Goals The learning goal of my class project is to understand how ideas from information theory characterize the long-run behavior of Markov processes. The use of entropy and relative entropy is closely related to the theory of stochastic processes. Chapter 4 and Section 16.8 of our textbook [CT05] have some great examples that reveal this close connection between information theory and random processes. This project focuses on Markov chains, and looks at how the ideas of information theory help characterize the rate at which chains converge to equilibrium. Through the project I can gain an in-depth understanding about relative entropy as a statistical distance, and learn some interesting information inequalities as well.

The most important questions I would like to answer through my project include:

- What does information-theoretic concepts evaluate in a Markov process? Why do people sometimes choose relative entropy as the measure of distance for a finite irreducible Markov chain from stationarity?
- Why do people define the entropy in (3)[1] differently from the Shannon entropy? Is it a more widely used concept in probability and statistics?
- How do information theory concept (entropy) connect and compare to pure probability concept (variance) in the analysis of mixing times? Is there a deeper underlying motivation for people to use entropy in the analysis of mixing, or is it just for mathematical convenience in the analysis?

All these questions will be discussed at the end of the report.

Background/Context The mixing time of Markov chains is a highly active area of research at the intersection of probability theory, statistical physics, and computer science. Our discussion will focus on continuous-time Markov chains[2], and we will stick with the row-stochastic convention in this report. We assume that the reader is familiar with everything covered in the STAT 157 class this semester, and some very elementary knowledge of Hilbert spaces.

The continuous-time Markov chain (CTMC) is a simple extension of the discrete-time chain: we now have an independent random time $T_j \sim \text{Exponential}(r)$ for each transition j to take place (so that the number of transitions per unit time follows $\text{Poisson}(r)$). In our exposition we fix $r = 1$.

Let $\mathcal{X}$ be the state space and P be the transition matrix (as in discrete time). Define the *heat kernel* H_t of a CTMC by

$$\begin{aligned} H_t(x, y) &= \mathbf{P}(X_t = y \mid X_0 = x) \\ &= \sum_{k=0}^{\infty} \mathbf{P}(k \text{ transitions before time } t)\, P^k(x, y) \\ &= \sum_{k=0}^{\infty} \frac{e^{-t} t^k}{k!} P^k(x, y). \end{aligned}$$

Therefore $H_t = e^{t(P-I)}$.

We state the continuous-time version of the convergence theorem.

[LPW17, Theorem 20.1]. *Let P an irreducible transition matrix, and let H_t be its heat kernel. Define $d(t) = \max_{x \in \mathcal{X}} \| H_t(x, \cdot) - \pi \|_{\text{TV}}$. Then there exists a unique distribution π such that $\pi H_t = \pi$ for all $t \geq 0$ and*

$$d(t) \to 0 \quad \text{as } t \to \infty. \tag{1}$$

[1] If we use the term "entropy" without prefix, it will always mean the entropy defined in (3).

[2] This setup makes the theory much easier to present. For discussions of the case for discrete–time chains, see for example [Sal97, Section 1.3.1] and [BT06].

From now on we will assume P to be irreducible. We remark that $\|H_t(x,\cdot)-\pi\|_{\mathrm{TV}}$ is monotonically decreasing in t. To characterize the rate of mixing in (1), one could ask for any $\epsilon\in(0,1)$, how

$$t_{\mathrm{mix}}(\epsilon):=\inf\{t\geq 0: d(t)\leq\epsilon\}$$

depends on the time t.

Here we recall from [CT05, Section 4.4] that in the discrete case, $D_{\mathrm{KL}}(H_t(x,\cdot)\,\|\,\pi)$ is monotonically decreasing in t as well, and it converges to 0 as $t\to\infty$. We can prove the same result in the continuous case by the same reasoning; notice that $H_{t+s}=H_tH_s$.

We also remark that $D_{\mathrm{KL}}(H_t(x,\cdot)\,\|\,\pi)\to 0$ directly implies (1) by Pinsker's inequality

$$\|\mu-\nu\|_{\mathrm{TV}}\leq\sqrt{\frac{1}{2}D_{\mathrm{KL}}(\mu\,\|\,\nu)}, \tag{2}$$

which appeared as Lemma 11.6.1 in [CT05].[3]

The relative entropy technique is a well-known method used to bound the mixing time of certain Markov chains. It is often compared to the spectral gap technique, which is similar in form, but the relative entropy technique provides better bounds on the mixing time. From (2) we know we may bound the relative entropy between $H_t(x,\cdot)$ and π to upper bound t_{mix}.

