TCS@UCSB Summer School on Spectral Independence **August 8-12, 2022**

Lecture 1: *August 8th, 2022*

Lecturer: *Pietro Caputo*

Markov Chain Fundamentals

These notes were prepared by Amanda Priestley based on the lecture by Pietro Caputo. This is part of the 2022 Summer School on *New tools for optimal mixing of Markov chains: Spectral independence and entropy decay*, which was held at the University of California, Santa Barbara (UCSB) from August 8, 2022 to August 12, 2022. More information on the summer school is available at: https://sites.cs.ucsb.edu/~vigoda/School/

1.1 Introduction

In this lecture, Professor Caputo introduces the models and definitions relevant for the rest of the summer school. This includes an introduction to *relative entropy*, and how it is used in the context of Markov chains.

1.2 Relative Entropy

Let Ω be a finite set and let $P(\Omega)$ denote the collection of probability measures on Ω. For two measures $\nu, \mu \in P(\Omega)$ the *relative entropy* or *Kullback–Leibler (KL) divergence* is defined as follows:

$$H(\nu \mid \mu) := \nu\left[\log\left(\frac{\nu}{\mu}\right)\right] = \sum_{\sigma\in\Omega} \nu(\sigma)\log\left(\frac{\nu(\sigma)}{\mu(\sigma)}\right) \tag{1.1}$$

Note, by convention: $0\log 0 = 0\log\frac{0}{0} = 0$. Furthermore, $H(\nu \mid \mu) = \infty$ when there exists some σ such that $\mu(\sigma) = 0$ and $\nu(\sigma) \neq 0$ (i.e, when ν is not absolutely continuous with respect to μ).

1.2.1 Facts

Here are some elementary facts regarding relative entropy which demonstrate it is a reasonable notion of distance between probability measures:

1. **Non-negative**: $H(\nu \mid \mu) \geq 0$ for all $\nu, \mu \in P(\Omega)$, and $H(\nu \mid \mu) = 0 \iff \mu = \nu$.
2. **Convexity**: For $\nu_i \in P(\Omega)$, $\alpha_i \geq 0$, and $\sum_i \alpha_i = 1$
$$H\left(\sum_i \alpha_i \nu_i \mid \mu\right) \leq \sum_i \alpha_i H(\nu_i \mid \mu).$$
3. **Pinsker's Inequality**: This provides the following bound on the total variation distance in terms of the relative entropy:
$$\|\nu - \mu\|_{\mathrm{TV}}^2 \leq \frac{1}{2} H(\nu \mid \mu).$$
Recall, $\|\nu - \mu\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega} |\nu(x) - \mu(x)|$.

4. **Functional form**:
$$H(\nu \mid \mu) = \sup_{g:\Omega\to\mathbb{R}} \{\ \nu[g] - \log\left(\mu[e^g]\right)\ \},$$
where $\nu[g] = \sum_{x\in\Omega} \nu(x)g(x)$ and $\mu[e^g] = \sum_{x\in\Omega} \mu(x)\exp(g(x))$.

Note, relative entropy is not symmetric and does not necessarily satisfy the triangle inequality, and hence it is not a metric.

Proof of Fact 1. Letting $h \coloneqq \frac{\nu}{\mu}$ we have that $H(\nu \mid \mu) = \mu\left[h \log h\right]$. Note that $\mu[h] = \sum_{x\in\Omega} \mu(x)h(x) = \sum_x \nu(x) = 1$. Since $h \log h$ is convex, applying Jensen's Inequality we have the following:
$$H(\nu \mid \mu) = \mu\left[h \log h\right] \geq \mu\left[h\right] \log(\mu\left[h\right]) = \log 1 = 0. \tag{1.2}$$
Equality holds in (1.2) iff the function h is constant. Since ν and μ are probability distributions and hence sum to 1, then h is constant iff $\nu = \mu$. □

Proof of Fact 2. Notice that we are taking a convex combination of the ν_i and hence the result follows by applying Jensen's inequality. □

