Approximation with Sampling

Daniel Khashabi

Summer 2013
Last Update: April 27, 2015

1 Introduction

In all of the learning problems, after the parametric modelling we need to devise a way to learn the optimal parameters, using some training samplings, or some indirect rules, with respect to some criterion, e.g. a defined loss-function. Usually this can be cast as maximizing (or minimizing, with an additional negative sign) a function of parameters, training data, and the prior knowledge, commonly known as MAP or *maximum a posteriori*.

$$\mathcal{L} = \log p(\mathcal{D}|\Theta) \rightarrow \Theta^* = \max_{\Theta} \log p(\mathcal{D}|\Theta)$$

Since the posterior distribution (or function, if not normalized) is usually a complicated function, it is not straightforward to maximize it directly with respect to model parameters. One approach can be approximating this function and finding the sub-optimal parameters. The other approach which is mostly studied here, is statistical sampling methods, which take many samples of the model, to simulate the behaviour of the model. These methods are usually slow, and exact asymptotically (if they run long enough).

Before starting on learning based on sampling, we should first learn how to sample complicated distributions. Usually it could be assumed that we know how to sample a uniform distribution, and we aim at generalizing it to sampling other complicated distributions.

2 Sampling a proper distribution

In theory, there is an easy way to sample any distribution, by finding an invertible parametric form which converts the variables in two distributions. Let's say we know how to sample $p(x)$, our goal is to get the distribution $p(z)$ by finding a parametric form for $x \rightarrow z$. We get the distribution $p(z)$ by the following conversion between two distributions,

$$p(x) = p(z) \left| \frac{dz}{dx} \right|.$$

In Figure 1 a probability distribution $p(x)$, and its cumulative distribution $P(X \leq x) = \int_{-\infty}^{x} p(x')dx'$ is depicted. If we sample the y-axis uniformly, find the corresponding points in the CDF curve,

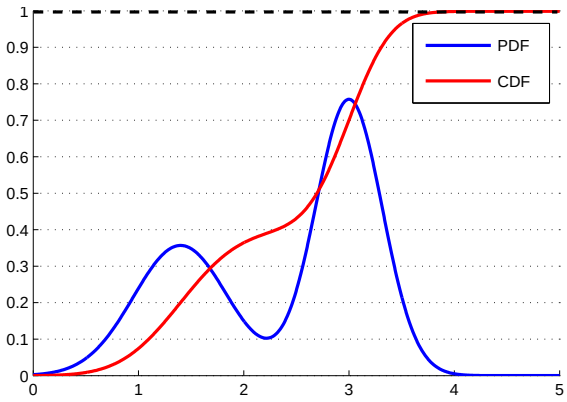

Figure 1: A sample distribution, and its cumulative distribution.

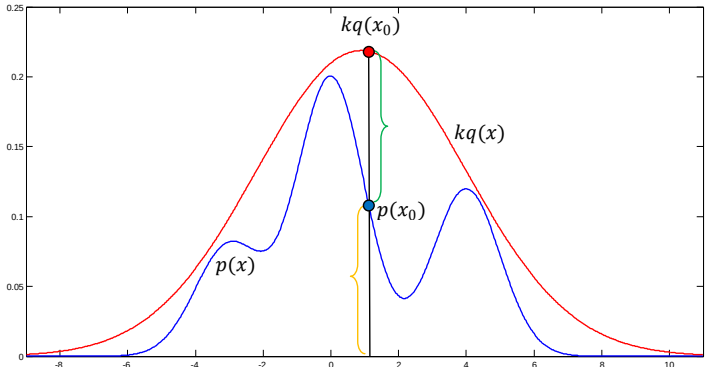

Figure 2: A sample distribution, and its cumulative distribution.

and map them on the the x-axis, the corresponding points are distributed according to $p(x)$. In mathematical form, this can be explained in the following form,

$$p(x) = p(z)\left|\frac{dz}{dx}\right| \ , p(z) = 1 \Rightarrow z = h(x) = P(X \leq x) = \int_{-\infty}^{x} p(x^{'})dx^{'} \Rightarrow x = h^{-1}(z).$$

2.1 Rejection sampling

Since in many modelling problems, it is not easy to find an analytical for the CDF, we prefer to find a way for sampling an arbitrary distribution, without the need for finding its CDF. One of these methods is called *rejection sampling*.

