?

Related problem: what is the normalizing constant in the Ising model?
\textbf{\textit{q}-coloring of a graph}

Wish to estimate number $N$ of $q$-colorings of a graph $G = (V, E)$

- **Naive approach: rejection method**
  - Assign each vertex to a color, independently of other vertices
  - Each configuration arises wp $1/q^k$, $k = |V|$

\begin{equation*}
Y = \begin{cases} 
1 & \text{if } q \text{ coloring} \\
0 & \text{otherwise}
\end{cases} \implies \mathbb{P}(Y = 1) = \frac{N}{q^k}
\end{equation*}

Repeat experiment and use $q^k \overline{Y}$ to estimate $N$

Suppose in $n$ trials, we do not get a $q$-coloring $\implies$ bad estimate

\begin{equation*}
\mathbb{P}(Y = 1) \leq \left(\frac{q-1}{q}\right)^k \quad [\text{connected graph}]
\end{equation*}
\[ \implies \mathbb{P}(\text{at least one coloring}) \leq n \left( \frac{q - 1}{q} \right)^k \]

To have at least $1$ $q$-coloring wp $\geq 50\%$ we need at least

\[ n \geq \frac{1}{2} \left( \frac{q}{q - 1} \right)^k \]

Exponential growth in $|V|$ $\implies$ useless for large graphs

To estimate a small probability requires a lot of simulations!
Approximate counting

Suppose each note has at most $d$ neighbors

$$|V| = k \quad |E| = m \leq dk$$

Enumerate edges $E = \{e_1, \ldots, e_m\}$

Subgraphs $G_0, G_1, \ldots, G_m$

$G_0 = (V, \emptyset)$ nodes $V$ & no edges

$G_1 = (V, \{e_1\})$ note $G_m = G$

$\vdots$

$G_j = (V, \{e_1, \ldots, e_j\})$

$N_j : \# \text{ of } q\text{-colorings of } G_j$

$$N_m = \frac{N_m}{N_{m-1}} \times \frac{N_{m-1}}{N_{m-2}} \times \ldots \times \frac{N_1}{N_0} \times N_0$$

If we can estimate each factor, could get a reasonable estimate of $N = N_m$
\[ N_0 = q^k \]

Consider \( \frac{N_j}{N_{j-1}} \)

Let \( v \) & \( v' \) be the end vertices of \( e_j \)

\[ \{ q \text{-colorings of } G_j \} = \{ q \text{ colorings of } G_{j-1} \text{ with} \]
\[ \text{color}(v) \neq \text{color}(v') \} \]

\( \frac{N_j}{N_{j-1}} \) proportion of colorings of \( G_{j-1} \) obeying \( \text{color}(v) \neq \text{color}(v') \)

\[ \implies \frac{N_j}{N_{j-1}} = \mathbb{P}(X(v) \neq X(v')) \quad X \text{ random } q \text{ coloring of } G_{j-1} \]

Simulate a random coloring of \( G_{j-1} \) several times

\( Y_j : \text{Fraction of time colors of } v \text{ & } v' \text{ differ } \implies \text{good estimate of } \frac{N_j}{N_{j-1}} \)

Apply procedure for each factor : \( \hat{N} = q^k \prod_{j=1}^{m} Y_j \)
\[ \hat{N} = q^k \prod_{j=1}^{m} Y_j \]

How accurate is this?

Suppose that \( \text{wp at least } 1 - \delta \)

\[ \frac{N_j}{N_{j-1}} \left( 1 - \frac{\epsilon}{2m} \right) \leq Y_j \leq \frac{N_j}{N_{j-1}} \left( 1 + \frac{\epsilon}{2m} \right) \quad \forall j \quad (\ast) \]

\[ \implies 1 - \epsilon \leq \frac{\hat{N}}{N} \leq 1 + \epsilon \quad \text{wp at least } 1 - \delta \]

\[ \iff |\hat{N} - N| \leq \epsilon N \quad \text{wp at least } 1 - \delta \quad \text{which is what we want} \]

\[ (\ast) \iff \left| Y_j - \frac{N_j}{N_{j-1}} \right| \leq \frac{\epsilon}{2m} \frac{N_j}{N_{j-1}} \]

