ON THE LINK BETWEEN MARKOV CHAIN MIXING TIME AND GRADIENT DESCENT CONVERGENCE BOUND

LINA PIAO

ABSTRACT. The main aim of this paper is to establish the relations between optimization and sampling, more specifically, the mixing time of reversible Markov chains and the convergence bound of gradient descent.

Keywords. markov chain, reversible markov chain, markov chain mixing time, gradient descent, gradient descent convergence bound, spectral theorem, optimization, sampling

CONTENTS

1. INTRODUCTION

1.1. **Optimization and Sampling.**

This paper, on a high level, shows the unexpected and close relationship between *optimization* and *sampling*. We will now briefly explore these two topics.

To understand the idea of *optimization*, consider a road trip. You want to get to your destination as quickly as possible, but you also want to avoid toll roads, minimize fuel consumption, and perhaps enjoy some scenic views along the way. The process of deciding the best route that balances these factors is a form of *optimization*. In this case, you're trying to optimize your route based on time, cost, and experience.

At its core, *optimization* is about finding the best possible solution to a problem within a set of given constraints. Formally, this is modelled mathematically with a real-valued function called an objective function defined over a set of valid solutions. The goal is, for example, to find the solution that minimizes the function value. In this example, the set of solutions is all feasible ways to solve the problem (i.e., all the paths you can take), whereas the objective function gives the "cost" (real number) associated to each path.

In particular, if the objective function and set of solutions are convex, one powerful algorithm is gradient descent, which we will explore further.

To understand the idea of *sampling* in the context of the paper, consider a dance class. In this class, people are paired up with someone different each week. However, we would like to ensure that any pairing have compatible dance styles. Now, we must decide how to construct a set of pairings such that it works for everyone. If we had a way to *sample* a set of pairings from the set of all possible sets of pairings, we could use this algorithm to construct a set of compatible and random pairings each week. The randomness will ensure that the pairings are sufficiently different each week.

Therefore, the task of *sampling* is to find an algorithm that constructs a sample from a given distribution.

Here is how an example of this algorithm works. We maintain a set of pairings which is initially empty. Now, we choose a viable pairing (from the set of all pairs of people that are compatible dance partners) at random, and compare it to the set of pairings we maintain. If the two people in the proposed pairing are already partnered with each other, we will remove that pairing from the set of pairings. If neither has partners, we partner them together, and finally, if one of them has a partner, we will partner them with the proposed partner instead of their current partner. We repeat this process for some number of times until the distribution of the set of pairings is sufficiently random. This is an example of an algorithm that is called Markov Chain Monte Carlo.

Clearly at the start, the set of pairings we maintain is not random, but rather deterministic. This is far from the distribution on the set of viable sets of pairings we want. However, as we repeat the process, the outcome becomes more random, and in fact, the distribution of viable sets of pairings gets closer to the desired distribution. The theory of mixing time helps us quantify how long we need to run this process until the distribution output is close enough to the distribution we want. By close enough, we use the total variation distance, which is a standard measure of distance between probability distributions used throughout statistics.

This act of removing and adding pairings is an example of a Markov chain. At any time, the distribution of the next set of pairings only depends on the current set of pairings and not any previous pairings. This property is called the Markov property, as explored later.

From a high level, in Markov Chain Monte Carlo, we perform a random walk over the state space of all possible outcomes, starting from any fixed outcome of the distribution. Each step of this walk performs some random local transformations to the current outcome to get another (random) one. Importantly, these random local transformations are usually much easier to implement, and it turns out that if this process is designed correctly, then letting it run long enough will produce a sample that is distributed almost as if we sampled from the desired distribution.

As we discussed in the pairing example, the random walk is modeled by an object in probability called a Markov chain.

1.2. **Connection between Sampling and Optimization.**

While the two subjects, sampling and optimization, might at first glance seem disconnected from each other, a lot of research has now been built on the interesting relation between the two. One of the foremost examples of the connection between sampling and optimization was given in [JKO98]. In this paper, we will show that a version of this connection can prove useful even in the case of discrete time, finite space Markov chains. For our purposes, the connection between Markov chains and gradient descent comes from the following observations:.

First, we will see from the basic theory of Markov chains that given some starting distribution on the states of the chain, the distribution after running the Markov chain one step can be found with a matrix multiplication. On the other hand, the gradient of a quadratic form is itself a matrix multiplication, so gradient descent also performs matrix multiplication. Thus, gradient descent on the right objective will produce iterates that exactly equal the distributions of the Markov chain, and so we can transfer statements about convergence of gradient descent to statements about convergence of Markov chains to their stationary distributions.

