Lattice Gaussian Sampling with Markov Chain Monte Carlo (MCMC)

Cong Ling
Imperial College London

joint work with Zheng Wang (Huawei Technologies Shanghai)

September 20, 2016

Outline

Lattice Gaussian Distribution

- Lattice

$$\Lambda = \mathcal{L}(\mathbf{B}) = \{\mathbf{Bx} : \mathbf{x} \in \mathbb{Z}^n\}$$

- Continuous Gaussian distribution

$$\rho_{\sigma,\mathbf{c}}(\mathbf{z}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\|\mathbf{z}-\mathbf{c}\|^2}{2\sigma^2}}$$

- Discrete Gaussian distribution over lattice Λ

$$\begin{aligned} D_{\Lambda,\sigma,\mathbf{c}}(\mathbf{x}) &= \frac{\rho_{\sigma,\mathbf{c}}(\mathbf{Bx})}{\rho_{\sigma,\mathbf{c}}(\Lambda)} \\ &= \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}-\mathbf{c}\|^2}}{\sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}-\mathbf{c}\|^2}} \end{aligned}$$

where $\rho_{\sigma,\mathbf{c}}(\Lambda) \triangleq \sum_{\mathbf{Bx}\in\Lambda} \rho_{\sigma,\mathbf{c}}(\mathbf{Bx})$

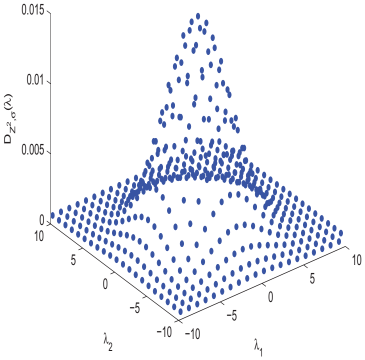


Fig. 1. Discrete Gaussian distribution over $\mathbb{Z}^2$.

Why Does It Matter?

- **Decoding**

The shape of $D_{\Lambda,\sigma,\mathbf{c}}(\mathbf{x})$ suggests that a lattice point $\mathbf{B}\mathbf{x}$ closer to $\mathbf{c}$ will be sampled with a higher probability

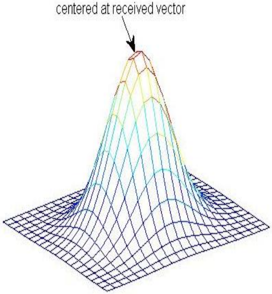
- solve the CVP and SVP problems [Aggarwal et al. 2015, Stephens-Davidowitz 2016]
- decoding of MIMO systems [Liu, Ling and Stehlé 2011]

Closest Vector Problem (CVP)

Given a lattice basis $\mathbf{B} \in \mathbb{R}^{n\times n}$ and a target point $\mathbf{c} \in \mathbb{R}^n$, find the closest lattice point $\mathbf{Bx}$ to $\mathbf{c}$

Shortest Vector Problem (SVP)

Given a lattice basis $\mathbf{B} \in \mathbb{R}^{n\times n}$, find the shortest nonzero vector of $\mathbf{B}$



Why Does It Matter?

- **Mathematics**
  - prove the transference theorem of lattices [Banaszczyk 1993]
- **Coding**
  - obtain the full shaping gain in lattice coding [Forney, Wei 1989, Kschischang, Pasupathy 1993]
  - capacity achieving distribution in information theory: Gaussian channel [Ling, Belfiore 2013], Gaussian wiretap channel [Ling, Luzzi, Belfiore and Stehlé 2013], fading and MIMO channels [Campello, Ling, Belfiore 2016]...
- **Cryptography**
  - propose lattice-based cryptosystems based on the worst-case hardness assumptions [Micciancio, Regev 2004]
  - underpin the fully-homomorphic encryption for cloud computing [Gentry 2009]

How to sample from lattice Gaussian distribution?

The problem that lattice Gaussian sampling aims to solve

State of the Art

- Klein's algorithm [Klein 2000]: works if

$$\sigma \geq \omega(\sqrt{\log\, n}) \cdot \max_{1 \leq i \leq n} \|\widehat{\mathbf{b}}_i\|$$

  where $\widehat{\mathbf{b}}_i$'s are Gram-Schmidt vectors [Gentry, Peikert, Vaikuntanathan 2008].
- Aggarwal et al. 2015: works for arbitrary σ, exponential time, exponential space.
- Markov chain Monte Carlo [Wang, Ling 2014]: arbitrary σ, polynomial space, how much time?
- For special lattices: Construction A, B etc., very fast [Campeloo, Belfiore 2016]; polar lattices, quasilinear complexity [Yan et al.'14].