Results/Effort Given a random variable X and a convex function φ, consider the expression

$$\mathbf{E}\varphi(X)-\varphi(\mathbf{E}X),$$

which by Jensen's inequality is nonnegative. If we choose $\varphi(x)=x^2$, then the expression is exactly $\mathrm{Var}(X)$.

Now choose $\varphi(x)=x\log x$, and we define the *entropy*[4] of a *nonnegative* random variable X to be

$$\operatorname{Ent}X=\mathbf{E}(X\log X)-(\mathbf{E}X)\log(\mathbf{E}X)=\mathbf{E}\left(X\log\left(\frac{X}{\mathbf{E}X}\right)\right)=\mathbf{E}\big(D_\varphi(X\,\|\,\mathbf{E}X)\big), \tag{3}$$

where D_φ is the Bregman divergence associated to the continuously differentiable strictly convex function φ[5]. In comparison, $\mathrm{Var}(X)=\mathbf{E}(X-\mathbf{E}X)^2$. It is also worth emphasizing that the entropy functional is homogeneous, i.e., $\operatorname{Ent}(aX)=a\operatorname{Ent}X$, in contrast to $\mathrm{Var}(aX)=a^2\mathrm{Var}(X)$.

If we replace X by $\frac{d\mu}{d\nu}$, the Radon-Nikodym derivative of μ with respect to ν, and take the expectations in (3) with respect to ν, then

$$\operatorname{Ent}_\nu\left(\frac{d\mu}{d\nu}\right)=\mathbf{E}_\nu\left(\frac{d\mu}{d\nu}\log\left(\frac{d\mu}{d\nu}\right)\right)=\int_{\mathcal{X}}\log\left(\frac{d\mu}{d\nu}\right)d\mu=D_{\mathrm{KL}}(\mu\,\|\,\nu), \tag{4}$$

exactly the relative entropy of μ against ν.

When initialized at state x, let $p_x^{(t)}(\cdot)=H_t(x,\cdot)$ be the more compact notation, and $\pi(\cdot)$ be the stationary distribution. By (4) we have

$$\operatorname{Ent}_\pi\left(\frac{p_x^{(t)}}{\pi}\right)=D_{\mathrm{KL}}(p_x^{(t)}\,\|\,\pi), \tag{5}$$

[3] We will use μ and ν to refer to two probability measures defined on the same measurable space, such that μ is absolutely continuous with respect to ν, throughout the report. Feel free to treat them as probability mass and density functions. Note when $\mu\not\ll\nu$, $D_{\mathrm{KL}}(\mu\,\|\,\nu)$ is defined to be $+\infty$. We ignore the singularity case in our discussion.

[4] This is clearly not the Shannon entropy. Sometimes it is called the concentration entropy because it is connected to the concentration of measure inequalities in probability and statistics. We will briefly touch on this at the end of the report.

[5] See Wikipedia for a brief introduction; it measures the differences between two points, but is not a metric. Given $\varphi=x\log x$, we compute that for two points $a,b\in\mathbf{R}^+$, $D_\varphi(a\,\|\,b)=a(\log a-\log b)-(a-b)$.

and therefore if we could bound $\mathrm{Ent}_\pi\left(\frac{p_x^{(0)}H_t}{\pi}\right)$*, then this will give us a bound on* t_{mix}. The analysis of convergence rate has become a purely analytic one.

We now look at the Hilbert space $\ell^2(\pi)$ with inner product $\langle f, g\rangle = \sum_{x\in\mathcal{X}} f(x)g(x)\pi(x)$. For any function $f:\mathcal{X}\to\mathbf{R}$, we can define the linear operators P and H_t[6] on f by

$$Pf(x) = \sum_{y\in\mathcal{X}} P(x,y)f(y) \quad \text{and} \quad H_t f(x) = \sum_{y\in\mathcal{X}} H_t(x,y)f(y). \tag{6}$$

Define the *Dirichlet form* of a transition matrix P by

$$\mathcal{E}_P(f,g) = \langle f, (I-P)g\rangle.$$

Define P^* to be the time reversal of P, i.e., let $P^*(x,y) = \frac{\pi(y)}{\pi(x)}P(y,x)$. One may easily check P^* is indeed the adjoint of P, and thus $\mathcal{E}_{P^*}(f,g) = \mathcal{E}_P(f,g)$. Also define H_t^* to be the time reversal (adjoint) of H_t with $H_t^*(x,y) = \frac{\pi(y)}{\pi(x)}H_t(y,x)$. It is elementary to check that

$$\frac{p_x^{(0)}H_t}{\pi} = H_t^*\left(\frac{p_x^{(0)}}{\pi}\right). \tag{7}$$

Why the Dirichlet form? One way to understand is to note that $L := I - P$ is the Laplacian matrix of the transition matrix P.[7] The Dirichlet form characterizes the decay of functionals as time progresses, as we will see below.