Need to make sure empirical proportions approximate true proportions
Two sources of error

(i) Gibbs sampler after $T$ steps has not reached equilibrium

(ii) Finitely many simulations $\implies Y_j \neq \frac{N_j}{N_{j-1}}$

Remark: $N_j/N_{j-1} \geq 1/2$ (do it!)

\[ \therefore \text{ suffices to make sure that } -\frac{\epsilon}{4m} \leq Y_j - \frac{N_j}{N_{j-1}} \leq \frac{\epsilon}{4m} \]
With exact uniform sampling

**Chernoff bound:** \( X_1, \ldots, X_n \) iid Ber\((p)\)

\[
\mathbb{P}(\mid X_1 + \ldots + X_n - np \mid \geq tnp) \leq 2e^{-np \frac{t^2}{3}}
\]

\( n \) simulations for \( Y_j \) \((\mathbb{E} Y_j = p_j)\)

\[
\mathbb{P} \left( \mid Y_j - p_j \mid \geq \frac{\epsilon}{2m} p_j \right) \leq 2e^{-\frac{np_j \epsilon^2}{12m^2}} \leq 2e^{-\frac{n\epsilon^2}{24m^2}}
\]

\[
\mathbb{P}(\mid Y_j - p_j \mid \geq \frac{\epsilon}{2m} p_j \text{ for some } j) \leq 2m \exp(-\frac{n\epsilon^2}{24m^2}) \leq \delta \text{ if }
\]

\[
n \geq \frac{24m^2}{\epsilon^2 \log(\delta/2m)}
\]

\( \implies (\epsilon, \delta) \) approximation algorithm
Analysis using Gibbs sampler

$t$ steps of chain:

\[
P(Y = 1) = \mu_t \{\text{color}(v) \neq \text{color}(v')\} = p_j^{(t)}
\]

\[
|Y_j - p_j| \leq |Y_j - p_j^{(t)}| + |p_j^{(t)} - p_j|
\]

Sufficient to establish

(i) \( |Y_j - p_j^{(t)}| \leq \frac{\epsilon}{8m} \)

(ii) \( |p_j^{(t)} - p_j| \leq \frac{\epsilon}{8m} \)

(i) As before

\[
P \left( |Y_j - p_j^{(t)}| \geq \frac{\epsilon}{8m} \right) \leq 2\exp \left( -\frac{n}{3} p_j^{(t)} \left( \frac{\epsilon}{8m} \right)^2 \right) \leq 2\exp \left( -\frac{n}{6} p_j^{(t)} \left( \frac{\epsilon}{8m} \right)^2 \right) \leq \delta/m
\]

Number of trials must obey \( n \geq C_0 \left( \frac{m}{\epsilon} \right)^2 \log(2m/\delta) \quad C_0 = 6 \times 8^2 \)
(ii) How many steps for Gibbs sampler?

Sufficient to have

$$\|\mu_t - \text{unif}\|_{TV} \leq \frac{\epsilon}{8m}$$

**Theorem (Thm 5.7 (textbook))**

Assume $m \leq dk$. Then

$$q > 3d \implies t_{mix}(\epsilon) \leq C_1 k[\log k + \log 1/\epsilon] \quad C_1 = \frac{1}{1 - 3d/q}$$
Summary

- $m$ fractions $Y_j$ to compute
- Each one uses $C_0 \frac{m^2}{\epsilon^2} \log \left( \frac{2m}{\delta} \right)$ trials
- Each trial requires no more than $C_1 k [2 \log k + \log \frac{1}{\epsilon} + \log 8]$ steps of Gibbs sampler

Total complexity: $m \times C_0 \frac{m^2}{\epsilon^2} \log \left( \frac{2m}{\delta} \right) \times C_1 k [2 \log k + \log \frac{1}{\epsilon} + \log 8]$

Polynomial in $k$ or $m$, $1/\epsilon$, $\log(1/\delta)$