Our goal for this paper is to use this idea to show the standard spectral mixing time bound for Markov chain.

2. Preliminaries

In this section we will talk about the basic definitions and theorems we will be using in the paper. We will only be giving a high level overview, but for depth, the reader may consult [BH14] [Ove23] [Sig09] for Markov chains and [Gup19][Kak15][Rya15] for gradient descent.

2.1. **Markov Chains.**

Example 2.1. To introduce the idea of Markov chains, consider a bunny.

Let this bunny exist in a world where there are three fields they can migrate between. Now, the likelihood of the bunny moving between any two of the three fields has fixed values. Given these conditions, there are some questions that we may ask.

(1) What is the probability that after n migrations, the bunny is in a given field?
(2) Does the probability that the bunny is in a given field converge? If the probability were to converge
    (a) Is there the kth migration in which it converges?
    (b) Can we find how close we are to the convergence (later defined as stationary distribution) after n migrations?

These are all questions which Markov chains allow us to answer.

Now, let us define Markov chains.

Definition 2.2 (Markov Chain)**.** A Markov chain is a stochastic process that acts on a set of states Ω in which the probability of each event (migration to a specific field, Ex 2.1) only depends on the state (field, Ex 2.1) of the previous event.

That is, if we let X_t represent the state at time t and let $X_t = x$ such that $x \in \Omega$ (where Ω is the set of states), we know that the Markov chain makes a transition from one state (X_t) to the other (X_{t+1}) by

$$P(X_{t+1} = y | X_t = x)$$

where $P(X_{t+1} = y|X_t = x)$ is the conditional probability of getting y given that in the previous step we had x.

As seen in the definition above, one of the most important properties of a Markov chain is the *Markov property*, which implies that the state at time $t + 1$ depends solely on the state at time t, in other words:

$$P(X_{t+1}|X_0, \cdots, X_t) = P(X_{t+1}|X_t)$$

.

Now, we will define an important concept of Markov chain that will be continuously referred to within the paper.

Definition 2.3 (Stationary Distribution)**.** A probability distribution π of a finite-space, discrete-time Markov chain is considered a stationary distribution of the transition matrix P if $\pi P = \pi$.

Essentially, the stationary distribution represents the equilibrium or steady-state behavior of the chain, where the probability distribution does not change as the Markov chain evolves over time.

Now, we will explore two specific types of Markov chains.

Definition 2.4 (Reversible Markov chains)**.** A Markov chain is said to be reversible if for all $i, j \in \Omega$,

$$\pi(i)P(i|j) = \pi(j)P(j|i),$$

where π is the stationary distribution.

Essentially, time-reversible Markov chains satisfy the detailed balance condition, which implies that for any two states i, j in the Markov chain, the probability flux from i to j is matched by that from j to i. Hence, time-reversible Markov chains do not exhibit a net flow of probability in any particular direction between states once it reaches the stationary distribution.

Note that it follows from the definition that a Markov chain with a transition matrix P is reversible with respect to $\pi \in \Delta_n = \{\nu \in \mathbb{R}^n_{\geq 0} | \sum_{i=1}^n \nu_i = 1\}$ (stationary distribution) if

$$\Pi P = P^T \Pi,$$

where Π is a diagonal matrix such that $\Pi = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n \end{bmatrix}$.

Now, it can be seen that ΠP is symmetric as $P^T \Pi^T = P^T \Pi$ given that Π is a diagonal matrix, and hence, it follows that

$$\begin{aligned} \Pi P &= P^T \Pi \\ &= P^T \Pi^T \\ &= (\Pi P)^T. \end{aligned}$$

Definition 2.5 (Irreducible Markov chains)**.** A Markov chain with transition matrix P is said to be irreducible if for any two states i, j, there exists a finite number of steps with positive probability to go from i to j. In other words, for each two states, there is some time where the probability of going from i to j is positive.

Essentially, if every state is accessible from any other state of the Markov chain, it is an irreducible Markov chain, and for the purpose of this paper, we will only be exploring reversible and irreducible Markov chains.

Now, before we define a crucial idea of the paper, we will first explore the fundamental theorem of Markov chain as it deeply relates to this idea. However, before that, we will define a crucial measure of length used in statistics.

Definition 2.6 (Total Variation Distance)**.** Let u, v be probability distributions.

We define the total variation distance as

$$||u - v||_{TV} := \sum_{x:u(x)>v(x)} (u(x) - v(x)) = \frac{1}{2}||u - v||_1 = \max_{A \in \Omega} u(A) - v(A)$$

where Ω is the set of states.

