Optimistic Estimation of Convergence in Markov Chains with the Average Mixing Time

Pierre Alquier

ESSEC BUSINESS SCHOOL

SciCADE, July 15, 2024, Singapore

Talk is based on the preprint :

Wolfer, G. and Alquier, P. (2024). Optimistic Estimation of Convergence in Markov Chains with the Average Mixing Time. Preprint arXiv :2402.10506.

Geoffrey WOLFER

Waseda University, Tokyo

In this talk, $(X_1, X_2, X_3 \dots)$ is an ergodic Markov chain on a finite or countable space $\mathcal{X}$, with transition kernel P and invariant probability π.

$$d(t) = \sup_{x \in \mathcal{X}} \|\delta_x P^t - \pi\|_{\mathrm{TV}}.$$

Mixing time

$$t_{\mathrm{mix}}(\xi) = \min\{t : d(t) \leq \xi\} \in \{1, 2, \dots\} \cup \{+\infty\}.$$

Multiplicative property

$$t_{\mathrm{mix}}(\xi) \leq \lceil \log_2(1/\xi) \rceil \, t_{\mathrm{mix}}(1/4).$$

$$t_{\mathrm{mix}} := t_{\mathrm{mix}}(1/4).$$

Hsu, D., Kontorovich, A., Levin, D. A., Peres, Y., Szepesvári, C. & Wolfer, G. (2019). Mixing time estimation in reversible Markov chains from a single sample path. Annals of Applied Probability.

Wolfer, G. & Kontorovich, A. (2024). Improved estimation of relaxation time in nonreversible Markov chains. Annals of Applied Probability.

Based on a trajectory $(X_1, \dots, X_n)$ on a finite $\mathcal{X}$ we can

- estimate the spectral gap γ of P : with probability $1-\delta$,

$$|\hat{\gamma} - \gamma| \leq C \sqrt{\frac{\log\left(\frac{\mathrm{card}(\mathcal{X})}{\delta}\right) \log\left(\frac{n}{\pi_* \delta}\right)}{\pi_* \gamma n}}$$

  where $\pi_* = \min_x \pi(x)$;
- estimate the relaxation time $t_{\mathrm{rel}} = 1/\gamma$:
- finally,

$$(t_{\mathrm{rel}} - 1) \log 2 \leq t_{\mathrm{mix}} \leq t_{\mathrm{rel}} \log \frac{4}{\pi_*}.$$

The mixing time is a pessimistic notion :

- $t_{\mathrm{mix}}(\xi)$ can be infinite,
- $d(t) = \sup_{x\in\mathcal{X}} \|\delta_x P^t - \pi\|_{\mathrm{TV}}$.

This motivates :

$$d^\sharp(t) = \sum_{x\in\mathcal{X}} \pi(x) \|\delta_x P^t - \pi\|_{\mathrm{TV}}.$$

Average mixing time

$$t^\sharp_{\mathrm{mix}}(\xi) = \min\{t : d^\sharp(t) \leq \xi\}.$$

Proposition 1

$$t^{\sharp}_{\mathrm{mix}}(\xi) < +\infty.$$

Proposition 2

Assume $\mathrm{card}(\mathcal{X}) = 2$. Then, for any $M > 0$, there exists a transition kernel P such that

$$\frac{t_{\mathrm{mix}}(\xi)}{t^{\sharp}_{\mathrm{mix}}(\xi)} > M.$$

Remarks :

- no multiplicative property,
- this talk : $t^{\sharp}_{\mathrm{mix}}(\xi)$ is still informative and can be estimated.

A little detour to show that the average mixing time is a useful notion : let us explore its connection to mixing coefficients.

- Mixing coefficients are general tools developped to study the convergence of general stochastic processes.
- There are many notions : α-mixing, β-mixing and φ-mixing are the most popular.