Let's say we want to sample a distribution $p(x)$ which has a complicated form, and we can't find its CDF. IJn rejection sampling, we find another distribution $q(x)$ which supports the target distribution, i.e. $p(x)$. In other words, for any $x^{'}$ in the domain of the distributions, $q(x^{'}) > p(x^{'})$. Note that, in general $q(x)$ doesn't have to be a proper distribution, in the sense that the area under it sum up to one, but it needs to be of the forms which is easy to sample from. Then $kq(x),\ k \in \mathbb{R}_{++}$

which is easy to sample from, and supports $p(x)$ can be used. In Figure 2 a complicated target distribution$p(x)$, and a supporting distribution $kq(x)$ are shown.

The procedure for rejection sampling is as following: first we sample a point x_0 from the distribution $kq(x)$. Because $k > 0$, we know that $kq(x) > 0$. We create a uniform distributions on $[0, kq(x)]$, and sample a point from that. If the point is greater that $p(x_0)$ we accept it as a sample of $p(x)$, if not, we reject it. It can be shown that in long-run the accepted samples will have distribution according to $p(x)$ (proof?).
To decrease the ratio of the rejected samples it is necessary to choose the supporting distribution $kq(x)$ as close as possible to $p(x)$, though it might need might be hard to find such a distribution when handling high-dimensional distributions. Also it can be shown, roughly speaking, the probability of a sample being accepted diminishes exponentially with the number of the dimensions. This makes rejection sampling very hard to use in high-dimensional problems, and with a very complicated form, which are hard to visualize. There are a few works which aim at finding better supporting distribution adaptively by using peace-wise exponential functions, or log-concave families (see [2, 1])

Usually in probabilistic inference problems, we are dealing with a real ratio of the target distribution $p(x)$. In other words, if we assume that $p(x) = \frac{1}{\mathcal{Z}}\tilde{p}(x)$, where $\mathcal{Z} = \int p(x)dx$ is a normalizing constant, we usually only have $\tilde{p}(x)$, and it is hard to normalize. Thus, there is a big motivation for finding methods which can use the unnormalized function $\tilde{p}(x)$, and give samples of $p(x)$ without directly having it.

2.2 Sampling for approximating integrals

Let's say we want to approximate the following integration, $\int f(z)p(z)dz$ which is equivalent to the following expectation, $\mathbb{E}_p[f]$. We can use sample mean as an estimator of the statistical mean, and we can approximate the above expectation by sampling from $p(x)$,

$$\mathbb{E}_p[f] \approx \frac{1}{L}\sum_{i=1}^{L} f(z^i), \quad x^i \sim p(x).$$

Note that in general this trick could be used for approximating any integration with a proper choice of $p(x)$. Also it can easily verified that sample mean is an unbiased estimator the statistical mean.

Let's say we don't know how to sample $p(x)$ and we want to approximate $\mathbb{E}_p[f]$. We choose a distribution $q(x)$ which we know how to sample from, and change the expectation using it,

$$\mathbb{E}_p[f] = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \approx \frac{1}{L}\sum_{i=1}^{L} f(z^i)\frac{p(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

This let's us sample from $q(x)$ for approximating the value of the sample expectation. This trick is usually called *importance sampling*. Practical usage of this trick demands careful considerations. One important point is that, to get realistic answers, $q(x)$ must be non-zero (or not very small) wherever $p(x)f(x)$ is not zero. This trick has many interesting applications; for example one can

use use this trick to calculate expectation of events happening when their probability is very small, e.g. calculating "bit error rate" in a communication system [4].