In other words, total variation distance takes the largest distance between the two states.

Theorem 2.7 (Fundamental Theorem of Markov Chain)**.** *Any irreducible, reversible and aperiodic Markov chain has a unique stationary distribution, π, and for all states x, y, as t (steps) approaches infinity,*

$$P^t(x|y) \to \pi(y)$$

Furthermore, for any $\epsilon > 0$, there exists $t > 0$ such that

$$||P^t(x|.) - \pi||_{TV} \leq \epsilon.$$

Therefore, the fundamental theorem of Markov chain states that starting from any place we will eventually get to the stationary distribution, and the mixing time quantifies how fast this happens.

Now, for state $x \in \Omega$ let

$$\tau_x(\epsilon) = \min\{t : ||P^t(x|.) - \pi||_{TV} \leq \epsilon\}$$

be the first time that the total variation distance between the chain started at x and the stationary distribution drops below ϵ.

Furthermore, define

$$\tau(\epsilon) = \max_x \tau_x(\epsilon).$$

That is, how long it takes (how many steps it takes), in the worst case, for the Markov chain to get close to the stationary distribution, no matter where it starts.

Definition 2.8 (Markov Chain Mixing Time)**.** Markov chain mixing time is defined as $\tau(1/2\epsilon)$ as from above. That is, this is the time that the total variation distance of the chain started at the worst possible starting point drops below $1/2\epsilon$.

In other words,

the time (step) until the Markov chain is "close" to its stationary distribution

2.2. Gradient Descent.

Example 2.9. To introduce the idea of gradient descent, consider a bunny.

Let this bunny exist in a world where there are hills, and the objective of the bunny is to find the lowest point in the landscape, a valley, where it can rest and find the most food. Given these conditions, there are some questions that we may ask.

1. What is the best strategy for the bunny to find the lowest point?
2. How can the bunny reach the lowest point?
    - (a) How much should the bunny jump each time?
    - (b) How long will it take for the bunny to reach the valley?

These are all questions gradient descent can help answer.

Note that throughout the paper, x^* will be used as the notation for the x value of the minimum (optimal value).

Definition 2.10 (Gradient Descent)**.** Gradient Descent is an optimization algorithm used to find the minimum of a function.

It iteratively moves in the direction of the steepest descent, as defined by the negative of the gradient of the function at the current point.

The most common algorithm or formula of gradient descent that is seen is the update rule, which causes the iterative moves in the direction of steepest descent.

Definition 2.11 (Update Rule of Gradient Descent)**.** Let $f(x)$ be a multi-variable function that is differentiable and x_n be the value of the current value of x.

In gradient descent, the next iterate (x_{n+1}) is related to the previous one (x_n) by

$$x_{n+1} = x_n - \eta(\nabla f(x_n)),$$

where η is the step size (or learning rate) and $\nabla f(x_n)$ is the gradient of the function at x_n.

Note that $\nabla f(x_n)$ is a vector of partial derivatives that points in the direction of the steepest increase of the function.

Next, we will introduce two types of functions integrally related to gradient descent. However, before we do, we will quickly define the Euclidean norm used in the definition of the two functions.

Definition 2.12 (Euclidean Norm)**.** The Euclidean norm, $||\mathbf{v}||_2$, measures the standard Euclidean distance from the origin to the point $\mathbf{v} \in \mathbb{R}^n$, and is defined as

$$||\mathbf{v}||_2 = \sqrt{\sum_{i=1}^{n} v_i^2}.$$

We will also need other inner products in order to define our gradient descent algorithm. In particular, for a matrix M, we let $\langle x, y\rangle_M = \langle Mx, My\rangle$ be the inner product induced by M, and define $\|x\|_M = \sqrt{\langle x, x\rangle_M}$ as with the Euclidean inner product. We will use a version of gradient descent that is modified for this inner product with the update rule

$$x_{n+1} = x_n - \eta(M^T M)^{-1}\nabla f(x_n)$$

To analyze the performance of gradient descent, we will require the objective to satisfy certain properties, which we discuss now.

Definition 2.13 (μ-strongly convex)**.** A function f is μ-strongly convex with respect to the inner product $\langle\cdot,\cdot\rangle_M$ if for all x, x',

$$f(x') \geq f(x) + \langle \nabla f(x), (x'-x)\rangle + \frac{\mu}{2}||x-x'||_M^2.$$

Intuitively, this means the function does not change too rapidly and that its gradient is bounded by L. Note that the term $\langle \nabla f(x), (x'-x)\rangle$ uses the usual Euclidean inner product, not the inner product induced by M.