Klein Sampling

By sequentially sampling from the 1-dimension conditional Gaussian distribution $D_{\mathbb{Z},\sigma_i,\widetilde{x}_i}$ in a backward order from x_n to x_1, the probability of Klein is

$$P_{\text{Klein}}(\mathbf{x}) = \prod_{i=1}^{n} D_{\mathbb{Z},\sigma_i,\widetilde{x}_i}(x_i) = \frac{\rho_{\sigma,\mathbf{c}}(\mathbf{Bx})}{\prod_{i=1}^{n} \rho_{\sigma_i,\widetilde{x}_i}(\mathbb{Z})} \tag{1}$$

Klein's Algorithm

Input: $\mathbf{B}, \sigma, \mathbf{c}$
Output: $\mathbf{Bx} \in \Lambda$

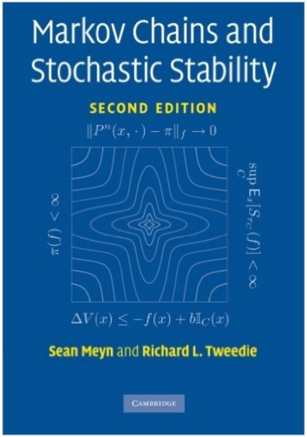
1. let $\mathbf{B} = \mathbf{QR}$ and $\mathbf{c}' = \mathbf{Q}^T\mathbf{c}$
2. **for** $i = n, \ldots, 1$ **do**
3. let $\widetilde{x}_i = \frac{c_i' - \sum_{j=i+1}^{n} r_{i,j} x_j}{r_{i,i}}$, $\sigma_i = \frac{\sigma}{|r_{i,i}|}$
4. sample x_i from $D_{\mathbb{Z},\sigma_i,\widetilde{x}_i}$
5. **end for**
6. return $\mathbf{Bx}$

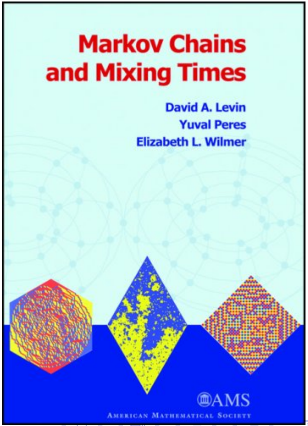
- $P_{\text{Klein}}(\mathbf{x})$ has been demonstrated in [GPV,2008] to be close to $D_{\Lambda,\sigma,\mathbf{c}}(\mathbf{x})$ within a negligible statistical distance if

  $$\sigma \geq \omega(\sqrt{\log n}) \cdot \max_{1 \leq i \leq n} \|\widehat{\mathbf{b}}_i\|$$

- The operation of Klein's algroithm has polynomial complexity $O(n^2)$ excluding QR decomposition

- Markov chain Monte Carlo (MCMC) methods were introduced into lattice Gaussian sampling for the range of σ beyond the reach of Klein's algorithm [Wang, Ling and Hanrot, 2014].
- MCMC methods attempt to sample from an intractable target distribution of interest by building a Markov chain, which randomly generates the next sample conditioned on the previous sample.





Gibbs Sampling

Gibbs Sampling

At each Markov move, perform sampling over a single component of $\mathbf{x}$

$$P(x_i^{t+1}|\mathbf{x}_{[-i]}^t) \quad = \quad \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{B}\mathbf{x}^{t+1}-\mathbf{c}\|^2}}{\sum_{x_i^{t+1}\in\mathbb{Z}} e^{-\frac{1}{2\sigma^2}\|\mathbf{B}\mathbf{x}^{t+1}-\mathbf{c}\|^2}}$$

where $\mathbf{x}_{[-i]}^t = (x_1^t, ..., x_{i-1}^t, x_{i+1}^t, ..., x_n^t)$.

Gibbs-Klein Sampling Algorithm [Wang, Ling and Hanrot, 2014]

At each Markov move, perform the sampling over a block of components of $\mathbf{x}$, while keeping the complexity at the same level as that of componentwise sampling

$$P(x_{\text{block}}^{t+1}|\mathbf{x}_{[-\text{block}]}^t) \quad = \quad \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{B}\mathbf{x}^{t+1}-\mathbf{c}\|^2}}{\sum_{x_{\text{block}}^{t+1}\in\mathbb{Z}^m} e^{-\frac{1}{2\sigma^2}\|\mathbf{B}\mathbf{x}^{t+1}-\mathbf{c}\|^2}}$$

Metropolis-Hastings Sampling

In 1970's, the original Metropolis sampling was extended to a more general scheme known as the Metropolis-Hastings (MH) sampling, which can be summarized as:

- Given the current state $\mathbf{x}$ for Markov chain $\mathbf{X}_t$, a state candidate $\mathbf{y}$ for the next Markov move $\mathbf{X}_{t+1}$ is generated from the proposal distribution $q(\mathbf{x}, \mathbf{y})$
- Then the acceptance decision ratio α about $\mathbf{y}$ is computed

