NOTES ON
STATISTICAL LEARNING THEORY

JIECAO CHEN
JIECA001@UMN.EDU

ABSTRACT. Most results of statistical learning theory take the form of so-called error bounds. This note contains introduction and brief summary of some key ideas and techniques to obtain those result.

1. INTRODUCTION

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, which includes gaining knowledge, making prediction or constructing models from the date set. As it is put into a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena.

V. Vapnik once described why we need theories:

Nothing is more practical than a good theory.

A theory of interference should be able to give the formal definition of words such as learning, generalization, overfitting and also to characterize the performance of learning algorithms so that it would eventually be able to help us to design better learning algorithms.

Two goals: make things more precise and derive new or improved algorithms

1.1. **Learning and Inference.** The process of inductive inference:

- Observe a phenomenon
- Construct a model of that phenomenon
- Make predictions using this model

The task of Machine Learning is to actually *automate* this process and the goal of Learning Theory is to *formalize* it.

To make our discussion simpler, in this note, we only consider a special case of above process which is the supervised learning framework for pattern recognition. In this special case, data set consists of instance-label pairs, where the value of label $\in \{-1, 1\}$. The task of a learning algorithm is to construct a function which is able to predict the corresponding label given the instance. A "Good" learning algorithm will give less errors when do such prediction.

A model that can exactly fit all instance-label pairs in a training data set may not always be the best model, as we know that the data set may sometime include noise, in such case, a model fitting all the data may give a worse prediction in the unseen data, such phenomenon is usually referred to as **overfitting**. A way to overcome **overfitting** is to look for **regularities**.

If there are many models available, and all of the well fit the data, we normally follow Occam's idea to choose the simplest model, which would more likely to be generalized from the training data to future data. This immediately raises the question of how to measure and qualify simplicity of a model.

There is no universal way of measuring simplicity and the choice of a specific measure inherently depends on the problem at hand.

A formula we believe:

Generalization = Data + Knowledge

Generalization can only come when one adds specific knowledge to the data. Each learning algorithm encodes specific knowledge.

1.2. **Assumptions.** Several more precise assumptions that made by the Statistical Learning Theory Framework:

- Probabilistic model of the phenomenon (or data generation process). Within this model, the relationship between past and future observations is that they both are sampled independently from the same distribution (i.i.d).

2. Formalization

Consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$. As we only consider our specific case (binary classification), we choose $\mathcal{Y} = -1, 1$.

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are random variables follow the an unknown distribution P. We observe a sequence of i.i.d pairs (X_i, Y_i) sampled according to P and the goal is to construct a function $g : \mathcal{X} \to \mathcal{Y}$ which can predict Y from X.

A criterion is need to choose the function g. It can be chosen as $P(g(X) \neq Y)$. We thus define the *risk* of g as

$$R(g) = P(g(X) \neq Y) = \mathbb{E}[\mathbb{I}_{g(X)\neq Y}] \tag{1}$$

Note that P can be decomposed as $P_X \times P(Y|X)$. Define the *regression function* $\eta(x) = \mathcal{E}[Y|X = x] = 2\mathcal{P}[Y = 1|X = x]-1$ and the *target function* (or Bayes classifier) $t(x) = sgn\eta(x)$. This function achieves the minimum risk over all possible measurable functions:

$$R(t) = \inf_g R(g) \tag{2}$$

Denote the value $R(t)$ as R^*, called the Bayes risk.

Define the *noise level* as $s(x) = \min(\mathbb{P}[Y = 1|X = x], 1 - \mathbb{P}[Y = 1|X = x]) = (1 - \eta(x))/2$ and this gives $R^* = \mathbb{E}s(X)$.

Our goal is thus to find this function t. but since P is unknown we cannot directly measure the risk. Let's define *empirical risk*:

$$R_n(g) = \frac{1}{b}\sum_{i=1}^{n} \mathbb{I}_{g(X_i)\neq Y_i} \tag{3}$$

It is common to use this quantity as a criterion to select an estimate of t.

2.1. **Algorithms. Overfitting**. When the input space is infinite, one can always construct a function g_n which perfectly predicts the labels of the training data but behaves on the other points as the opposite of the target function t. So one would have minimum empirical risk but maximum risk.

