Sample Averages for Reversible Markov Chains

Xiaoyu Chen

This article is a simple rewrite of the results that appear in [Ald87] in a language that I am familiar with. The main purpose is to estimate $\mathbb{E}[f]$ by sampling while showing that our estimate is concentrated on $\mathbb{E}[f]$.

1 Estimate the Expectation of A Function

Given a distribution π over Ω, and a function $f : \Omega \to \mathbb{R}$, we would like to estimate $\bar{f} := \mathbb{E}[f]$ using independent experiments.

The most naive method to achieve this is to sample i.i.d. random variables $X_1, X_2, \cdots, X_n$ according to π. Then, let $\hat{f} := \frac{1}{n}\sum_i f(X_i)$ as our estimation for $\bar{f}$.

We have the following facts.

Fact 1.1.

$$\mathbb{E}[\hat{f}] = \mathbb{E}[\frac{1}{n}\sum_i f(X_i)] = \frac{1}{n}\sum_i \mathbb{E}[f] = \mathbb{E}[f] = \bar{f}$$

Fact 1.2.

$$\begin{aligned}
\mathrm{Var}\hat{f} &= \mathrm{Var}(\frac{1}{n}\sum_i f(X_i)) \\
&= \frac{1}{n^2}\mathrm{Var}(\sum_i f(X_i)) \\
&= \frac{1}{n^2}\sum_i \mathrm{Var} f(X_i) \qquad \textit{by independence} \\
&= \frac{1}{n}\mathrm{Var}_\pi f
\end{aligned}$$

So, from Chebyshev inequality, we have

$$\Pr[|\hat{f} - \bar{f}| > t] \leq \frac{\frac{1}{n}\mathrm{Var}_\pi f}{t^2}$$

In many cases, by setting an apropriate t, we may conclude the event which we want to happen really happens with high probability.

2 Estimate $\mathbb{E}[f]$ by A Reversible Markov Chain

Suppose we have an reversible Markov chain P with its unique stationary π and we want to estimate $\mathbb{E}[f]$. Surprisingly, it was shown in [Ald87] that, instead of sampling independent X_i according to π using P, we could run P from stationary for n steps and get $X_1, X_2, \cdots, X_n$ to estimate $\mathbb{E}[f]$ with a quite good result.

Theorem 2.1. *Suppose we runs P for N steps from stationary distribution (i.e. $X_0 \sim \pi$) and get $X_1, X_2, \cdots, X_N$, then we have*

- $\mathbb{E}[\hat{f}] = \bar{f}$
- $\mathrm{Var}[\hat{f}] \leq \alpha(\gamma N)$, *where* $\alpha(x) = \frac{2}{x^2}(e^{-x} + x) = \frac{2}{x}(xe^{-x} + 1)$

, where $\gamma := 1 - \lambda_2$.

Proof. $\mathbb{E}[\hat{f}] = \bar{f}$ is trivial, so we only prove the second part here. For convenience, we assume that $\bar{f} = 0$, which means $\langle f, \mathbf{1} \rangle_\pi = 0$ (i.e. $f \perp_\pi \mathbf{1}$). Then, Note that

$$\begin{aligned}
\mathbb{E}[f(X_0)f(X_t)] &= \sum_{x\in\Omega}\sum_{y\in\Omega} \pi(x)f(x)P^t(x,y)f(y) \\
&= \sum_{x\in\Omega} \pi(x)f(x) \sum_{y\in\Omega} P^t(x,y)f(y) \\
&= \sum_{x\in\Omega} \pi(x)f(x)P^t f(x) \\
&= \langle f, P^t f \rangle_\pi
\end{aligned}$$

Since P is time reversible, it is also a self-adjoint operator accroding to $\langle \cdot, \cdot \rangle_\pi$ (see [Som] for example). And, moreover, P has an eigenbasis according to $\langle \cdot, \cdot \rangle_\pi$. And we denote its eigenbasis as $f_1 = \mathbf{1}, f_2, \cdots, f_n$. So, we have

$$\begin{aligned}
\mathbb{E}[f(X_0)f(X_t)] &= \alpha_1^2 \langle f_1, P^t f_1 \rangle_\pi + \alpha_2^2 \langle f_2, P^t f_2 \rangle_\pi + \cdots \alpha_n^2 \langle f_n, P^t f_n \rangle_\pi \\
&\quad (\text{let } \alpha_i = \langle f_i, f \rangle_\pi) \\
&= \sum_i \alpha_i^2 \lambda_i^t \langle f_i, f_i \rangle_\pi = \sum_i \alpha_i^2 \lambda_i^t
\end{aligned}$$