Let's consider the case where we don't have the distribution $p(x)$ but we only have a positive ratio of that. In other words, if $p(z) = \frac{1}{Z}\tilde{p}(x)$, we only have $\frac{1}{Z}\tilde{p}(x)$ and calculation of the normalizing constant is too costly that we don't want to do it. We can simplify the previous formulations as following,

$$\mathbb{E}_p[f] = \int f(z)p(z)dz = \frac{1}{Z_p}\int f(z)\tilde{p}(z)dz = \frac{1}{Z_p}\int f(z)\frac{\tilde{p}(z)}{q(z)}q(z)dz = \frac{1}{Z_p}\sum_{i=1}^{L} f(z^i)\frac{\tilde{p}(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

A similar thing can be done to find an estimation of the normalizing constant,

$$1 = \mathbb{E}_p[1] = \int p(z)dz = \frac{1}{Z_p}\int \tilde{p}(z)dz = \frac{1}{Z_p}\int \frac{\tilde{p}(z)}{q(z)}q(z)dz = \frac{1}{Z_p}\sum_{i=1}^{L} \frac{\tilde{p}(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

$$\Rightarrow Z_p = \sum_{i=1}^{L} \frac{\tilde{p}(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

Now using the above unbiased estimations, the estimation for the expectation is as following,

$$\Rightarrow \mathbb{E}_p[f] = \frac{\sum_{i=1}^{L} f(z^i)\frac{\tilde{p}(z^i)}{q(z^i)}}{\sum_{i=1}^{L} \frac{\tilde{p}(z^i)}{q(z^i)}}, \quad z^i \sim q(z).$$

This estimator is *biased* estimator of the target expectation (proof?), it is not always the case that the ratio of any two unbiased estimators is biased estimator (example?).

2.3 Gibbs sampling

Let's say we want to sample from a multivariate distribution $p(x, y)$. Since sampling jointly sample from (x, y) we can sample for each variable, from the marginal distributions,

$$\begin{cases} x_t \sim p(x|y_{t-1}) \\ y_t \sim p(y|x_t) \end{cases}$$

In general this can be applied to any distribution with any number of the variables. More details on convergence proof and properties could be found at [7, 6]. The idea of Gibbs sampling in statistics is very similar to "coordinate descent" optimization of multivariate objective functions in optimization(more?).

2.4 Markov Chain Monte Carlo(MCMC)

MCMC methods *implicitly* create makov chains which have the stationary distributions the same as that of the target distribution. At each step a new sample $x^{(i)}$ is proposed using a *transition distribution*, $\mathcal{P}(x, x')$,

$$x^{(i-1)} \xrightarrow{\mathcal{P}} x^{(i)}.$$

There are many other names used to call this function, e.g. *Jumping Distribution*, *Proposal Distribution*, *Candidate Generating Distribution*.

One of the families of MCMCs is Metopolis-Hastings methods which introduced in [5, 3]. Let's say we want to sample from $p(x) = \frac{1}{\mathcal{Z}}\tilde{p}(x)$, and let's assume that we don't have the normalization constant $\mathcal{Z}$. We define a transition distribution,

$$q(z_2|z_1) = \Pr(z_1 \to z_2).$$

The steps of the algorithm are shown in Algorithm 1.

Algorithm 1: Metropolis-Hastings algorithm

Start with random samples z_0, s.t. $p(z_0) > 0$.

repeat

genrate random sample, z_* from proposal distribution, $z_* \sim q(Z|z_{t-1})$, given the sandom sample of the previous iteration z_t.

Calculate: $\alpha(z_*, z_{t+1}) = \min\left\{1, \frac{\tilde{p}(z_*)q(z_{t-1}|z_*)}{\tilde{p}(z_{t-1})q(z_*|z_{t-1})}\right\}$.

$$\alpha(z_*, z_{t+1}) = \begin{cases} \geq 1 & : \text{Accept the sample: } z_t = z_* \\ < 1 & : \text{Accept the sample with probability of } \alpha. \end{cases}$$

until *TERMINATION-CONDITION*;

To get a good approximation of the samples found from the above method, it is necessary to throw away the samples until a time *burn-in* period k where the samples $x_{k+1}, x_{k+2}, \ldots$ get closer to realistic samples of the target distribution, or the markov chain gets close enough to its stationary distribution.