There are essentially two ways to prevent this overfitting situation:

- restrict the class of functions in which the minimization is performed
- modify the criterion to be minimized (e.g. adding a penalty for complicated function)

Empirical Risk Minimization. This algorithm is one of the most straightforward, yet it is usually efficient:

$$g_n = arg\min_{g\in\mathcal{G}} R_n(g) \tag{4}$$

where $\mathcal{G}$ is the function class we perform minimization.

Sometime (actually in most case) the *best* function in $\mathcal{G}$ is not necessary the overall best function. So one may want to enlarge the model as much as possible, while preventing overfitting.

Structural Risk Minimization. An infinite sequence $\{\mathbf{G_d}|d = 1, 2, \ldots\}$ of models of increasing size:

$$g_n = arg\min_{g\in\mathcal{G}_d, d\in\mathcal{N}} R_n(g) + \text{pen}(d, n) \tag{5}$$

The penalty $\text{pen}(d, n)$ gives preference to models where estimation error is small and measures the size of capacity of the model.

Regularization. Another, usually easier to implement approach consists in choosing a large model $\mathcal{G}$ and define on $\mathcal{G}$ a *regularizer*, typically a norm $\|g\|$:

$$g_n = arg\min_{g\in\mathcal{G}} R_n(g) + \lambda\|g\|^2 \tag{6}$$

Let (Ω, P, π) denote Markov kernel of a finite Markov chain on a finite state space Ω with a unique invariant measure π. That is

$$P(x, y) \leq 0, \forall x, y \in \Omega, and \sum_{y\in\Omega} P(x, y) = 1, \forall x \in \Omega$$

$$\sum_{x\in\Omega} \pi(x)P(x, y) = \pi(y), \forall y \in \Omega.$$

In this summary, we assume P is irreducible and π has full support (Ω).

Minimal holding probability $\alpha \in [0, 1]$ satisfies $\forall x \in \Omega : P(x, y) \geq \alpha$

Let $k_n^x(y) = P^n(x, y)/\pi(y)$ donate the density with respect to π at time $n \geq 0$, or simply $k_n(y)$ when start rate or start distribution is unimportant or clear from the context. Then $lim_{n\to\infty}k_n^x(y) = 1$

If μ is a probability distribution on Ω, then

- Total Variation Distance: $\|\mu - \pi\|_{TV} = \frac{1}{2}\|\frac{\mu}{\pi} - 1\| = \frac{1}{2}\sum_{y\in\Omega} |\mu(y) - \pi(y)|$
- Variance: $Var_\pi(\mu/\pi) = |\frac{\mu}{\pi} - 1\|_{2,\pi}^2 = \sum_{y\in\Omega} \pi(y)(\frac{\mu(y)}{\pi(y)} - 1)^2$
- Informational divergence: $D(P^n(x, \cdot)\|\pi) = Ent_\pi(k_n^x) = \sum_{y\in\Omega} P^n(x, y) \log \frac{P^n(x,y)}{\pi(y)}$
  Where the entropy $Ent_\pi(f) = E_\pi f \log \frac{f}{E_\pi f}$.

Each of these distance are convex, which means given $s \in [0, 1], dist((1 - s)\mu + s\nu, \pi) \leq (1 - s)dist(\mu, \pi) + s\ dist(\nu, \pi)$. A convex distance satisfies the condition

$$dist(\sigma P^n, \pi) = dist(\sum_{x\in\Omega} \sigma(x)P^n(x, \cdot), \pi) \leq \sum_{x\in\Omega} \sigma(x)dist(P^n(x, \cdot), \pi) \leq max_{x\in\Omega}dist(P^n(x, \cdot), \pi) \tag{7}$$

so the distance is maximized when the initial distribution concentrate at a point.