Definition 2.14 (L-smooth)**.** A function f is L-smooth with respect to $\langle\cdot,\cdot\rangle_M$ if for all x, x',

$$f(x') \leq f(x) + \langle \nabla f(x), (x'-x)\rangle + \frac{L}{2}||x-x'||_M^2.$$

Intuitively, this means that the function has a strong "curvature" with a lower bound on how steep it can be. In the case of an objective $f(x) = \frac{1}{2}x^TAx$ for a symmetric matrix A, it turns out that the smoothness and strong convexity are related to the eigenvalues of the matrix $M^{-T}AM^{-1}$:

Claim 2.15. *$f(x)$ defined above is $\lambda_{\max}(M^{-T}AM^{-1})$-smooth with respect to $\langle\cdot,\cdot\rangle_M$ and $\lambda_{\min}(M^{-T}AM^{-1})$-strongly convex with respect to $\langle\cdot,\cdot\rangle_M$.*

This claim follows from the characterization of smoothness and strong convexity based on the eigenvalues of the Hessian (the second derivative of f), see e.g. [Faw13], and a change of basis argument.

Now, we will state the convergence theorem for gradient descent on smooth and strongly convex functions. We will use this later to prove the spectral mixing time bound of Markov chains.

Theorem 2.16 (Gradient Descent Convergence Bound)**.** *Fix an inner product $\langle\cdot,\cdot\rangle_M$. Let f be μ-strongly convex and differentiable and L-smooth with respect to M. Then if we run gradient descent with respect to this inner product for t iterations with a fixed step size $\eta \leq \frac{1}{L}$, it yields a solution which satisfies the following bound:*

$$\|x_t - x^*\|_M^2 \leq (1-\eta\mu)^t \|x_0 - x^*\|_M^2$$

where x_t is the t^{th} x iterate, x_1 is the first iterate, and $x^* = \arg\min_x f(x)$.

This shows that gradient descent, with an appropriate step size, converges exponentially to the optimal solution. Our convergence bound is slightly more general than the usual gradient descent convergence bound because it works for any inner product rather than just the standard inner product. This theorem is proved by successive application of the following property, which is called the contraction property.

Lemma 2.17 (Contraction Property)**.** *Let f be L-smooth and μ-strongly convex with respect to the inner product $\langle\cdot,\cdot\rangle_T$, then for gradient descent with respect to this inner product $x' = x - \eta(M^TM)^{-1}\nabla f(x)$ with $\eta = \frac{1}{L}$, we have:*

$$||x' - x^*||_M^2 \leq (1-\eta\cdot\mu)||x-x^*||_M^2.$$

For a proof of this lemma, refer to Section 2.2 of [Gow18]. The proof given is in the Euclidean norm. However, it can be modified to accommodate any norm induced by an inner product.

3. Relations between Markov Chain Mixing Time and Gradient Descent Convergence Bound

Let P be the transition matrix of a reversible (with respect to π, the stationary distribution) and irreducible Markov chain, and Π be the diagonal matrix of π.

Consider the objective function

$$f(\nu) = \frac{1}{2}\nu^T \Pi^{-1}(I - P^T)\nu$$

where $\Pi^{-1}(i,i) = \frac{1}{\pi(i)}$. We will first explore a general intuition behind this objective function.

Consider the Laplacian matrix of the Markov chain, $I - P^T$. Since we know that the Markov chain is reversible, and hence, ΠP is symmetric, it follows that $\Pi^{-1}(I - P^T)$ is symmetric, as proven below.

Claim 3.1. *$\Pi^{-1}(I - P^T)$ is symmetric.*

Proof. Since $\Pi P = P^T \Pi$, if we multiply both sides by Π^{-1}:

$$\Pi^{-1}\Pi P \Pi^{-1} = \Pi^{-1} P^T \Pi \Pi^{-1}$$
$$P\Pi^{-1} = \Pi^{-1} P^T \ (*)$$

It follows that

$$\begin{aligned}(\Pi^{-1}(I - P^T))^T &= (I - P)\Pi^{-1} \\ &= \Pi^{-1} - P\Pi^{-1} \\ &= \Pi^{-1} - \Pi^{-1}P^T \\ &= \Pi^{-1}(I - P^T)\end{aligned}$$

Therefore, $\Pi^{-1}(I - P^T)$ is symmetric. □

Now, $\nu^T \Pi^{-1}(I - P^T)\nu$ is a quadratic form, which is common in optimization problems, and for a general quadratic function $f(\nu) = \frac{1}{2}\nu^T A \nu$, the gradient is $\nabla f(\nu) = A\nu$ where A is a matrix. Notice that there is a general similarity in both Markov chains and gradient descent for quadratic forms in that to get to the next state/point, they both multiply a matrix to the old state/point. Therefore, the following goal of this paper is to show that gradient descent allows us to achieve the mixing time of Markov chains.