Why running the algorithm 1 will result in convergence (detailed balance)? Suppose we are going from state x to state y and (without loss of generality) $p(x)q(y|x) > p(y)q(x|y)$:

$$\begin{cases} p(y|x) = q(y|x)\alpha(y|x) = q(y|x)\frac{p(y)q(x|y)}{p(x)q(y|x)} = q(x|y)\frac{p(y)}{p(x)} \\ p(x|y) = q(x|y)\alpha(x|y) = q(x|y) \end{cases} \Rightarrow p(y|x)p(x) = p(x|y)p(y)$$

2.4.1 Objective function

It can be shown that Metropolis-Hastings algorithm finds l_1-projection of $q(x|y)$ onto the space of reversible Markov chains with stationary distribution $p(x)$.

$$\min_{Q \in R(p)} \sum_x \sum_{y \neq x} |p(x)q(y|x) - p(y)Q(x|y)|$$

2.4.2 Convergence

Definition 1. *The conductance of a cut $S, \bar{S}$ is defined as*

$$\Phi(S) = \frac{\sum_{i \in S, j \in \bar{S}} a_{ij}}{\min\left\{a(S), a(\bar{S})\right\}}$$

Where $a(S) \triangleq \sum_{i,j \in S} a_{ij}$. The conductance of a graph (also known as Cheeger constant) is defined as $\Phi(G) = \min_{S \subseteq V} \Phi(S)$, or equivalently,

$$\Phi(G) = \min_{S \subseteq V, 0 \leq a(S) \leq a(V)/2} \frac{\sum_{i \in S, j \in \bar{S}} a_{ij}}{a(S)}$$

Theorem 1. *The ϵ-mixing time of a random walk on an undirected graph is*

$$O\left(\frac{\ln(1/q_{\min})}{\Phi^2\epsilon^2}\right)$$

where $q_{\min}$ is the minimum stationary probability of any state.

Example 1 (Mixing on a line graph)**.** *We have a line with N nodes, with self-loops at the two ends.*

1. *What is the stationary distribution of this walk? The transition matrix P is a all zeros, except subdiagonal and superdiagonal elements are all 0.5. Also $P_{11} = P_{nn} = 0.5$ (self-loops). We show tat the stationary distribution is uniform:*

   $$\pi = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$$

   It is easy to verify that: $\pi P = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$ which proves the result.

2. *What is the (normalized) conductance $\Phi = \min_{|S|\leq N/2} \Phi(S)$ of this graph?*

   $$\Phi = \min_{|S|\leq N/2} \frac{\frac{1}{n} \times \frac{1}{2}}{\min \pi(S)} \leq O\left(\frac{1}{n}\right)$$

3. *What is the bound on the mixing time on this graph?*

   $$T_{\max} \leq \frac{\frac{2}{\epsilon\sqrt{\pi_{\min}}}}{\Phi^2} \leq O(n^2 \log \frac{n}{\sqrt{\epsilon}})$$

4. *Show that for $\epsilon = 1/10$, the mixing time is $\Omega(N^2)$. To show this we use the fact that the largest hitting time on this graph $h_{1,N}$ is $\Omega(N^2)$ (since the mixing time needs to hold for any initialization of the masses, specifically when all mass is assigned to one of vertices at the ends). This shows that the bound on the mixing time is tight up to a logarithmic factor.*

Example 2 (Sampling independent sets)**.** *G is an n-node graph. We want to sample independent set S proportional* $\exp|S|$.

1. *Consider the following algorithm:*

   (a) *X_0 is an arbitrary initial independent set of $G(V, E)$*
   (b) *To compute X_{i+1}, the ith independent set:*
      i. *Randomly sample $v \in V$.*
      ii. *If $v \in X_i$ then $X_{i+1} \leftarrow X_i \setminus \{v\}$.*
      iii. *If $v \notin X_i$ then $X_{i+1} \leftarrow X_i \cup \{v\}$, if X_{i+1} still remains an independent set.*
      iv. *Otherwise $X_{i+1} = X_i$.*

First we claim that this algorithm visits any state, where state being an independent set. Assuming that the graph is not disconnected, any state (independent set) could be reached via another state (independent set) by addition/deletion of a node. More generally, to get from I to I', remove all vertices of I and then add all vertices of I' (irreducible).

Also, this sampling strategy is uniform over the state of the problem (the set of independent sets). First, the only source of randomness is in choosing v with probability $\frac{1}{|V|}$. Given a $v \in V$, everything is deterministic.

Fix $v \in V$, and consider the state X_{i+1}.