Definition 2.1. *The total variation, relative entropy and L^2 mixing times are defined as follows.*

- $\tau(\epsilon) = \min\{n : \forall x \in \Omega, \|P^n(x, \cdot) - \pi\|_{TV} \leq \epsilon\}$
- $\tau_D(\epsilon) = \min\{n : \forall x \in \Omega, D(P^n(x, \cdot)\|\pi) \leq \epsilon\}$
- $\tau_2(\epsilon) = \min\{n : \forall x \in \Omega, \|k_n^x - 1\|_{2,\pi} \leq \epsilon\}$

The time-reversal P^* is defined by $\pi(x)P^*(x,y) = \pi(y)P(y,x),\ x,y \in \Omega$, and we have $\langle f, Pg\rangle_\pi = \langle P^* f, g\rangle_\pi$

For $f, g : \Omega \to R$, let $\mathcal{E}(f,g) = \mathcal{E}_P(f,g)$ denote the Dirichlet form,

$$\mathcal{E}(f,g) = \langle f, (I-P)g\rangle_\pi \tag{8}$$

if $f = g$ then

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2 P(x,y)\pi(x) \tag{9}$$

and

$$\mathcal{E}_P(f,f) = \mathcal{E}_{P^*}(f,f) = \mathcal{E}_{\frac{P+P^*}{2}}(f,f) \tag{10}$$

A useful property of the reversal is that $k_n = P^* k_{n-1}$. If $P^* = P$, then P is said to be the time-reversible, or to satisfy the detailed balance condition. It's almost trivial to see that $\frac{P+P^*}{2}$ and PP^* are time-reversible. If P is reversible then $\mathcal{E}(f,g) = \mathcal{E}(g,f)$

Also, we borrow some notation from complexity theory, including O, Θ, Ω.

2.2. **Continuous Time.** Discrete Laplacian $\mathcal{L} = -(I-P)$. Then for $t \geq 0$, $H_t = e^{t\mathcal{L}}$ represents the continuized chain, corresponding to the discrete Markov kernel P. The contuized chain simply represents a Markov process $\{X_t\}_{t\geq 0}$. Also, let $h_t^x(y) = H_t(x,y)/\pi(y)$.

Lemma 2.2. *For any h_0 and all $t \geq 0$, $h_t = H_t^* h_0$. Consequently, for any $x \in \Omega$,*

$$\frac{dh_t(x)}{dt} = \mathcal{L}h_t(x)$$

Lemma 2.3.

$$\frac{d}{dt}Var(h_t) = -2\mathcal{E}(h_t, h_t) \tag{11}$$

$$\frac{d}{dt}Ent(h_t) = -\mathcal{E}(h_t, \log h_t) \tag{12}$$

Proof. Trivial to show. □

Spectral gap λ and **entropy constant** ρ_0

Definition 2.4. *Let $\lambda > 0$ and $\rho_0 > 0$ be the optimal constants in the inequalities:*

$$\lambda Var_\pi f \leq \mathcal{E}(f,f), \forall f : \Omega \to R. \tag{13}$$

$$\rho_0 Ent_\pi f \leq \mathcal{E}(f, \log f), \forall f : \Omega \to R_+. \tag{14}$$

Corollary 2.5. *Let $\pi_* = min_{x\in\Omega}\pi(x)$. Then in the continuous time,*

$$\tau_2(\epsilon) \leq \frac{1}{\lambda}\big(\frac{1}{2}\log\frac{1-\pi_*}{\pi_*} + \log\frac{1}{\epsilon}\big) \tag{15}$$

$$\tau_D(\epsilon) \leq \frac{1}{\rho_0}\big(\log\log\frac{1}{\pi_*} + \frac{1}{\epsilon}\big) \tag{16}$$

Proof. Simply solve the differential equations,

$$\frac{d}{dt}Var(h_t^x) = -2\mathcal{E}(h_t^x, h_t^x) \leq -2\lambda Var(h_t^x) \tag{17}$$

$$\frac{d}{dt}Ent(h_t^x) = -\mathcal{E}(h_t^x, \log h_t^x) \leq Ent(h_t^x) \tag{18}$$

and note that $Var(h_0) \leq \frac{1-\pi_*}{\pi_*}$ and $Ent(h_0) \leq \log \frac{1}{\pi_*}$ (e.q. by equation (6)) □

Proposition 2.6. *If $c > 0$ then*

(1) $Var_\pi(H_t f) \leq e^{-ct} Var_\pi f, \forall f \; and \; t > 0$*, if and only if* $\lambda \geq c$.
(2) $Ent_\pi(H_t f) \leq e^{-ct} Ent_\pi f, \forall f > 0 \; and \; t > 0$*, if and only if* $\rho_0 \geq c$

ρ_0 is rather challenging to estimate while there have been several techniques (linear-algebraic and functional-analytic) to help bound the spectral gap. The following inequality relating the two Dirichlet forms introduced above also motivates the study of classical logarithmic Sobolev inequality.