Now, consider the Laplacian matrix of the Markov chain ($L = I - P$).

From Claim 2.15, we know that the smoothness and strong convexity of $f(x) = \frac{1}{2}x^T Ax$ are the same as the largest and smallest eigenvalues of $M^{-T}AM^{-1}$, respectively. In this case, $M = \Pi^{-\frac{1}{2}}$ and $A = \Pi^{-1}(I - P^T)$, and hence, the smoothness and strong convexity of $f(x)$ are the largest and smallest eigenvalues of $\Pi^{-\frac{1}{2}}(I - P^T)\Pi^{\frac{1}{2}}$, which is, by definition, similar to $I - P^T$. Therefore, it has the same eigenvalues of the transpose of the Laplacian since $L^T = I - P^T$, and hence, has the same eigenvalues of the Laplacian.

Note that there is a case where $\Pi^{-1}(I - P^\top)\pi = 0$; however, for simplicity, we will simply ignore what happens in this eigenspace. (A formal argument can be made using quotient spaces.)

Now, since $L = I - P$ and we know that P is row-stochastic, if we let λ_j be any eigenvalue of P, we know that $-1 \leq \lambda_j \leq 1$ by the basic theory of Markov chains. Hence, if we let λ_i be any eigenvalue of L, we know that $0 \leq \lambda_i \leq 2$.

Therefore, if we order the eigenvalues of L, we get $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq 2$ where λ_1 corresponds to the stationary distribution of the Markov chain.

Here we introduce the spectral gap. The spectral gap is a property of the Markov chain that quantifies its connectivity, and we will show that it controls the mixing time. Now, the spectral gap of P (difference between the largest and the second-largest eigenvalues) is $\lambda_1(P) - \lambda_2(P)$ where the eigenvalues of P are ordered $\lambda_1(P) \geq \lambda_2(P) \geq \cdots \geq \lambda_n(P)$. As the largest eigenvalue corresponds to the stationary distribution, $\lambda_1(P) = 1$, it follows that the spectral gap is $\lambda_1(P) - \lambda_2(P) = 1 - \lambda_2(P)$. Now, since $L = I - P$, $1 - \lambda_2(P) = \lambda_2(L)$. Therefore, the spectral gap of P is $\lambda_2(L)$.

Now, we will prove the key lemma using connection between Markov chains and gradient descent.

Given a Markov chain P, associated with it is another Markov chain, which we call the 'lazy' version of the Markov chain. The transition matrix of the lazy version of the Markov chain is $P' = \frac{1}{2}(P+I)$. From the perspective of basic Markov chain theory, this ensures that the Markov chain is aperiodic, which is necessary for the fundamental theorem to hold. From our perspective, for the gradient descent contraction property to hold, we need to choose a step size that depends on the smoothness of the function we optimize. In this case, the step size we must choose means that gradient descent will be simulating the lazy version of the Markov chain, rather than the Markov chain itself.

Lemma 3.2. *Let P be the transition matrix of an irreducible Markov chain that is reversible with respect to π, and let Π be the diagonal matrix of π.*

Let $P' = \frac{1}{2}(P + I)$ be the lazy transition matrix for P. Then for any $\nu \in \Delta_n$ where Δ_n is the space of probability distribution, we have

$$||\nu_{t+1} - \Pi||^2_{\Pi^{-\frac{1}{2}}} \leq \left(1 - \frac{\lambda_2(L)}{2}\right) ||\nu_t - \Pi||^2_{\Pi^{-\frac{1}{2}}}.$$

Proof. Let P be the transition matrix of an irreducible Markov chain that is reversible with respect to π, and let Π be the diagonal matrix of π.

Let $P' = \frac{1}{2}(P + I)$ be the lazy transition matrix for P. Let the step size be $\eta = \frac{1}{2}$ and the objective function be $f(\nu) = \frac{1}{2}\nu^T \Pi^{-1}(I - P^T)\nu$ with the weighted inner product $\langle \cdot, \cdot \rangle_{\Pi^{-\frac{1}{2}}}$. First, the gradient of $f(\nu)$ with respect to ν is $\nabla f(\nu) = \Pi^{-1}(I - P^T)\nu$ and the update rule for gradient descent is

$$\begin{aligned}
\nu_{t+1} &= \nu_t - \frac{1}{2}(\Pi^{-T/2}\Pi^{-1/2})^{-1}\nabla f(\nu_t) \\
&= \nu_t - \frac{1}{2}\Pi\nabla f(\nu_t) = \nu_t - \frac{1}{2}\Pi\Pi^{-1}(I - P^T)\nu_t \\
&= \nu_t - \frac{1}{2}(I - P^T)\nu_t = (I - \frac{1}{2}(I - P^T))\nu_t \\
&= \left(\frac{1}{2}(I + P^T)\right)\nu_t
\end{aligned}$$