Definition : β-mixing coefficients

For two σ-fields $\mathcal{A}$ and $\mathcal{B}$,

$$\beta(\mathcal{A},\mathcal{B}) = \frac{1}{2} \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|,$$

sup over all partitions $(A_i)_{i=1}^{I}$ and $(B_j)_{j=1}^{J}$ of Ω by sets in $\mathcal{A}$ and $\mathcal{B}$ respectively ;

$$\beta(s) = \sup_{t,h,\ell} \beta(\sigma(X_t,\dots,X_{t+h}),\sigma(X_{t+h+s},\dots,X_{t+h+s+\ell})).$$

Let f be a measurable function with $\mathbb{E}_\pi(f) = 0$ and $\|f\|_\infty \leq 1$. Let n, B, s be three integers with $n = Bs$.

Theorem (Yu, 1984)

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{4s}\right) + 2(B-1)\beta(s).$$

Theorem (Mohri and Rostmizadeh, 2008)

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n} f(X_i) > \varepsilon + 2\mathfrak{Rad}_B(\mathcal{F})\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{4s}\right) + 2(B-1)\beta(s)$$

where $\mathfrak{Rad}_B(\mathcal{F})$ is the Rademacher complexity.

- Many classical processes satisfy $\beta(s) \leq \beta_0 \exp(-\beta_1 s^b)$ or $\beta(s) \leq \beta_1 / s^b$.
- In theory, it is possible to estimate $\beta(s)$ based on a trajectory $(X_1, \dots, X_n)$, see

Khaleghi, A. and Lugosi, G. (2023). Inferring the mixing properties of a stationary ergodic process from a single sample-path. IEEE Transactions on Information Theory.

- However, their procedure is hardly feasible in practice, and the convergence $\hat{\beta}(s) \xrightarrow[n\to\infty]{} \beta(s)$ can be arbitrarily slow...

Reminder

In this talk, $(X_1, X_2, X_3 \dots)$ is an ergodic Markov chain with transition kernel P and invariant probability π.

Theorem (Davydov, 1973)

Assume the chain has initial distribution μ.

$$\beta(s) = \sup_t \sum_{x \in \mathcal{X}} \mu P^t(x) \| \delta_x P^s - \mu P^{t+s} \|_{\mathrm{TV}},$$

$$\varphi(s) = \sup_t \sup_{x \in \mathcal{X}} \| \delta_x P^s - \mu P^{t+s} \|_{\mathrm{TV}}.$$

Thus if the chain is stationary ($\mu = \pi$),

$$\beta(s) = \sum_x \pi(x) \| \delta_x P^s - \pi \|_{\mathrm{TV}} = d^\sharp(s),$$

$$\varphi(s) = \sup_x \| \delta_x P^s - \pi \|_{\mathrm{TV}} = d(s).$$

Corollary

When the chain is stationary,

$$t^{\sharp}_{\mathrm{mix}}(\xi) = \min\{t : \beta(t) \leq \xi\}.$$

Corollary

Assume $\beta(s) \leq \beta_1/s^b$. Then

$$t^{\sharp}_{\mathrm{mix}}(\xi) \leq \left\lceil (\beta_1/\xi)^{\frac{1}{b}} \right\rceil.$$

For any $\varepsilon, \delta > 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) > \varepsilon\right) \leq \delta$$

as soon as

$$n \geq C\left(\frac{8}{\varepsilon^2}\log\frac{4}{\delta}\right)^{\frac{b+1}{b}} \left(2t^{\sharp}_{\mathrm{mix}}(\delta)\right)^{\frac{1}{b}}.$$

Estimation of $\beta(t)$ and $t^{\sharp}_{\mathrm{mix}}(\xi)$ for stationary, erdogic Markov chains from $(X_1, \ldots, X_n)$.

$$\beta(s) = \sum_x \pi(x) \|\delta_x P^s - \pi\|_{\mathrm{TV}}$$
$$= \frac{1}{2} \sum_{x,y} \pi(x) |P^s(x,y) - \pi(y)|$$

- Obvious empirical estimates for $\hat{\pi}(x)$ and $\widehat{P^s}(x,y)$.
- Leads to an estimator $\hat{\beta}(s)$ of $\beta(s)$ (see the paper for the exact formula).