Lemma 2.7. *If $f \geq 0$ then*

$$2\mathcal{E}(\sqrt{f}, \sqrt{f}) \leq \mathcal{E}(f, \log f) \tag{19}$$

Let $\rho_P > 0$ denote the logarithmic Sobolev constant of P defined as follows.

Definition 2.8.

$$\rho = \rho_P = inf_{Entf^2 \neq 0} \frac{\mathcal{E}(f, f)}{Entf^2}$$

Proposition 2.9. *For every irreducible chain P,*

$$2\rho \leq \rho_0 \leq 2\lambda$$

Proof. Using (18), the first inequality is immediate. The second follows from applying (13) to function $f = 1 + \epsilon g$, where $g \in L^2(\pi)$ and $E_\pi g = 0$. Assume $\epsilon \ll 1$, so that $f \geq 0$. Then using the Taylor approximation, $\log(1 + \epsilon g) = \epsilon g - 1/2(\epsilon)^2 g^2 + o(\epsilon^2)$, we have

$$Ent_\pi(f) = \frac{1}{2}\epsilon^2 \pi(g)^2 + o(\epsilon^2)$$

and

$$\mathcal{E}(f, \log f) = -\epsilon E_\pi((\mathcal{L}g)\log(1 + \epsilon g)) = \epsilon^2 \mathcal{E}(g, g) + o(\epsilon^2).$$

Thus starting from (13), and applying to f as above, we get

$$\epsilon^2 \mathcal{E}(g, g) \geq \frac{\rho_0}{2}\epsilon^2 E_\pi g^2 + o(\epsilon^2).$$

Canceling ϵ^2 and letting $\epsilon \downarrow 0$, yields the second inequality of the proposition, since $E_\pi g = 0$ □

2.3. **Discrete Time.** A mixing bound in terms of the spectral gap will be shown in a fashion similar to that in continuous time. There seems to be no discrete-time analog of the modified log-Sobolev bound on relative entropy.

In discrete time we consider two approaches to mixing time, both of which are equivalent. the first approach involves operator norms, and is perhaps the more intuitive of two methods.

Proposition 2.10.

$$\tau_2(\epsilon) \leq \lceil \frac{1}{1 - \|P^*\|} \log \frac{1}{\epsilon\sqrt{\pi_*}} \rceil \tag{20}$$

where $\pi_* = \min_{x\in\Omega} \pi(x)$ *and*

$$\|P^*\| = \sup_{f:\Omega\to R, Ef=0} \frac{\|P^* f\|_2}{\|f\|_2}$$

This results has appeared in mixing time literature in many equivalent forms.

Proof. Since $k_{i+1} - 1 = P^*(k_i - 1)$ and $E(k_i - 1) = 0$ for all i then

$$\|k_n - 1\|_2 = \|P^*\|^n \|k_0 - 1\|_2$$

Solving for when this expression drops to ϵ and using the approximations $\log x \leq -(1 - x)$ and $\|k_0 - 1\|_2 \leq \sqrt{\frac{1-\pi_*}{\pi_*}}$ □

In above proposition, the mixing bound followed almost immediately from the definition. However, there is an alternate approach to this problem which bear more of a resemblance to the continuous time result and is more convenient for showing refined bounds.

The discrete time analog of differentiating $Var(h_t)$ is take the difference $Var(k_n) - Var(k_{n-1})$, or more generally, $Var(P^* f) - Var(f)$.

Lemma 2.11. *Given Markov chain* P *and function* $f : \Omega \to R$*, then*

$$Var(P^* f) - Var(f) = -\mathcal{E}_{PP*}(f, f) \leq -Var(f)\lambda_{PP^*}$$

.