Proof of Fact 3. Left as an exercise. □

Proof of Fact 4. Once again let $h = \nu/\mu$ and consider $g = \log(h)$. As before, we have $\mu[\exp(g)] = \sum_x \nu(x) = 1$ and hence $\log(\mu[\exp(g)]) = 0$. Therefore, $H(\nu \mid \mu) \geq \nu[\log(h)]$. To complete the proof we want to show that this choice of g achieves the maximum in (4). Hence, it suffices to show that for any g:
$$\nu\left[\log h\right] \geq \nu[g] - \log(\mu[\exp(g)]).$$
Note,
$$\mu\left[h \log(\exp(g)/h))\right] = \nu[\log(\exp(g)/h)] = \nu[g] - \nu[\log h]. \tag{1.3}$$
Notice that
$$\mu\left[h \log\left(\frac{e^g}{h}\right)\right] = \nu\left[\log\left(\frac{e^g}{h}\right)\right] \leq \log\left(\nu\left[\frac{e^g}{h}\right]\right) = \log\left(\mu\left[e^g\right]\right), \tag{1.4}$$
where the inequality follows by Jensen's inequality. Combining these two calculations we have:
$$\begin{aligned} \nu[\log h] &= \nu[g] - \mu\left[h \log(\exp(g)/h))\right] && \text{by (1.3)} \\ &\geq \nu[g] - \log(\mu[\exp(g)]) && \text{by (1.4)}, \end{aligned}$$
which completes the proof. □

Use in The Context of Markov Chains

By thinking of ν as the (current) distribution of a Markov chain, and μ as the target distribution (i.e., the stationary distribution), we can use the relative entropy as a natural measure of convergence of a Markov chain to its target distribution.

1.3 Models, Definitions, and Notation

1.3.1 Gibbs Measure

For a graph $G = (V, E)$, label the vertex set by $V \coloneqq [n] = \{1, \ldots, n\}$. Let $\mu \in P(\Omega)$ be a Gibbs Measure with state space $\Omega \coloneqq \Omega_1 \times \cdots \times \Omega_n$. Typically, there is a common spin space S and $\Omega_i = S$ for all $1 \leq i \leq n$. For example in the case of the Ising model defined below $S = \{+1, -1\}$, for the Potts model $S = \{1, 2, \ldots, q\}$ for an integer $q \geq 3$, and for colorings then $S = \{1, \ldots, k\}$ for an integer $k \geq 2$. Configurations of the model are assignments $\sigma : V \to S$ of spins S to the n coordinates/vertices. The following are the classical examples of spin systems on graphs.

Example: Ising Model

For a graph $G = (V, E)$, label the vertex set by $V \coloneqq [m] = \{1, \ldots, m\}$. The Ising model on G has state space $\Omega \coloneqq \{+1, -1\}^{[n]}$, where $[n] = \{1, \ldots, n\}$. For a configuration $\sigma \in \Omega$ we can view σ as a vector of spins taking value ± 1, i.e., $\sigma \coloneqq (\sigma_x, x \in [n])$. Given a parameter $\beta \in \mathbb{R}$ (corresponding to the inverse temperature of the system), the Gibbs distribution is probability measure $\mu \coloneqq \mu_{G,\beta}$ where for $\sigma \in \Omega$:

$$\mu(\sigma) = \frac{1}{Z} \exp(\beta \sum_{x,y \in E} \sigma_x \sigma_y)$$

The normalizing factor $Z = Z_{G,\beta}$, known as the partition function, implies that μ is a probability distribution. Furthermore,

In addition, notice that β being positive forces μ to favor aligned spins (spins with the same sign), and β negative favors spins with opposite orientations along every edge (spins with opposing signs). β 0 forces μ into being a product measure corresponding to independent spins.

Example: Potts Model

A generalization of the Ising model is the Potts model. Each spin can take one of q values with $\sigma_x \in [q]$ and μ is defined to be

$$\mu(\sigma) \propto \exp(\beta \sum_{x,y \in E} \mathbb{1}\{\sigma_x = \sigma_y\}).$$

By taking $q = 2$ and setting β to $\beta/2$ we retrieve back the Ising model.