Let $S_N = \sum_{i=1}^N f(X_i) = n\hat{f}$, then we have

$$
\begin{aligned}
\mathrm{Var}(S_N) &= \mathbb{E}S_N^2 \\
&= \sum_i \sum_j \mathbb{E}[f(X_i)f(X_j)] \quad \text{since we assume that } \bar{f} = 0 \\
&= \sum_i \sum_j \left(\sum_x \sum_y \pi(x) f(x) P^{|j-i|}(x,y) f(y) \right) \\
&= \sum_i \sum_j \mathbb{E}[f(X_0) f(X_{|j-i|})] \\
&= N\mathbb{E}[f^2(X_0)] + \sum_{t=1}^{N-1} 2(N-i)\mathbb{E}[f(X_0)f(X_t)] \\
&= N\mathrm{Var}_\pi f + \sum_{t=1}^{N-1} 2(N-t) \sum_{i=1}^{n} \alpha_i^2 \lambda_i^t
\end{aligned}
$$

Then, it is easy to see that $\sum_{i=1}^n \alpha_i^2 = \langle f, f \rangle_\pi = \mathrm{Var}_\pi f$. Also, we have $f \perp_\pi \mathbf{1}$ and thus $\alpha_1 = 0$. So, we have

$$
\begin{aligned}
\mathrm{Var}(S_N) &\le N\mathrm{Var}_\pi f + \sum_{t=1}^{N-1} 2(N-t)\lambda_2^t \mathrm{Var}_\pi f \\
&= \left(N + \sum_{t=1}^{N-1} 2(N-t)\lambda_2^t \right) \mathrm{Var}_\pi f
\end{aligned}
$$

We know that

$$
\begin{aligned}
N &+ \sum_{t=1}^{N-1} 2(N-t)\lambda_2^t \\
&= \sum_{t=0}^{N-1} 2(N-t)\lambda_2^t - N \\
&= \frac{2}{(1-\lambda_2)^2}(\lambda_2^{N+1} - (N+1)\lambda_2 + N) - N \quad \text{by a lot of calc} \\
&\le \frac{2}{(1-\lambda_2)^2}((1-(1-\lambda_2))^{N+1} + N(1-\lambda_2) - \lambda_2) \\
&\le \frac{2}{(1-\lambda_2)^2}((1-(1-\lambda_2))^{N+1} + N(1-\lambda_2)) \\
&\le \frac{2}{(1-\lambda_2)^2}(e^{-N(1-\lambda_2)} + N(1-\lambda_2)) \\
&= \frac{2}{\gamma^2}(e^{-N\gamma} + N\gamma)
\end{aligned}
$$

Since $\mathrm{Var}\hat{f} = \frac{1}{N^2}\mathrm{Var}(S_N)$, so we have

$$\mathrm{Var}\hat{f} \le \frac{2}{(N\gamma)^2}(e^{-N\gamma} + N\gamma)\mathrm{Var}_\pi f \qquad \square$$

Remark 2.1. *Note that when x is small, we have $\alpha(x) \simeq \frac{2}{x}$, then we have*

$$\mathrm{Var}\hat{f} \le \frac{2}{N} \cdot \frac{1}{1-\lambda_2}\mathrm{Var}_\pi f$$

. So, by setting $N' = 2N(1-\lambda_2)^{-1}$, we get

$$\mathrm{Var}\hat{f} \le \frac{1}{N}\mathrm{Var}_\pi f$$

, which is the same effect as we sample $X_1, X_2, \cdots X_N$ i.i.d. variables.

On the other hand, it turns out that we have

$$\tau(\varepsilon) \le \frac{1}{1-\lambda_2}\log\left(\frac{1}{\varepsilon\pi_{\min}}\right)$$

Then, for example, if π is the uniform distribution on the basis of a matroid, then we could only upperbound $\frac{1}{\pi_{\min}}$ by n^r, and thus

$$\tau(\varepsilon) \le (1-\lambda_2)^{-1}(r\log n + \log\frac{1}{\varepsilon})$$

So, if N and r are at a same level, then the running time of our simulation is bounded by the mixing time. More generally, if N and $\log\frac{1}{\pi_{\min}}$ are at a same level, then the running time of our simulation is bounded by the mixing time of the chain.