Lemma 2.12. *Note, for any transition probability matrix* K*,* $E_\pi f = E_\pi(Kf)$*. Then*

$$Var(P^* f) - Var(f) = \langle P^* f, P^* f\rangle_\pi - \langle f, f\rangle_\pi = -\langle f, (I - PP^*)f\rangle_\pi$$

gives what we need.

Corollary 2.13. *A discrete time Markov chain* P *satisfies*

$$\tau_2(\epsilon) \leq \lceil \frac{1}{\lambda_{PP^*}} \log \frac{1}{\epsilon\sqrt{\pi_*}} \rceil \tag{21}$$

Proof.

$$Var(k_n) \leq Var(k_0)(1 - \lambda_{PP^*})^n \tag{22}$$

The result follows by solving for when variance drops to ϵ^2 and using the approximation $\log(1 - \lambda_{PP^*}) \leq -\lambda_{PP^*}$ □

It is preferable to work with P instead of PP^*. Here are several simplifications make this possible.

Corollary 2.14. *In discrete time, a Markov chain with holding probability α satisfies*

$$\tau_2(\epsilon) \leq \lceil \frac{1}{\alpha\lambda} \log \frac{1}{\epsilon\sqrt{\pi_*}} \rceil \tag{23}$$

For a reversible Markov chain,

$$\tau_2(\epsilon) \leq \lceil \frac{1}{1-\lambda_{max}} \log \frac{1}{\epsilon\sqrt{\pi_*}} \rceil \leq \lceil \frac{1}{\min\{2\alpha, \lambda\}} \log \frac{1}{\epsilon\sqrt{\pi_*}} \rceil \tag{24}$$

Where $\lambda_{max} = \max\{\lambda_1, |\lambda_{n-1}|\}$ when $\lambda_1 = 1 - \lambda$ is the largest non-trivial eigenvalue of P and $\lambda_{n-1} \geq -1$ is the smallest eigenvalue.

Now we will show that those two approaches to bounding discrete mixing in this section are equivalent.

Proposition 2.15.

$$1 - \lambda_{PP^*} = \|P^*\| \tag{25}$$

Proof. See the original paper, page 20. □

2.4. **Does Reversibility matter?** Many mixing result were originally shown only in the context of a reversible Markov chain. We can actually avoid this requirement in most cases. However, there are still some cases that this requirement is necessary. In this section, some difference would be covered, and some classical results of reversible Markov chain will be illustrated.

The difference between reversible and non-reversible results is most apparent when upper and lower bounds on distances are given. Let

$$d(n) = \max_x \|P^n(x, \cdot) - \pi(\cdot)\|_{TV}$$

denotes the worst variation distance after n steps. Then combining the above work, we have

Proposition 2.16. *if P is reversible:*

$$\frac{1}{2}\lambda_{max}^n \leq d(n) \leq \frac{1}{2}\lambda_{max}^n \sqrt{\frac{1-\pi_*}{\pi_*}} \tag{26}$$

if P is non-reversible:

$$x \tag{27}$$

Lemma 2.17. *If P is reversible and irreducible on state space of size $|\Omega| = n$, then it has a complete spectrum of real eigenvalues with magnitudes at most one, that is*

$$1 = \lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{n-1} \geq -1. \tag{28}$$

The Courant-Fischer theorem show the connection between eigenvalues and Dirichlet forms for a reversible Markov chain.

Lemma 2.18. *In a reversible Markov chain the second largest eigenvalue $lambda_1$ and the smallest eigenvalue λ_{n-1} satisfy*

$$1 - \lambda_1 = \inf_{Var(f) \neq 0} \frac{\mathcal{E}(f, f)}{Var(f)} = \lambda \tag{29}$$

$$1 + \lambda_{n-1} = \inf_{Var(f) \neq 0} \frac{\mathcal{F}(f, f)}{Var(f)} = \lambda \tag{30}$$

Where

$$\mathcal{F}(f, f) = \langle f, (I + P)f \rangle_\pi$$

Proof. See the original paper, page 25. □

3. Advanced Functional Techniques

In this section, it is shown that information on function of large variance, or on functions with small support, can be exploited to show better mixing time bounds.