- *If $v \in X_{i+1}$, we have transition to X_{i+1} essentially from $X_i = X_{i+1} \setminus \{v\}$ by adding v.*
- *If $v \notin X_{i+1}$ and adding v to X_{i+1} would not break any constraints, we must have come to X_{i+1} from $X_i = X_{i+1} \cup \{v\}$ by removing the vertex v.*
- *If $v \notin X_{i+1}$ and adding v to X_{i+1} would not break any constraints, we have come to X_{i+1} via its self-loop.*

First note that the Markov chain has nonzero self-loop probability, it is aperiodic.

Given the above three transitions for three different states, and given the fact that for any fixed $v \in V$ the transition is done from three different states, the contribution of each v over different states is uniform, which basically means that $P_{X_i,X_{i+1}} = \frac{1}{|V|}$.

Given the uniform distribution over states, irreduciblity and aperiodicity, MCMC is expected to converge to the right distribution.

Example 3 (Sampling matchings)**.** *Matchings are edge independent sets for graphs. Formally, for any graph $G(V, E)$, an edge subset is a matching if any distinct pair of do not share an endpoint. A matching M is said to be perfect if $|M| = |V|$ since it matches all the vertices. Consider the problem of uniformly sampling a perfect matching for a dense bipartite graph.*
Here we explain how to sample perfect matchings with uniform probability. We fix a bipartite graph $G(V1, V2, E)$ with $|V1| = |V2| = N$, and minimum degree at least $N/2$. Let $\mathcal{M}_k$ denote the set of distinct matchings of size k in G. Thus, $\mathcal{M}_N$ denote the set of the perfect matchings of the dense bipartite graph G. We can consider the following types of transitions closed within the space:

- *Reduce: For a graph $M \in \mathcal{M}_k$ and an edge $uv \in E$, remove it $E \leftarrow E \setminus \{uv\}$. The resulting graph belongs to $M \in \mathcal{M}_{k-1}$.*
- *Augment: For a graph $M \in \mathcal{M}_k$ and two vertices u and v unmatched in M, add uv as en edge $E \leftarrow E \cup \{uv\}$. The resulting graph belongs to $M \in \mathcal{M}_{k+1}$.*
- *Swap: For a graph $M \in \mathcal{M}_k$ with an edge $uv \in E$ and vertex w unmatched in M, $E \leftarrow E \cup \{uw\} \setminus \{uv\}$. The resulting graph belongs to $M \in \mathcal{M}_k$*

A Markov chain in state $\mathcal{M}_{k+1} \cup \mathcal{M}_k$ is defined as follow:

With probability $1/2$*, stay at the current state; otherwise uniformly choose a random edge, and if any one of the Reduce, Augment and Swap is applicable, apply the transformation; otherwise stay at the current state.*

This Markov chain is time-reversible, and the stationary distribution is the uniform distribution over the set of the perfect matchings. Since each transition of the Markov chain has the same associated probability of $\frac{1}{N^2}$*. Also any* $M \in \mathcal{M}_{k+1} \cup \mathcal{M}_k$ *can be transformed to any* $M' \in \mathcal{M}_{k+1} \cup \mathcal{M}_k$ *by a suitable sequence of augmentations, reductions, and swaps. So the Markov chain will eventually return a near-uniform matching.*

Now we suppose that the edges are weighted. Specifically, for any edge $(i, j) \in E$ *we have* $w_{ij} \in \mathbb{R}^+$*. The joint probability distribution of the matchings can be written as:*

$$\mu(\sigma) = \frac{1}{Z} \exp\left\{\sum_i w_{i\sigma(i)}\right\},$$

where $\sigma : V1 \to V2$ *is a mapping which encodes the matching.*
We know:

$$\mu(\sigma) = \frac{\exp\left\{\sum_i w_{i\sigma(i)}\right\}}{\sum_\sigma \exp\left\{\sum_i w_{i\sigma(i)}\right\}} \geq \frac{\exp\left\{\sum_i w_{i\sigma(i)}\right\}}{\sum_\sigma \exp\left\{\sum_i w^*\right\}} \geq \frac{1}{N! \exp\left\{N w^*\right\}},$$

where $w^* = \max_{i,j} w_{ij}$*.*

Suppose we are in matching defined by σ*. For two fixed vertices* $i, i' \in V$ $(i \neq i')$*, define the new matching* σ' *in the following form:*

$$\sigma'(i) = \sigma(i'), \quad \sigma'(i') = \sigma(i)$$

The probability of transition from σ *to* σ'

$$\begin{aligned} R &= \min\left\{1, \frac{\mu(\sigma)}{\mu(\sigma)}\right\} \\ &= \min\left\{1, \frac{\exp\left\{\sum_j w_{j\sigma(j)}\right\}}{\exp\left\{\sum_j w_{j\sigma'(j)}\right\}}\right\} \\ &= \min\left\{1, \frac{\exp\{w_{i\sigma(i)} + w_{i'\sigma(i')}\}}{\exp\{w_{i'\sigma(i)} + w_{i\sigma(i')}\}}\right\} \end{aligned}$$