Hence, the update rule for gradient descent on the objective function can be rewritten

$$\nu_{t+1} = (P')^T \nu_t.$$

Second, we will prove that the objective function is μ-strongly convex and L-smooth. Let $A = \Pi^{-1}(I - P^T)$. The smallest positive eigenvalue of A corresponds to μ and the largest eigenvalue of A corresponds to L. Now, since $A = \Pi^{-1}(I - P^T) = \Pi^{-\frac{1}{2}}(I - P^T)\Pi^{\frac{1}{2}}$, it follows that the eigenvalues are the same as those of L. Therefore, μ corresponds to the smallest positive eigenvalue of L, $\lambda_2(L)$ and L corresponds to the largest eigenvalue of L, 2. Hence, it follows that $f(\nu)$ is $\lambda_2(L)$-strongly convex and 2-smooth. Applying Lemma 2.15, using $\eta = \frac{1}{2} = \frac{1}{L}$, it follows that

$$\begin{aligned} ||\nu_{t+1} - \pi||^2_{\Pi^{-\frac{1}{2}}} &\leq (1 - \eta\mu)||\nu_t - \pi||^2_{\Pi^{-\frac{1}{2}}} \\ &= \left(1 - \frac{\lambda_2(L)}{2}\right)||\nu_t - \pi||^2_{\Pi^{-\frac{1}{2}}}. \end{aligned}$$

□

Therefore, this lemma shows that the weighted distance to the stationary distribution decreases geometrically, with a contraction factor dependent on the spectral gap $\lambda_2(L)$.

Now, we will prove the key theorem of this paper, but first, we will give intermediate lemmas that lead to this theorem.

Theorem 3.3 (Spectral Mixing Time Bound). *For any starting starting distribution on the states $\nu_0 \in \Delta$, after running the Markov chain with the lazy transition matrix P' for $t = \frac{4\log\left(\frac{1}{\epsilon\sqrt{\pi(x^*)}}\right)}{\lambda_2(L)}$ steps, we have $||\nu_t - \pi||_1 \leq \epsilon$.*

First, we will relate the total variation distance, the more common measure of distance, to the distance in the $\Pi^{-\frac{1}{2}}$ norm, the one used in this paper.

Lemma 3.4. *Let ν_t be the distribution on the states after running the 'lazy' Markov chain t times. Then,*

$$||\nu_t - \pi||_1 \leq ||\nu_t - \pi||_{\Pi^{-\frac{1}{2}}}.$$

Proof. From the Cauchy-Schwarz inequality and the definition of the total variation distance, we know that

$$\begin{aligned} ||\nu_t - \pi||_1 = \sum_x |\nu_t(x) - \pi(x)| &\leq \sqrt{\sum_x \frac{(\nu_t(x) - \pi(x))^2}{\pi(x)} \sum_x \pi(x)} \\ &= \sqrt{\sum_x \frac{(\nu_t(x) - \pi(x))^2}{\pi(x)}} \qquad \text{since } \sum_x \pi(x) = 1 \\ &= ||\nu_t - \pi||_{\Pi^{-\frac{1}{2}}}. \end{aligned}$$

□

Now, we will bound the $\|\nu_0 - \pi\|^2_{\Pi^{-1/2}}$ distance by the worst-case starting distance, which happens for distributions that are concentrated at the least likely state. In other words,

Lemma 3.5. *Let ν_0 be any distribution on the states and x^* be $x^* = \arg\min_x \pi(x)$. Then,*

$$||\nu_0 - \pi||^2_{\Pi^{-1/2}} \leq \frac{1}{\pi(x^*)}.$$

Proof. For any state x, write $\nu_0(x) = \sum_y \nu_0(y)\mathbf{1}_{\{x=y\}}$, where $\mathbf{1}_{\{x=y\}}$ means 1 when $x = y$ and 0 otherwise.