Example: Coloring

Another model of interest that of is proper colorings. Again we are given a graph $G = (V, E)$, and we are also given $q \in \mathbb{Z}$ colors which we can assign to the vertices. The state space is $\Omega \coloneqq [q]^n$ with $\Omega_0 \subset \Omega$ being the subset of proper colorings defined as

$$\Omega_0 \coloneqq \{\sigma \in \Omega \ : \ \sigma_x \neq \sigma_y, \forall x \neq y \in E\}.$$

One measure of interest is taking μ to be the uniform measure on the set of all proper colorings. Note also that the coloring model can be retrieved from the Potts model by taking $\beta \to -\infty$.

Example: Permutations

We let $\Omega \coloneqq [n]^n$ and $\Omega_0 \subset \Omega$ being the set of permutations of n objects:

$$\Omega_0 \coloneqq \{\sigma \in \Omega \ : \ \sigma_x \neq \sigma_y, \forall x \neq y\}\,.$$

One measure of interest is taking μ to be the uniform measure on Ω_0.

As a general remark, note that the Ising and Potts Models are such that all configurations are given positive probability, whereas the coloring and permutation models have hard constraints, and thus have configurations with probability 0.

1.3.1.1 Boundary Conditions/Pinnings

For a graph $G = (V, E)$, for $A \subset V$, let $A^c = V \setminus A$. Given $\Omega \coloneqq \Omega_1 \times \cdots \times \Omega_n$ and a Gibbs measure μ, for a subset of the vertices A we introduce a *boundary condition* or *pinning* $\tau \coloneqq \sigma_{A^c} = \{\sigma_x, x \in A^c\}$ which is a fixing of the spins in the set A^c. Thus, we have that the conditional probability is defined to be

$$\mu_A^\tau(\eta) \coloneqq \mu^\tau(\sigma_A = \eta \mid \sigma_{A^c} = \tau)\,.$$

Furthermore, given a function $f : \Omega \to \mathbb{R}$ we use the following notation

$$\mu_A^\tau[f] \coloneqq \sum_{\sigma_A} \mu_A^\tau(\sigma_A) f(\sigma_A \circ \tau)$$

To denote the conditional expectation of the function f, where Ω_A is the set of all possible assignments of the spins in A, and $\sigma_A \circ \tau \coloneqq (\sigma_x, x \in A, \tau_y, y \in A^c)$.

In addition we use the notation

$$\mu_A[f] : \tau \to \mu_A^\tau[f]$$

for the function mapping boundary conditions to the expectation conditioned on that choice of boundary condition. Notice that taking the expectation over all choices of τ sampled according to the measure μ gives:

$$\mu[\ \mu_A[f]\ ] = \mu[f]\,.$$

As a general fact we also have that

$$\mu_B \mu_A = \mu_B \quad \forall A \subset B\,. \tag{1.5}$$

1.3.2 Functionals

Covariance

Given two functions $f, g : \Omega \to \mathbb{R}$, we have

$$\mathrm{Cov}_A^\tau(f, g) \coloneqq \mu_A^\tau[f, g] - \mu_A^\tau[f]\mu_A^\tau[g]\,.$$

Furthermore:

$$\mathrm{Cov}_A(f, g) : \tau \to \mathrm{Cov}_A^\tau(f, g).$$

In addition when $f = g$, we have

$$\mathrm{Var}_A(f) \coloneqq \mathrm{Cov}_A(f, f).$$

Entropy

For non-negative functions f, we define

$$\mathrm{Ent}_A^\tau[f] \coloneqq \mu_A^\tau[f \log f] - \mu_A^\tau[f] \log\left(\mu_A^\tau[f]\right).$$