We then prove that:

$$\mathbb{P}\left(\sigma \to \sigma'\right) = \frac{1}{N^2} \min\left\{1, \frac{\exp\{w_{i\sigma(i)} + w_{i'\sigma(i')}\}}{\exp\{w_{i'\sigma(i)} + w_{i\sigma(i')}\}}\right\} \geq \frac{1}{N^2} \min\left\{1, \frac{\exp\{0+0\}}{\exp\{w^* + w^*\}}\right\} \geq \frac{1}{N^2} \frac{1}{\exp\{2w^*\}}$$

Given the results above,

$$\begin{aligned} \Phi &= \min_S \frac{\sum_{\sigma \in S, \sigma' \in S'} \mu(\sigma) \mathbb{P}\left(\sigma \to \sigma'\right)}{\mu(S)\mu(S^c)} \\ &\geq \frac{1}{N^2 \exp\{2w^*\}} \times \frac{1}{N! \exp\left\{N w^*\right\}} \end{aligned}$$

where $\mu(S) \triangleq \sum_{\sigma \in S} \mu(\sigma)$.

We know that the $T \leq \frac{2 \log \frac{2}{\epsilon \sqrt{\pi_{\min}}}}{\Phi^2} \leq 2 \log \frac{2}{\epsilon \sqrt{\pi_{\min}}} \times N^2 N! \exp\{w^*(N+2)\}$

2.5 Notions from random walks

[Intro:TODO]

Example 4 (Cover time of n-node clique)**.** *1. Cover time of* n*-node clique: Since there is a path to from any node to another, the problem is exactly the same as coupon collector problem:*

> *Coupon Collector Problem: A set of* n *urns each containing infinite number of coupons with the same sign. Each time we choose a coupon from one urn. What is the expected number of trials until we get at least one coupon from each urn?*

Claim: The expected time to see each coupon at least once is $O(n \log n)$*. Let* T *be the time needed to collect all of the coupons and* t_i *the time to collect* i*th new coupon, after* $i-1$ *unique coupons are collected. We know:* $T = \sum_i t_i$*. The probability of observing new coupon after observing* $i-1$ *is* $\frac{n-(i-1)}{n}$*. Therefore expectation of* t_i *is* $1/p_i$*.*

$$\mathbb{E}T = \sum_i \mathbb{E}t_i = \sum_i \frac{1}{p_i} = \sum_i \frac{n}{n-(i-1)} = n \sum_i \frac{1}{i} = n \log n$$

2. A walk of length $2n \log n$ *on a* n*-node clique has probability at last* $1 - 1/n$ *of visiting all nodes. For a fixed unseen element the probability of not seeing it after* k *iterations:*

$$\mathbb{P}\left(\textit{not seeing a fixed element for } k \textit{ trials}\right) = \left(1 - \frac{1}{n}\right)^k \leq \exp\left(-\frac{k}{n}\right)$$

$$\Rightarrow \mathbb{P}\left(\textit{Not seeing all elements untll } k \textit{ trials}\right) = \bigcup_{i \in [n]} \mathbb{P}\left(\textit{not seeing a fixed element for } k \textit{ trials}\right)$$
$$\leq n \exp\left(-\frac{k}{n}\right)$$

If $k = 2n \log n$*:*

$$\Rightarrow \mathbb{P}\left(\textit{Not seeing all elements untll } 2n \log n \textit{ trials}\right) \leq n \exp\left(\log n^{-2}\right) = \frac{1}{n}$$

which concludes the proof.