Now, Jensen's inequality states that for any function ϕ that is convex, then for any probability distribution ν_0, the inequality holds:

$$\phi(\sum_y \nu_0(y)z_y) \leq \sum_y \nu_0(y)\phi(z_y).$$

Now, consider the convex function $\phi(z) = (z-1)^2$ and $z_y = \frac{\mathbf{1}_{\{x=y\}}}{\pi(x)}$. By Jensen's inequality and convexity of $\phi(z) = (z-1)^2$, we get

$$\left(\frac{\sum_y \nu_0(y)\mathbf{1}_{\{x=y\}}}{\pi(x)} - 1\right)^2 \leq \sum_y \nu_0(y)\left(\frac{\mathbf{1}_{\{x=y\}}}{\pi(x)} - 1\right)^2.$$

Now, since $\|\pi - \nu_0\|^2_{\Pi^{-1/2}} = \sum_x \pi(x)(\frac{\nu_0(x)}{\pi(x)} - 1)^2$, and by the Jensen inequality we know that

$$\left(\frac{\nu_0(x)}{\pi(x)} - 1\right)^2 \leq \sum_y \nu_0(y)\left(\frac{\mathbf{1}_{\{x=y\}}}{\pi(x)} - 1\right)^2,$$

it follows that

$$\sum_x \pi(x)\left(\frac{\nu_0(x)}{\pi(x)} - 1\right)^2 \leq \sum_x \pi(x)\sum_y \nu_0(y)\left(\frac{\mathbf{1}_{\{x=y\}}}{\pi(x)} - 1\right)^2.$$

Changing the order of summation we get

$$\sum_x \pi(x)\sum_y \nu_0(y)\left(\frac{\mathbf{1}_{\{x=y\}}}{\pi(x)} - 1\right)^2 = \sum_y \nu_0(y)\sum_x \pi(x)\left(\frac{\mathbf{1}_{\{x=y\}}}{\pi(x)} - 1\right)^2 = \sum_y \nu_0(y)\frac{1}{\pi(y)}.$$

Now, to bound this expression more clearly, we can bound the sum by the inverse of the smallest $\pi(x)$, $\pi(x^*)$, and we get $\sum_y \nu_0(y)\frac{1}{\pi(y)} \leq \frac{1}{\pi(x^*)}$.

Therefore,

$$||\nu_0 - \pi||^2_{\Pi^{-1/2}} \leq \frac{1}{\pi(x^*)}.$$

□

Proof of Theorem 3.3. Following Claim 3.1, we know that the lazy transition state P' satisfies

$$||\nu_{t+1} - \Pi||^2_{\Pi^{-\frac{1}{2}}} \leq \left(1 - \frac{\lambda_2(L)}{2}\right)||\nu_t - \Pi||^2_{\Pi^{-\frac{1}{2}}}.$$

Hence, after t steps

$$||\nu_t - \Pi||^2_{\Pi^{-\frac{1}{2}}} \leq \left(1 - \frac{\lambda_2(L)}{2}\right)^t ||\nu_0 - \Pi||^2_{\Pi^{-\frac{1}{2}}}.$$

This implies that

$$||\nu_t - \Pi||_{\Pi^{-\frac{1}{2}}} \leq \left(1 - \frac{\lambda_2(L)}{2}\right)^{\frac{t}{2}} ||\nu_0 - \Pi||_{\Pi^{-\frac{1}{2}}}.$$

Taking this conclusion with Lemma 3.4, it follows that

$$||\nu_t - \pi||_1 \leq ||\nu_t - \pi||_{\Pi^{-\frac{1}{2}}} \leq \left(1 - \frac{\lambda_2(L)}{2}\right)^{\frac{t}{2}} ||\nu_0 - \Pi||_{\Pi^{-\frac{1}{2}}}.$$

Since we know that $||\nu_0 - \pi||^2_{\Pi^{-1/2}} \leq \frac{1}{\pi(x^*)}$ from Lemma 3.5 and that $||\nu_t - \pi||_1 \leq \left(1 - \frac{\lambda_2(L)}{2}\right)^{\frac{t}{2}} ||\nu_0 - \Pi||_{\Pi^{-\frac{1}{2}}}$ from above, it follows that

$$||\nu_t - \pi||_1 \leq \left(1 - \frac{\lambda_2(L)}{2}\right)^{\frac{t}{2}} \frac{1}{\pi(x^*)}.$$

Now, to ensure that $||\nu_t - \pi||_1 \leq \epsilon$, we want

$$\frac{(1 - \frac{\lambda_2(L)}{2})^{\frac{t}{2}}}{\sqrt{\pi(x^*)}} \leq \epsilon.$$

First, multiply both sides by $\sqrt{\pi(x^*)}$ to get,

$$\left(1 - \frac{\lambda_2(L)}{2}\right)^{\frac{t}{2}} \leq \epsilon\sqrt{\pi(x^*)}.$$