While:

$$\mathrm{Ent}_A(f) : \tau \to \mathrm{Ent}_A^\tau[f]$$

1.3.2.1 Elementary Properties

Notice that sampling the boundary condition according to μ gives

$$\mu\left[\mathrm{Ent}_A(f)\right] = \mu\left[f \log\left(\frac{f}{\mu_A f}\right)\right]$$

Proof.

$$\begin{aligned}
\mu\left[\mathrm{Ent}_A(f)\right] &= \sum_{\tau \in \Omega_A^c} \mu(\tau)\left(\mu_A^\tau[f \log f] - \mu_A^\tau[f] \log\left(\mu_A^\tau[f]\right)\right) \\
&= \sum_{\tau \in \Omega_A^c} \mu(\tau)\left(\mu_A^\tau\left[f \log\left(\frac{f}{\mu_A^\tau[f]}\right)\right]\right) \\
&= \mu\left[\mu_A\left[f \log\left(\frac{f}{\mu_A f}\right)\right]\right]
\end{aligned} \tag{1.6}$$

Fact 1.5 yields the conclusion. □

1.3.2.2 Relation to Relative Entropy

Notice that $\mathrm{Ent}(f)$ is equivalent to the Relative Entropy. In particular when $A = [n]$, we have

$$\mu\left[\mathrm{Ent}_A(f)\right] = \mathrm{Ent}(f) = \mu\left[f \log\left(\frac{f}{\mu[f]}\right)\right] = H(\nu \mid \mu) \cdot \mu[f] \tag{1.7}$$

with $\nu \coloneqq \frac{f}{\mu[f]} \cdot \mu$.

1.4 Markov Chains

We will now look at Markov chains with stationary measure which is one of the Gibbs Measures introduced in Section 1.3.1

In particular let P be a stochastic matrix such that $P(\sigma, \eta) \geq 0$, and $\sum_{\eta \in \Omega} P(\sigma, \eta) = 1$, and assume that

(a) **Invariant:** μ is an invariant measure, i.e., $\mu P = \mu$ and for all $\eta \in \Omega$

$$\sum_\sigma \mu(\sigma) P(\sigma, \eta) = \mu(\eta)$$

(b) **Ergodicity:** There exists $k > 0$, for all $\sigma, \eta \in \Omega$,

$$P^k(\sigma, \eta) > 0.$$

Under these assumptions there exists a unique stationary distribution, namely μ, and

$$P^t(\sigma, \eta) \to \mu(\eta) \qquad \text{as } t \to \infty, \quad \forall \sigma, \eta \in \Omega.$$

The question of interest is the speed of the above convergence which is referred to as the mixing time.

1.4.1 Mixing Time

The *ε-mixing time* of a Markov chain is defined as, starting from the worst initial state σ, the minimum number of steps t so that the distribution of the chain is within total variation distance $\leq \varepsilon$ of stationarity:

$$T_{\text{mix}}(P, \varepsilon) \coloneqq \max_{\sigma \in \Omega} \min_t \{t \in \mathbb{N} : \|P^t(\sigma, \cdot) - \mu\|_{\text{TV}} \leq \varepsilon\},$$

where $P^t(\sigma, \cdot)$ is the distribution of the Markov chain after t steps and the maximum is taken over all initial conditions $\sigma \in \Omega$.

Often one considers $\varepsilon = \frac{1}{4}$. Let

$$T_{\text{mix}} \coloneqq T_{\text{mix}}(P, \frac{1}{4}).$$

The mixing time satisfies a submultiplicative property, namely for any k we have:

$$\|P^k(\sigma, \cdot) - \mu\|_{\text{TV}} \leq 2^{-\left\lfloor \frac{k}{T_{\text{mix}}} \right\rfloor}. \tag{1.8}$$

and hence the ε-mixing time decays exponentially in k/T_{mix} and the following holds for all $\varepsilon > 0$,

$$T_{\text{mix}}(P, \varepsilon) \leq T_{\text{mix}} \cdot \lceil \log(1/\epsilon) \rceil.$$

This is proved using a simple coupling argument, see, e.g., [LP17, Chapter 4.5].