Lemma 1. *For a given function* $\varphi:\mathcal{X}\to\mathbf{R}^+$*, define* $\varphi_t = \frac{\varphi H_t}{\pi} = H_t^*\left(\frac{\varphi}{\pi}\right)$*. Then*

$$\frac{d}{dt}\mathrm{Ent}_\pi\,\varphi_t = -\mathcal{E}_P(\varphi_t, \log\varphi_t).$$

Proof. Since $H_t = e^{-tL}$, we have $H_t^* = e^{-tL^*}$. By the definition in (6) this gives $\frac{d}{dt}\varphi_t = -L^*\varphi_t$. We may now compute

$$\begin{aligned}
\frac{d}{dt}\mathrm{Ent}_\pi\,\varphi_t &= \frac{d}{dt}\left[\mathbf{E}_\pi(\varphi_t\log\varphi_t)\right] \quad (\text{since } \mathbf{E}_\pi\varphi_t = \mathbf{E}_\pi\varphi \text{ is constant})\\
&= \sum_x \pi(x)\left(\left[\frac{d}{dt}\varphi_t(x)\right]\cdot\log\varphi_t(x) + \varphi_t(x)\cdot\frac{1}{\varphi_t(x)}\cdot\left[\frac{d}{dt}\varphi_t(x)\right]\right)\\
&= \sum_x \pi(x)\cdot\left[\frac{d}{dt}\varphi_t(x)\right]\cdot\log\varphi_t \quad (\text{since } \frac{d}{dt}\mathbf{E}_\pi\varphi_t = 0)\\
&= \langle -L^*\varphi_t, \log\varphi_t\rangle\\
&= \langle \varphi_t, -L(\log\varphi_t)\rangle = -\mathcal{E}_P(\varphi_t,\log\varphi_t),
\end{aligned}$$

as desired. □

With Lemma 1 in mind, we may define the *modified log-Sobolev constant* of a given P by

$$\rho = \inf_{\mathrm{Ent}_\pi\,\varphi\neq 0} \frac{\mathcal{E}_P(\varphi,\log\varphi)}{\mathrm{Ent}_\pi\,\varphi}, \tag{8}$$

[6] This operator H_t is known as the Markov semigroup in more advanced literature.

[7] Another way is to interpret $I - P$ as the negative *generator* of the continuous Markov process.

which is minimal decay speed for the $\mathrm{Ent}_\pi\,\varphi$ over all φ. Lemma 1 should then give us

$$\frac{d}{dt}\,\mathrm{Ent}_\pi\,\varphi_t \le -\rho\,\mathrm{Ent}_\pi\,\varphi,$$

which implies

$$\mathrm{Ent}_\pi\,\varphi_t \le \exp(-\rho t)\,\mathrm{Ent}_\pi\,\varphi,$$

a nice exponential decay in the entropy. Now for any $\delta > 0$, let φ be $p_x^{(\delta)}$, which is strictly positive at each state. Then by (5) and (7) the above becomes

$$D_{\mathrm{KL}}(p_x^{(t)} \,\|\, \pi) \le \exp\big(-\rho(t-\delta)\big) D_{\mathrm{KL}}(p_x^{(\delta)} \,\|\, \pi).$$

Here we take $\delta \to 0^+$, and by continuity we may conclude

$$D_{\mathrm{KL}}(p_x^{(t)} \,\|\, \pi) \le \exp(-\rho t) D_{\mathrm{KL}}(p_x^{(0)} \,\|\, \pi) = \exp(-\rho t)\log\Big(\frac{1}{\pi(x)}\Big). \tag{9}$$

(Note $p_x^{(0)} = 1$ only at x and $= 0$ everywhere else.) Recall our aim is to find the smallest t such that $\|p_x^{(t)} - \pi\|_{\mathrm{TV}} \le \epsilon$. By (2) it suffices to let

$$D_{\mathrm{KL}}(p_x^{(t)} \,\|\, \pi) \le 2\epsilon^2,$$

and by (9) it suffices to further let

$$\exp(-\rho t)\log\Big(\frac{1}{\pi(x)}\Big) \le 2\epsilon^2,$$

i.e.,

$$t \ge \frac{1}{\rho}\left(\log\frac{1}{2\epsilon^2} + \log\log\frac{1}{\pi(x)}\right).$$

Hence

$$t_{\mathrm{mix}}(\epsilon) \le \frac{1}{\rho}\left(\log\frac{1}{2\epsilon^2} + \log\log\frac{1}{\min_{x\in\mathcal{X}}\pi(x)}\right). \tag{10}$$