The argument is simple. Recall that we have $\frac{d}{dt}Var(h_t) = -2\mathcal{E}(h_t, h_t)$. If $\mathcal{E}(f, f) \geq G(Var(f))$ for some $G : R_+ \to R_+$ and $f : \Omega \to R_+$ with $Ef = 1$, the it follows that $\frac{d}{dt}Var(h_t) = -2\mathcal{E}(h_t, h_t) \leq -2G(Var(f))$, so we have the following inequality:

$$\tau_2(\epsilon) = \int_0^{\tau_2(\epsilon)} 1dt \leq \int_{Var(h_0)}^{\epsilon^2} \frac{dI}{-2G(I)} \tag{31}$$

When $G(r) = \lambda r$, it simply gives the bound we got in previous chapter. More generally, in this chapter we derive such functions G in terms of the log-Sobolev constant, Nash inequalities, spectral profile, or via comparison to another Markov chain.

In the discrete time, replace $\frac{d}{dt}Var(h_t) = -2\mathcal{E}(h_t, h_t)$ with $Var(k_n) - Var(k_{n-1}) = -\mathcal{E}_{PP*}(k_n, k_n)$(is it correct? shouldn't it be k_{n-1}). Then if $\mathcal{E}_{PP*}(f, f) \geq G_{PP^*}(Var(f))$, and $I(n) = Var(k_n)$ and $G_{PP^*}(r)$ are non-decreasing, the piecewise linear extension of $I(n)$ to $t \in R_+$ will satisfy

$$\frac{dI}{dt} \leq -G_{PP^*}(I). \tag{32}$$

Then

$$\tau_2(\epsilon) = \int_0^{\tau_2(\epsilon)} 1dt \leq \int_{Var(h_0)}^{\epsilon^2} \frac{dI}{-G_{PP^*}(I)} \tag{33}$$

Since we have

$$\mathcal{E}_{PP^*}(f, f) \leq 2\alpha\mathcal{E}(f, f) \leq 2\alpha G(Var(f))$$

then we may take $G_{PP^*}(r) = 2\alpha G(r)$

3.1. Log-Sobolev and Nash Inequalities.

Some of the best bounds on L^2 mixing times were shown by use of the log-Sobolev constant. Following is a example. (But I can't talk about what's the bound is in this case).

Example 3.1. *A matroid $\mathcal{M}$ is given by a ground set $E(\mathcal{M})$ with $|E(\mathcal{M})| = n$ and a collection of bases $\mathcal{B}(\mathcal{M}) \subseteq 2^{\mathcal{B}(\mathcal{M})}$. The bases $\mathcal{B}(\mathcal{M})$ must all have the same cardinality r, and $\forall X, Y \in \mathcal{B}(\mathcal{M}), \forall e \in X, \exists f \in Y : X \cup \{f\} \setminus \{e\} \in \mathcal{B}(\mathcal{M})$. One choice of a Markov chain on matroids is, given state $X \in \mathcal{B}(\mathcal{M})$ half the time do nothing, and otherwise choose $e \in X, f \in \mathcal{B}(\mathcal{M})$ and transition to state $X - e + f$ if it is also a basis.*

Equation (30) and (31) will bound mixing time in terms of log-Sobolev constant if we can show a relation between the Dirichlet form and a function of the variance. The following lemma establishes this connection.

Lemma 3.2. *If f is non-negative then*

$$Ent(f^2) \geq Ef^2 \log \frac{Ef^2}{(Ef)^2}$$

and in particular, if $Ef = 1$ then

$$\mathcal{E}(f, f) \geq \rho Ent(f^2) \geq \rho(1 + Var(f)) \log(1 + Var(f)).$$

Proof. See the original paper on page 29. □

In some case, Nash inequality can be used to supplement the log-Sobolev result and improve the bound. Unfortunately, however, Nash inequalities are notoriously difficult to establish. We now show that the Dirichlet form can also be lower bounded in terms of variance by using Nash inequality.