1.4.2 Relationship to Entropy

If the relative entropy decays at a constant rate then that implies very fast mixing time.

Definition 1.1. *Let $\delta > 0$. For an ergodic Markov chain with transition matrix P on state space Ω and stationary distribution μ, we say that the* relative entropy decays at rate *δ if for all $\nu \in P(\Omega)$,*

$$H(\nu P \mid \mu) \leq (1 - \delta) H(\nu \mid \mu).$$

This implies the following bound on the mixing time, which will yield optimal bounds in many upcoming examples.

Lemma 1.2. *Suppose a Markov chain has relative entropy decay at rate $\delta > 0$, then the mixing time can be bounded as follows:*

$$T_{\text{mix}}(P, \varepsilon) \leq 1 + \frac{1}{\delta} \left[\log \frac{1}{2\varepsilon^2} + \log \log \frac{1}{\mu_*} \right]$$

where $\mu_* \coloneqq \min\{\mu(\sigma), \sigma \in \Omega\}$.

For hard constraint models such as colorings, in the definition of μ_* we further restrict to the subset of Ω which is positive, i.e., Ω_0. Note, for colorings then $\mu_* \geq 1/k^n$ and thus: $T_{\text{mix}} = O(\frac{\log n}{\delta})$. In contrast, if one considers contraction of variance (instead of entropy) then the analog of Lemma 1.2 has $\log(1/\mu_*)$ in place of $\log\log(1/\mu_*)$.

Proof. Let $k = 1 + \frac{1}{\delta}\left[\log\frac{1}{2\varepsilon^2} + \log\log\frac{1}{\mu_*}\right]$. The proof follows by applying Pinsker's Inequalty to $P^k(\sigma,\cdot)$, and then iteratively applying the inequality $H(\nu P \mid \mu) \leq (1-\delta)H(\nu \mid k)$. Using the notation $\nu P^k(\eta) \coloneqq \sum_\sigma \nu(\sigma)P^k(\sigma,\eta)$, we have

$$\begin{aligned}
\|\nu P^k(\sigma,\cdot) - \mu\|_{\text{TV}}^2 &\leq \frac{1}{2}H(\nu P^k \mid \mu) && \text{by Pinsker's Inequality}\\
&\leq \frac{1}{2}(1-\delta)^k H(\nu \mid \mu) && \text{by the entropy decay rate}\\
&\leq \frac{1}{2}(1-\delta)^k \frac{1}{\mu_*} && \text{since } H(\nu \mid k) \leq \log(1/\mu_*)\\
&\leq \varepsilon^2 && \text{by the choice of } k.
\end{aligned}$$

Notice that the worst case for $H(\nu \mid \mu)$ is when ν is concentrated at a single state, by convexity. Hence we can apply the bound $H(\nu \mid k) \leq \log(1/\mu_*)$ in the above proof.

□

1.4.2.1 Relation to Entropy Functional

Recall that, for $f \geq 0, \mu[f] = 1$ and $\nu = f\mu$, we have that

$$H(\nu \mid \mu) = \text{Ent} f\,.$$

Furthermore we have that

$$\nu P = g\,u$$

where $g = (P^* f)(a) = \sum_\eta P^*(\sigma,\eta)f(\eta)$, and $P^*(\sigma,\eta) = P(\eta,\sigma)\frac{\mu(\eta)}{\mu(\sigma)}$ is the adjoint of P in $L^2(\mu)$. As an aside when $P = P^*$, then P is a *reversible* Markov chain.

With this in mind we have that the following two notions of *entropy contraction* are equivalent:

$$H(\nu P \mid \mu) \leq (1-\delta)H(\nu \mid k) \iff \text{Ent}P^* f \leq (1-\delta)\text{Ent} f \quad \forall f \geq 0 \tag{1.9}$$

This statement is stronger than that given by studying the *spectral gap* which requires contraction of variance rather than entropy. However contraction of variance can be much easier to prove.

References

[LP17] D. A. Levin and Y. Peres. *Markov chains and mixing times.* American Mathematical Society, 2017.