Our discussion of the entropy method above is similar to the ones given in [Sin23] and [BT06]. The iterated logarithmic dependence on the stationary distribution π is a highly desirable result. We remarked at the beginning about the similarity between the variance and the entropy functionals. In fact our deductions above also work for $\mathrm{Var}_\pi\big(\frac{p_x^{(t)}}{\pi}\big)$. We first note that by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|p_x^{(t)} - \pi\|_{\mathrm{TV}}^2 &= \frac{1}{4}\Big(\sum_{y\in\mathcal{X}}\big|p_x^{(t)}(y) - \pi(y)\big|\Big)^2 \\
&= \frac{1}{4}\Big(\sum_{y\in\mathcal{X}}\pi(y)\Big|\frac{p_x^{(t)}(y)}{\pi(y)} - 1\Big|\Big)^2 \\
&\le \frac{1}{4}\cdot 1^2 \cdot \left(\frac{p_x^{(t)}(y)}{\pi(y)} - 1\right)^2 = \frac{1}{4}\,\mathrm{Var}_\pi\left(\frac{p_x^{(t)}}{\pi}\right),
\end{aligned} \tag{11}$$

i.e., $\|p_x^{(t)} - \pi\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{\mathrm{Var}_\pi\big(\frac{p_x^{(t)}}{\pi}\big)}$. This reduces bounding the total variance distance to bounding the variance functional. Compared to Lemma 1, we have the following well-known result (which is also the key reason behind the important definition of Dirichlet form):

Lemma 2. *For any function* $\varphi: \mathcal{X} \to \mathbf{R}$, $\frac{d}{dt} \operatorname{Var}_\pi(H_t\varphi) = -2\mathcal{E}_P(H_t\varphi, H_t\varphi)$.

If we define the *Poincaré constant*[8]

$$\alpha = \inf_{\text{nonconstant } \varphi} \frac{\mathcal{E}_P(\varphi,\varphi)}{\operatorname{Var}_\pi \varphi}, \tag{12}$$

then one have

$$\operatorname{Var}_\pi(H_t\varphi) \le \exp(-2\alpha t) \operatorname{Var}_\pi(\varphi). \tag{13}$$

Since $\mathcal{E}_P(\varphi,\varphi) = \mathcal{E}_{P^*}(\varphi,\varphi)$, the Poincaré constant is the same for P and P^*. It follows from (11) and (13) that

$$4\|p_x^{(t)} - \pi\|_{\mathrm{TV}}^2 \le \operatorname{Var}_\pi\left(\frac{p_x^{(t)}}{\pi}\right) = \operatorname{Var}\left(H^{*t}\left(\frac{p_x^{(0)}}{\pi}\right)\right) \le \exp(-2\alpha t) \operatorname{Var}_\pi\left(\frac{p_x^{(0)}}{\pi}\right) < \exp(-2\alpha t) \cdot \frac{1}{\pi(x)}.$$

One can thus get

$$t_{\mathrm{mix}}(\epsilon) \le \frac{1}{2\alpha}\left(\log\frac{1}{4\epsilon^2} + \log\frac{1}{\min_{x\in\mathcal{X}} \pi(x)}\right). \tag{14}$$

We remark that the variance technique we described above is a more basic result in the literature. Indeed it is clear that that the entropy method (10) has a sharper dependence on π compared to the variance method (14). This is crucial in applications. [BT06] and [Goe04] contain references to some applications of the entropy methods to random walks on graphs and on groups, where bounding the modified log-Sobolev constant is relatively easy.

Discussion and conclusions We have presented two important techniques in the mixing time literature in a self-contained manner above. Apart from the modified log-Sobolev constant technique, there is a *standard* log-Sobolev constant technique involving relative entropy that can be found in [Sal97]. However, it is less useful in practice compared to the modified one, and we omit it in our report.

We provide our answers to the questions at the start. First, the relative entropy, along with the total variation distance, both evaluates the convergence rate of the time-t distribution $H_t(x,\cdot)$ to the stationary measure π. We remark that the use of total variation distance in characterizing the rate of mixing is standard in the literature. However, alternative distances such as the relative entropy and the ℓ^p distances are also considered. In fact we have a list of inequalities that can relate all these distances; see [Sal97, Section 2.4].

The entropy rate discussed in Chapter 4 and Section 16.8 of [CT05] are important results, but entropy rate is defined only for discrete-time chains. Therefore it has limited applications.