Example 5. *Does adding an edge to a grad reduce the cover time? No. We show this with an example. Suppose we have* n *node. If we add edges and make it a line graph the cover time will be* $O(n^2)$*. If we add further edges to make it a lollipop graph (Figure 3) the cover time will increase to* $O(n^3)$*. Further addition of edges to make it a complete graph will reduce the cover time to* $O(n \log n)$*.*

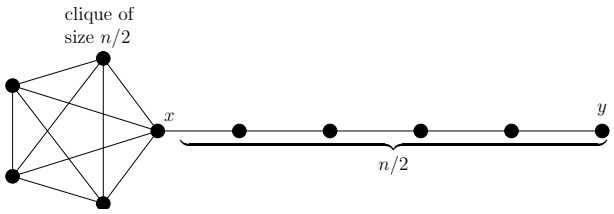

Figure 3: Lollipop graph has $O(n^3)$ cover time.

2.5.1 Hitting time

[hittingTimeIntro:TODO]

Example 6 (Hitting time in a clique)**.** *Hitting time for two (adjacent) nodes in n-node clique. Let's denote the hitting time with h. Since in a clique every two nodes are connected to each other the each times for any pair of nodes are the same.*

$$h = \frac{1}{d} \times 1 + \frac{d-1}{d}(1+h) \Rightarrow \boxed{h = d = n-1}$$

Example 7 (Hitting time in a cycle)**.** *Hitting time of of two adjacent nodes u and v on a n-node cycle. Claim: Let's denote the hitting time of two nodes with distance d with $H(d)$. The hitting time is $H(d) = d(n-d)$. To prove this without loss generality, suppose n is even. Then:*

$$H\left(\frac{n}{2}\right) = 0.5\left(H\left(\frac{n}{2}-1\right)+1\right) + 0.5\left(H\left(\frac{n}{2}+1\right)+1\right) = H\left(\frac{n}{2}-1\right)+1, \tag{1}$$

since $H(d) = H(n-d)$.

$$\Rightarrow \boxed{H\left(\frac{n}{2}\right) = H\left(\frac{n}{2}-1\right)+1}$$

Similarly:

$$H\left(\frac{n}{2}-1\right) = 0.5\left(H\left(\frac{n}{2}-2\right)+1\right) + 0.5\left(H\left(\frac{n}{2}\right)+1\right)$$

Combining it with Eq. 1 we get:

$$H\left(\frac{n}{2}-1\right) = 0.5\left(H\left(\frac{n}{2}-2\right)+1\right) + 0.5\left(H\left(\frac{n}{2}-1\right)+2\right)$$

$$\Rightarrow 2H\left(\frac{n}{2}-1\right) = H\left(\frac{n}{2}-2\right) + H\left(\frac{n}{2}-1\right) + 3$$

$$\Rightarrow \boxed{H\left(\frac{n}{2}-1\right) = H\left(\frac{n}{2}-2\right)+3}$$

$$H\left(\frac{n}{2}-2\right) = 0.5\left(H\left(\frac{n}{2}-3\right)+1\right) + 0.5\left(H\left(\frac{n}{2}-1\right)+1\right)$$

$$\Rightarrow H\left(\frac{n}{2}-2\right) = 0.5\left(H\left(\frac{n}{2}-3\right)+1\right) + 0.5\left(H\left(\frac{n}{2}-2\right)+4\right)$$

$$\Rightarrow \boxed{H\left(\frac{n}{2}-2\right) = H\left(\frac{n}{2}-3\right)+5}$$

Similarly we can prove (with induction) that:

$$\boxed{H\left(\frac{n}{2}-i\right)=H\left(\frac{n}{2}-i-1\right)+2i+1}$$

The hitting time for two adjacent vertices is:

$$\boxed{H\left(1\right)=2n-1},$$

since $H\left(0\right)=0$.

3 Bibliographical notes

More proofs on properties of MCMC could be found at [3].

References

[1] W.R. Gilks, NG Best, and KKC Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, pages 455–472, 1995.

[2] W.R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

[3] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[4] M. Jeruchim. Techniques for estimating the bit error rate in the simulation of digital communication systems. *Selected Areas in Communications, IEEE Journal on*, 2(1):153–170, 1984.

[5] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[6] A.E. Raftery and S. Lewis. How many iterations in the gibbs sampler. *Bayesian statistics*, 4(2):763–773, 1992.

[7] A.F.M. Smith and G.O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.