Now, taking the log of both sides, we get $\frac{t}{2}\log(1 - \frac{\lambda_2(L)}{2}) \leq \log(\epsilon\sqrt{\pi(x^*)})$ and since $\log(1-x) \leq -x$ for $x < 1$, we get

$$\begin{aligned} \frac{t}{2}\left(-\frac{\lambda_2(L)}{2}\right) &\leq \log(\epsilon\sqrt{\pi(x^*)}) \\ -\frac{t\lambda_2(L)}{4} &\leq \log(\epsilon\sqrt{\pi(x^*)}) \\ t &\geq \frac{4\log(\frac{1}{\epsilon\sqrt{\pi(x^*)}})}{\lambda_2(L)} \end{aligned}$$

□

Thus, the mixing time can be achieved by the convergence bound of the gradient descent on the objective function in relation to the Markov chain.

4. The error of approximation

Now, we will explore why in Section 3 we used the $\Pi^{-\frac{1}{2}}$ norm instead of the 2 norm.

Existing proofs of the spectral mixing time bound use the $\Pi^{-\frac{1}{2}}$ norm, and we can modify the usual gradient descent convergence bound to accommodate the $\Pi^{-\frac{1}{2}}$ norm. This allows us, with the right choice of objective function, to run a gradient descent that simulates a Markov Chain. The way that we use the $\Pi^{-\frac{1}{2}}$ norm is reminiscent of a technique in optimization called preconditioning.

However, the intuition for the reasoning can be as follows:.

(1) Optimization

From the perspective of gradient descent, changing the norm induced by the inner product norm of the gradient descent is equivalent to doing regular gradient descent in different coordinates. Since the performance of gradient descent depends on the conditioning of the function we are optimizing, changing the norm of the gradient descent can make the problem better conditioned so we get faster convergence. In this particular case, the best conditioning was achieved with the $\Pi^{-\frac{1}{2}}$ norm.

(2) Sampling

The $\Pi^{-\frac{1}{2}}$ norm is seen when showing spectral mixing time bounds using χ^2 distance, which is equivalent to the distance in the $\Pi^{-\frac{1}{2}}$ norm and hence is a natural norm that may be suitable (and was) for gradient descent. This is because gradient descent convergence bound guarantees that $||\nu_t - \nu^*||$ is decreasing in a particular norm and it should be the norm we want to decrease.

5. Acknowledgements

It is a pleasure to thank my mentor, Max Ovsiankin, for his endless support. I truly could not have written my paper if it weren't for Max.

6. References

[BH14] Joseph Blitzstein and Jessica Hwang. *Introudction to Probability.* CRC Press, 2014.

[Faw13] Hamza Fawzi. "Lecture 3". In: *Topics in Convex Optimization Lecture Notes (University of Cambridge)* (2013). URL: `https://www.damtp.cam.ac.uk/user/hf323/L23-III-OPT/lecture3.pdf`.

[Gow18] Robert M Gower. "Convergence Theorems for Gradient Descent". In: *Convergence Theorems for Gradient Descent* (2018). URL: `https://perso.telecom-paristech.fr/rgower/pdf/M2_statistique_optimisation/grad_conv.pdf`.

[Gup19] Anupam Gupta. "Lecture 20 - Gradient Descent". In: *Lecture notes for 15-451/651: Design & Analysis of Algorithms (Carnegie Mellon University)* (2019). URL: `https://www.cs.cmu.edu/~15451-s19/lectures/lec20-gradient.pdf`.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17.

[Kak15] Sham Kakade. "Optimization 1: Gradient Descent". In: *Lecture Notes for CSE 546: Machine Learning (University of Washington)* (2015). URL: `https://www.columbia.edu/~ks20/stochastic-I/stochastic-I-Time-Reversibility.pdf`.

[Ove23] Shayan Oveis Gharan. "Lecture 15 - Introduction to Markov Chains". In: *Lecture notes for CSE 525 Randomized Algorithms (University of Washington)* (2023). URL: `https://courses.cs.washington.edu/courses/cse525/23sp/525-lecture-15.pdf`.

[Rya15] Tibshirani Ryan. "Lecture 6". In: *Lecture Notes for 10-725: Optimization (Carnegie Mellon University)* (2015). URL: `https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf`.

[Sig09] Karl Sigman. "Lecture 5 - Time-reversible Markov chains". In: *Lecture Notes for Stochastic Modeling I (Columbia Univeristy)* (2009). URL: `https://www.columbia.edu/~ks20/stochastic-I/stochastic-I-Time-Reversibility.pdf`.