Lemma 3.3. *Given a Nash Inequality*

$$\|f\|_2^{2+1/D} \leq C\left[\mathcal{E}(f,f) + \frac{1}{T}\|f\|_2^2\right]\|f\|1^{1/D}$$

which holds for every function $f : \Omega \to R$ and some constants $C, D, T \in R_+$, then whenever $f \geq 0$ and $Ef = 1$ then

$$\mathcal{E}(f,f) \geq (1 + Var(f))\left(\frac{(1+Var(f))^{1/D}}{C} - \frac{1}{T}\right)$$

Proof. Simply rewrite the inequality and use the facts that $\|f\|_1 = 1$ and $Var(f) = \|f\|_2^2 - 1$. □

3.2. **Spectral profile.** In previous section it was found that log-Sobolev bounds on mixing time can on spectral gap results, by replacing the $\log(1/\pi_*)$ term with $\log\log(1/\pi_*)$. However, the log-Sobolev is much more difficult to bound than the spectral gap and, to date, bounds on it are know for only a handful of problems. Moreover, sometimes even log-Sobolev is not strong enough. In this section, the method of Spectral Profile will be introduced. The main idea is to improve the basic relation $\mathcal{E}(f,f) \geq \lambda Var(f)$ by considering the support of f.

Definition 3.4. *For a non-empty subset $S \subset \Omega$ the first Dirichlet eigenvalue on S is given by*

$$\lambda_1(S) = \inf_{f \in c_0^+(S)} \frac{\mathcal{E}(f,f)}{Var(f)} \tag{34}$$

where $c_0^+(S) = \{f \geq 0 : supp(f) \subset S\}$ is the set of non-negative function supported on S. The spectral profile $\Lambda : [\pi_*, \infty) \to R$ is given by $\Lambda(r) = \inf_{\pi_* \leq \pi(S) \leq r} \lambda_1(S)$

The spectral profile is a natural extension of spectral gap λ, and will we now see that it can improve on the basic bound $\mathcal{E}(f,f) \geq \lambda Var(f)$ used earlier.

Lemma 3.5. *For every non-constant function $f : \Omega \to R_+$*

$$\mathcal{E}(f,f) \leq \frac{1}{2}\Lambda\left(\frac{4(Ef)^2}{Varf}\right)Var(f) \tag{35}$$

Proof. Proof can be founded in the original paper, page 35. I can also talk about that. □

Applying this formula together with (30), we obtain following inequality.

$$\tau_2(\epsilon) \leq \int_{Var(h_0)}^{8} \frac{dI}{-I\Lambda(4/I)} + \int_8^{\epsilon^2} \frac{dI}{-2\lambda I}$$

Let σ be the initial distribution, and let $h_0(x) = \frac{\sigma(x)}{\pi(x)}$, a change of variables to $r = 4/I$ gives the mixing time bound (in continuous time).

$$\tau_2(\epsilon) \leq \int_{4/Var(\sigma/\pi)}^{1/2} \frac{dr}{r\Lambda(r)} + \frac{1}{\lambda}\log\frac{2\sqrt{2}}{\epsilon} \tag{36}$$

Similar things can be done in discrete-time case.

The theorem, with the trivial bound $\Lambda(r) \geq \lambda$ sometimes produces worse bound than we got in previous chapter, however, when $\Lambda(r) \gg \lambda$ when r are small, we could expect the obtain sharper bound.

Spectral profile is a fairly new tool, and it has not been widely studied yet. However, some researchers still got some results to bound spectral profile, for example, it can be shown that log-Sobolev constant and a Nash inequality induce the following lower bounds:

$$\Lambda(r) \geq \rho \frac{\log(1/r)}{1-r}$$

and

$$\Lambda(r) \geq \frac{1}{Cr^{1/2D}} - \frac{1}{T}.$$

3.3. **Comparison methods.** It sometimes happened that a Markob chain is difficult to study, but a related chain is more manageable. In this situation, the comparison method has been widely used to bound spectral gap, log-Sobolev constant and Nash inequalities.

Before deriving comparison results for the quantities in this chapter, a preliminary result is needed.

