

# 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  $\pi$  over  $\Omega$ , and a function  $f : \Omega \rightarrow \mathbb{R}$ , 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  $X_1, X_2, \dots, X_n$  according to  $\pi$ . 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.**

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

**Fact 1.2.**

$$\begin{aligned} \text{Var} \hat{f} &= \text{Var}\left(\frac{1}{n} \sum_i f(X_i)\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_i f(X_i)\right) \\ &= \frac{1}{n^2} \sum_i \text{Var} f(X_i) \quad \text{by independence} \\ &= \frac{1}{n} \text{Var}_\pi f \end{aligned}$$

So, from Chebyshev inequality, we have

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

In many cases, by setting an appropriate  $t$ , 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  $\pi$  and we want to estimate  $\mathbb{E}[f]$ . Surprisingly, it was shown in [Ald87] that, instead of sampling independent  $X_i$  according to  $\pi$  using  $P$ , we could run  $P$  from stationary for  $n$  steps and get  $X_1, X_2, \dots, X_n$  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  $X_1, X_2, \dots, X_N$ , then we have*

- $\mathbb{E}[\hat{f}] = \bar{f}$
- $\text{Var}[\hat{f}] \leq \alpha(\gamma N)$ , 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 according 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  $f_1 = \mathbf{1}, f_2, \dots, f_n$ . So, we have

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

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

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

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

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

We know that

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

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

$$\text{Var} \hat{f} \leq \frac{2}{(N\gamma)^2} (e^{-N\gamma} + N\gamma) \text{Var}_\pi f \quad \square$$

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

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

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

$$\text{Var} \hat{f} \leq \frac{1}{N'} \text{Var}_\pi f$$

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

**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  $f : \Omega \rightarrow \mathbb{R}$ , and time reversible  $P$  with stationary distribution  $\pi$ , we have

$$\text{Var}_\pi(P^t f) \leq (1 - \gamma_*)^{2t} \text{Var}_\pi f$$

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

$$\begin{aligned} \text{Var}_\pi(P^t f) &= \text{Var}_\pi\left(\sum_{i=1}^n \alpha_i f_i \lambda_i^t\right) \\ &= \text{Var}_\pi\left(\sum_{i=2}^n \alpha_i f_i \lambda_i^t\right) \\ &\leq \text{Var}_\pi\left((1 - \gamma_*) \sum_{i=2}^n \alpha_i f_i\right) \\ &= (1 - \gamma_*)^{2t} \text{Var}_\pi f \quad \square \end{aligned}$$

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

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

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

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

Then we have

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

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

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

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

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

Finllay, we have for any  $x, y$ :

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

□

## 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](#)] .