Definition 2.1. *Let h_t^x be a vector, such that $h_t^x(y) = \frac{P^t(x,y)}{\pi(y)}$*

Definition 2.2 (spectral gap)**.** *We let $\gamma = 1-\lambda_2$ as* ***spectral gap****. And we let $\gamma_* = 1 - \max\{\lambda_2, |\lambda_n|\}$ as the* ***absolute spectral gap****.*

Fact 2.1. *For any $f : \Omega \to \mathbb{R}$, and time reversible P with stationary distribution π, we have*

$$\mathrm{Var}_\pi(P^t f) \le (1-\gamma_\star)^{2t}\mathrm{Var}_\pi f$$

Proof. Recall that $\mathrm{Var}_\pi(X+c) = \mathrm{Var}_\pi(X)$, so we assume $\mathbb{E}[f] = 0$ for convenience. Let $\alpha_i = \langle f, f_i\rangle_\pi$, then

$$\begin{aligned}
\mathrm{Var}_\pi(P^t f) &= \mathrm{Var}_\pi(\sum_{i=1}^n \alpha_i f_i \lambda_i^t) \\
&= \mathrm{Var}_\pi(\sum_{i=2}^n \alpha_i f_i \lambda_i^t) \\
&\le \mathrm{Var}_\pi((1-\gamma_\star)\sum_{i=2}^n \alpha_i f_i) \\
&= (1-\gamma_\star)^{2t}\mathrm{Var}_\pi f
\end{aligned} \qquad \square$$

Theorem 2.2. *If we first run P for N_0 steps (**from any distribution**), then run P for N_1 steps to generates $X_{1+N_0}, X_{2+N_0}, \cdots, X_{N_1+N_0}$ to estimate $\bar{f}$. Then we have*

$$\mathbb{E}(\hat{f}-\bar{f})^2 \leq (1+\frac{1}{\pi_{\min}}e^{-N_0\gamma_\star})\alpha(N_1\gamma)\mathrm{Var}_\pi f$$

Note that, currently, we may not have $\mathbb{E}[\hat{f}] = \bar{f}$.

Proof. Let $\rho(N) = \max_{x,y} \frac{P^N(x,y)}{\pi(y)}$ and

$$b_x = \mathbb{E}\left[(\frac{1}{N_1}\sum_{i=1}^{N_1} f(X_{i+N_0}) - \bar{f})^2 | X_{N_0} = x \right]$$

Then we have

$$\begin{aligned}
\mathbb{E}[(\hat{f}-\bar{f})^2|X_0 = x] &= \sum_y P^{N_0}(x,y)b_y \\
&\leq \rho(N_0)\sum_y \pi(y)b_y \qquad \text{recall the definition of } \rho \\
&\leq \rho(N_0)\alpha(N_1\gamma)\mathrm{Var}_\pi f \qquad \text{refer to Theorem 2.1}
\end{aligned}$$

So, it is suffice to prove that $\rho(N) \leq 1 + \frac{1}{\pi_{\min}}e^{-N\gamma_\star}$ Recall Fact 2.1, for any x we have

$$\begin{aligned}
\mathrm{Var}_\pi h_t^x = \mathrm{Var}_\pi(P^t h_0^x) &\leq (1-\gamma_\star)^{2t}\mathrm{Var}_\pi h_0^x \\
&\leq (1-\gamma_\star)^{2t}\frac{1-\pi_{\min}}{\pi_{\min}} \\
&\leq (1-\gamma_\star)^{2t}\frac{1}{\pi_{\min}}
\end{aligned}$$

Moreover, since $\mathbb{E}[h_t^x] = 1$, we have

$$\begin{aligned}
\pi(y)(\frac{P^t(x,y)}{\pi(y)} - 1)^2 &\leq \sum_y \pi(y)(\frac{P^t(x,y)}{\pi(y)} - 1)^2 \\
&= \mathrm{Var}_\pi h_t^x
\end{aligned}$$

Finlay, we have for any x, y:

$$\begin{aligned}
\pi(y)(\frac{P^t(x,y)}{\pi(y)} - 1)^2 &\leq (1-\gamma_\star)^{2t}\frac{1}{\pi_{\min}} \\
\pi^{1/2}(y)(\frac{P^t(x,y)}{\pi(y)} - 1) &\leq (1-\gamma_\star)^{t}\pi_{\min}^{-1/2} \\
\frac{P^t(x,y)}{\pi(y)} &\leq (1-\gamma_\star)^{t}\pi^{-1/2}(y)\pi_{\min}^{-1/2} + 1 \\
&\leq (1-\gamma_\star)^{t}\pi_{\min}^{-1} + 1 \\
&\leq e^{-t\gamma_\star}\pi_{\min}^{-1} + 1
\end{aligned}$$

□

References

[Ald87] David Aldous. On the markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Probability in the Engineering and Informational Sciences*, 1(1):33–46, 1987.

[Som] Some notes for inner products. [local link] [online link] .