Theorem 3.6. *Consider two Markov chains P and $\hat{P}$ on the same state space Ω, and for every $x \neq y \in \Omega$ with $\hat{P}(x,y) \geq 0$ define a directed path γ_{xy} from x to y along edges in P. Let Γ denote the set of all such paths. Then*

$$\mathcal{E}_P(f,f) \geq \frac{1}{A}\mathcal{E}_{\hat{P}}(f,f), \tag{37}$$

$$Var_\pi(f) \leq M Var_{\hat{\pi}}(f),\ Ent_\pi(f^2) \leq M Ent_{\hat{\pi}}(f^2) \tag{38}$$

Where $M = \max_x \frac{\pi(x)}{\hat{\pi}(x)}$, and

$$A = A(\Gamma) = \max_{a \neq b; P(a,b) \neq 0} \frac{1}{\pi(a)P(a,b)} \sum_{x \neq y; (a,b) \in \gamma_{xy}} \hat{\pi}(x)\hat{P}(x,y)|\gamma_{xy}|$$

Proof. see page 39 in original paper. □

Corollary 3.7.

$$\lambda_P \geq \frac{1}{MA}\lambda_{\hat{P}},\ \rho_P \geq \frac{1}{MA}\rho_{\hat{P}},\ \Lambda_P(r) \geq \frac{1}{MA}\Lambda_{\hat{P}}(r)$$

In the case of a reversible chain, it is also possible to compare λ_{n-1} if the paths are of odd length. First, consider the preliminary result.

Theorem 3.8. *Consider two Markov chains P and $\hat{P}$ on the same state space Ω, and for every $x \neq y \in \Omega$ with $\hat{P}(x,y) \geq 0$ define a directed path γ_{xy} of odd length $|\lambda_{x,y}|$ from x to y along edges in P. Let Γ^* denote the set of all such paths. Then*

$$\mathcal{F}_P(f,f) \geq \frac{1}{A^*}\mathcal{F}_{\hat{P}}(f,f), \tag{39}$$

Where $M = \max_x \frac{\pi(x)}{\hat{\pi}(x)}$, and

$$A^* = A^*(\Gamma^*) = \max_{a \neq b; P(a,b) \neq 0} \frac{1}{\pi(a)P(a,b)} \sum_{x \neq y; (a,b) \in \gamma_{xy}} \hat{\pi}(x)\hat{P}(x,y)|\gamma_{xy}|r_{xy}(a,b)$$

where $r_{xy}(a,b)$ is the number of times the edge (a,b) appears in path γ_{xy}.

If P and $\hat{P}$ are reversible then

$$1 - \lambda_{max}(P) \geq \frac{1}{MA^*}(1 - \lambda_{max}(\hat{P}))$$

The most widely used example of these comparison results is the canonical path theorem

Corollary 3.9. *Given a Markov chain P on state space Ω, and directed paths λ_{xy} between every pair of vertices $x \neq y \in \Omega$, then*

$$\lambda \geq (\max_{a \neq b; P(a,b) \neq 0} \frac{1}{\pi(a)P(a,b)} \sum_{x \neq y; (a,b) \in \gamma_{xy}} \pi(x)\pi(y)|\gamma_{xy}|)^{-1}$$

Proof. Let $\hat{P}(x,y) = \pi(y), \hat{\pi} = \pi$, then $M = 1$, note that

$$\mathcal{E}_{\hat{P}}(f,f) = \frac{1}{2} \sum_{x,y \in \Omega} (f(x) - f(y))^2 \pi(x)\pi(y) = \text{Var}_\pi(f).$$

□

Corollary 3.10. *Consider two Markov chains P on the state space Ω, and a set of cycles γ_x of odd length from each vertex $x \in \Omega$ to itself. Then the smallest eigenvalue λ_{n-1} of P satisfies the relation*

$$1 + \lambda_{n-1} \geq 2\left(\max_{a \neq b; P(a,b) \neq 0} \frac{1}{\pi(a)P(a,b)} \sum_{x \neq y; (a,b) \in \gamma_{xy}} \pi(x)|\gamma_{xy}|r_{xy}(a,b)\right)^{-1}$$

REFERENCES

[1] Montenegro, Ravi, and Prasad Tetali. ***Mathematical aspects of mixing times in Markov chains.*** Now Pub, 2006.

[2] Stein, Elias M., and Rami Shakarchi. ***Real analysis: measure theory, integration, and Hilbert spaces***. Vol. 3. Princeton University Press, 2010.