Shannon entropy works nicely for discrete distributions, but does not generalize well to continuous distributions. As we have seen in class, the differential entropy

$$h(X) = -\int_{\mathcal{X}} f(x)\log f(x)\,dx$$

for a random variable X with density f (see [CT05, Chapter 8]), although preserving some useful properties of the discrete Shannon entropy, is no longer an intrinsic measure of uncertainty of X. Also the density is undefined for certain distributions. Compared to the Shannon entropy and differential entropy, relative entropy/KL divergence appears much more often in the probability and statistics literature. It is measure-theoretically always defined, and thus has the generality that probabilists/statisticians desire.

It is true that the techniques we have discussed are analytic, and I have tried my best to provide intuition behind the techniques. In the literature the variance and entropy techniques are referred to as

[8] This is also known as *the spectral gap* when P is reversible, because in this case $\alpha = 1 -$ the second largest eigenvalue of P, which is indeed a surprising fact.

"analytic tools," in contrast to probabilistic techniques such as coupling and strong stationary times (see [LPW17]). To summarize, the convergence to stationarity in total variation can be reduced to the decay in entropy/variance of $\frac{p_x^{(t)}}{\pi}$. The (modified) log-Sobolev constant ρ and the Poincaré constant α that we defined in (8) and (12) are respectively the smallest constants that satisfy the (modified) log-Sobolev inequality and the Poincaré inequality. These inequalities characterize the behavior of distributions studied in high-dimensional probability and Markov chains (see [Han14] and [BGL14]).

Despite the techniques being very analytic, we emphasize that it is the "$x \log x$" form that appears in the quantities of information theory that ensures nice properties (including nonnegativity and homogeneity) of our entropy functional. In particular, the entropy closely resembles the variance, while the entropy turns out to be a stronger tool than variance in the analysis of Markov chains mixing times. We refer the readers back to the start of Results/Effort for the comparison between $\mathrm{Ent}(\cdot)$ and $\mathrm{Var}(\cdot)$. These analytic tools are powerful and somewhat abstract, but they arise from elementary notions that characterize the deviation of a random variable from its mean.

Moving forward, [Han14], [Wai19], and [Tro22] discusses the importance of variance and entropy functionals in high dimensional probability, another key area of my interest. These sources inspired my exposition of the entropy definition at the beginning. The mixing time literature I have surveyed provides little motivation to why variance and entropy are used, and why entropy is defined as in (3). [Han14, Chapter 2 & 3] establishes the connection between the high dimensional probability and mixing time literature at an extended level.

For the analysis of Markov operators and their functional inequalities (such as the the modified log-Sobolev inequality and the Poincaré inequality), see [BGL14]. For further studies of modern analytic tools in the mixing time literature, see [MT06].

My semester project is a continuation of STAT 157 last semester, where I first learned about the theory of mixing times and became very interested in it. However, we did not have time to learn about the spectral and analytic techniques beyond the probabilistic techniques. Later I heard about the use of entropy in the analysis of certain Markov chains and learned that it is Pinsker's inequality that relates the total variation distance to the relative entropy. This is a fairly challenging project for me in a semester's time, but it turns out to be extremely worthwhile. I learned tremendously about information theory and tied it to my areas of interest in probability theory, and I improved my skills at source-gathering and self-study as well.

Later in graduate school I would also be happy to explore large deviation theory (where relative entropy is extremely important, for example in Sanov's theorem [CT05, Chapter 11]), which I believe is a bit too technical for me at the moment and for such a project.

References

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing, 2014.

[BT06] Sergey G. Bobkov and Prasad Tetali. "Modified Logarithmic Sobolev Inequalities in Discrete Settings". *Journal of Theoretical Probability* 19.2 (2006), pp. 289–336.

[CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, John Wiley & Sons, 2005.

[Goe04] Sharad Goel. "Modified logarithmic Sobolev inequalities for some models of random walk". *Stochastic Processes and their Applications* 114.1 (2004), pp. 51–79.

[Han14] Ramon van Handel. "APC550 Lecture Notes: Probability in High Dimension". 2014.

[LPW17] David Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2017.

[MT06] Ravi Montenegro and Prasad Tetali. "Mathematical aspects of mixing times in Markov chains". *Foundations and Trends in Theoretical Computer Science* 1.3 (2006), pp. 237–354.

[Sal97] Laurent Saloff-Coste. "Lectures on finite Markov chains". *Lectures on Probability Theory and Statistics*. Springer Berlin Heidelberg, 1997, pp. 301–413.

[Sin23] Alistair Sinclair. "Lecture Notes for CS294 Partition Functions". 2023.

[Tro22] Joel A. Tropp. "ACM 217: Probability in High Dimensions". 2022.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.