Theorem

For any $p \geq 1$,

$$\mathbb{E}|\hat{\beta}(s) - \beta(s)| \leq 4\sqrt{\frac{s\mathcal{B}_p^{(s)}\mathcal{I}_p^{(s)}}{n-s-1}}, \text{ where}$$

$$\mathcal{B}_p^{(s)} = \sum_{t=0}^{\infty} \beta(st)^{\frac{1}{p}} \text{ and } \mathcal{I}_p^{(s)} = \sum_{x,y\in\mathcal{X}} [\pi(x)P^s(x,y)]^{1-\frac{1}{p}}.$$

- The term $\mathcal{B}_p^{(s)}$ is bounded explicitly and finite under the standard assumptions $\beta(s) \leq \beta_0 \exp(-\beta_1 s^b)$ or $\beta(s) \leq \beta_1/s^b$, with conditions on p and b.
- Compare $\mathcal{I}_p^{(s)}$ to $\sum_y \sqrt{P(x,y)}$ that already appears when bounding $\|\widehat{P} - P\|_\infty$:

Wolfer, G. (2024). Empirical and Instance-Dependent Estimation of Markov Chain and Mixing Time. Scandinavian Journal of Statistics.

Study of $\mathcal{I}_p^{(s)} = \sum_{x,y\in\mathcal{X}} [\pi(x)P^s(x,y)]^{1-\frac{1}{p}}$.

- Case $s = 0$:
$$\mathcal{I}_p^{(0)} = \sum_{x\in\mathcal{X}} [\pi(x)]^{1-\frac{1}{p}}.$$
- When $s \to \infty$, $P^s(x,y) \to \pi(y)$. Thus, at least when $\mathcal{X}$ is finite, we expect :

$$\mathcal{I}_p^{(s)} \xrightarrow[s\to\infty]{} \sum_{x,y\in\mathcal{X}} [\pi(x)\pi(y)]^{1-\frac{1}{p}} \leq \sum_{x\in\mathcal{X}} [\pi(x)]^{2\left(1-\frac{1}{p}\right)}.$$

Theorem

Assume :

- $\mathcal{X}$ is finite, then $\mathcal{I}_p^{(s)} \leq [\mathrm{card}(\mathcal{X})]^{2/p}$.
- the chain is V-geometrically ergodic, then, under suitable moment conditions on $1/V$ and π, then

$$\sup_{s \in \mathbb{N}} \mathcal{I}_p^{(s)} < +\infty.$$

- various weak conditions : $\mathcal{I}_p^{(s)}$ sub-linear in s.

Estimation of $\beta(t)$ and $t^{\sharp}_{\mathrm{mix}}(\xi)$ for stationary, uniformly erdogic Markov chains from $(X_1, \dots, X_n)$.

Theorem

For any $p \geq 1$,

$$\mathbb{E}|\hat{\beta}(s) - \beta(s)| \leq 8\sqrt{\frac{\left(s + \frac{t_{\mathrm{mix}} p}{\log(2)}\right)\mathcal{I}_p^{(s)}}{n-1}}.$$

In this case, we can say more...

Theorem

Let $p > 1$ and assume $\sup_s \mathcal{I}_p^{(s)} = \mathcal{I}_p < +\infty$. Put $\tilde{\beta}(s) = \hat{\beta}(s) 1_{\{s \leq S\}}$ for some $S = S(n, \varepsilon, \delta, p, \mathcal{I}_p)$. Then, as soon as

$$n \gtrsim \frac{t_{\mathrm{mix}} p \log \frac{1}{\varepsilon}}{\varepsilon^2} \max \left[\mathcal{I}_p, \log \left(\frac{t_{\mathrm{mix}} p \log \frac{1}{\varepsilon}}{\varepsilon^2} \right) \right]$$

we have, with probability $1 - \delta$,

$$\sup_s |\tilde{\beta}(s) - \beta(s)| \leq \varepsilon.$$

Put $\hat{t}^\sharp_{\mathrm{mix}}(\xi) = \min\{s : \tilde{\beta}(s) \leq \xi\}$. Then, for n (explicitly) large enough,

$$\hat{t}^\sharp_{\mathrm{mix}}(\xi) \in \left[t^\sharp_{\mathrm{mix}}(\xi(1+\varepsilon)), t^\sharp_{\mathrm{mix}}(\xi(1-\varepsilon)) \right].$$

Thank you !