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ 1, \frac{\pi(\mathbf{y}) q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x}) q(\mathbf{x}, \mathbf{y})} \right\}, \tag{2}$$

  where $\pi(\mathbf{x})$ is the target invariant distribution
- $\mathbf{y}$ and $\mathbf{x}$ will be accepted as the state by $\mathbf{X}_{t+1}$ with probability α and $1 - \alpha$, respectively

In MH sampling, $q(\mathbf{x}, \mathbf{y})$ can be any fixed distribution. However, as the dimension goes up, finding a suitable $q(\mathbf{x}, \mathbf{y})$ could be difficult

Independent Metropolis-Hastings-Klein Sampling

- In [Wang, Ling 2015], Klein's sampling is used to generate the state candidate $\mathbf{y}$ for the Markov move $\mathbf{X}_{t+1}$, namely, $q(\mathbf{x}, \mathbf{y}) = P_{\text{Klein}}(\mathbf{y})$

The generation of $\mathbf{y}$ for $\mathbf{X}_{t+1}$ does not depend on the previous state $\mathbf{X}_t$

$q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y})$ is a special case of MH sampling known as independent MH sampling [Tierney, 1991]

Independent MHK sampling algorithm

- Sample from the independent proposal distribution through Klein's algorithm to obtain the candidate state $\mathbf{y}$ for $\mathbf{X}_{t+1}$

$$q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}) = P_{\text{Klein}}(\mathbf{y}) = \frac{\rho_{\sigma,\mathbf{c}}(\mathbf{By})}{\prod_{i=1}^{n} \rho_{\sigma_i, \widetilde{y}_i}(\mathbb{Z})},$$

  where $\mathbf{y} \in \mathbb{Z}^n$
- Calculate the acceptance ratio $\alpha(\mathbf{x}, \mathbf{y})$

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})}\right\} = \min\left\{1, \frac{\pi(\mathbf{y})q(\mathbf{x})}{\pi(\mathbf{x})q(\mathbf{y})}\right\},$$

  where $\pi = D_{\Lambda,\sigma,\mathbf{c}}$
- Make a decision for $\mathbf{X}_{t+1}$ based on $\alpha(\mathbf{x}, \mathbf{y})$ to accept $\mathbf{X}_{t+1} = \mathbf{y}$ or not

Ergodicity

- A Markov chain is ergodic if there exists a limiting distribution $\pi(\cdot)$ such that

$$\lim_{t\to\infty} \|P^t(\mathbf{x};\cdot) - \pi(\cdot)\|_{TV} = 0$$

where $\|\cdot\|_{TV}$ is the total variation distance

- All the afore-mentioned Markov chains are ergodic

Modes of ergodicity

- Polynomial Ergodicity $\quad \|P^t(\mathbf{x};\cdot) - \pi(\cdot)\|_{TV} = M \cdot \frac{1}{f(t)}$
- Uniform Ergodicity $\quad \|P^t(\mathbf{x};\cdot) - \pi(\cdot)\|_{TV} = M(1-\delta)^t$
- Geometric Ergodicity $\quad \|P^t(\mathbf{x};\cdot) - \pi(\cdot)\|_{TV} = M(\mathbf{x})(1-\delta)^t$

$f(t)$ is a polynomial function of t, $M < \infty$, $0 < \delta < 1$

Mixing Time of a Markov Chain

$$t_{\mathrm{mix}}(\epsilon) = \min\{t : \max\|P^t(\mathbf{x},\cdot) - \pi(\cdot)\|_{TV} \le \epsilon\}.$$

Ergodicity of Independent MHK

- The transition probability $P(\mathbf{x},\mathbf{y})$ of the independent MHK algorithm is

$$P(\mathbf{x},\mathbf{y})=q(\mathbf{x},\mathbf{y})\cdot\alpha(\mathbf{x},\mathbf{y})=\begin{cases}\min\left\{q(\mathbf{y}),\frac{\pi(\mathbf{y})q(\mathbf{x})}{\pi(\mathbf{x})}\right\} & \text{if } \mathbf{y}\neq\mathbf{x},\\ q(\mathbf{x})+\sum_{\mathbf{z}\neq\mathbf{x}}\max\left\{0,q(\mathbf{z})-\frac{\pi(\mathbf{z})q(\mathbf{x})}{\pi(\mathbf{x})}\right\} & \text{if } \mathbf{y}=\mathbf{x}.\end{cases} \tag{3}$$

Lemma 1

Given the invariant distribution $D_{\Lambda,\sigma,\mathbf{c}}$, the Markov chain induced by the independent MHK algorithm is ergodic

$$\lim_{t\to\infty}\|P^t(\mathbf{x};\cdot)-D_{\Lambda,\sigma,\mathbf{c}}\|_{TV}=0$$

for all states $\mathbf{x}\in\mathbb{Z}^n$.

- For a countably infinite state space Markov chain, ergodicity is achieved by irreducibility, aperiodicity and reversibility

Proof:

The Markov chain produced by the proposed algorithm is inherently reversible $(\mathbf{x}\neq\mathbf{y})$

$$\begin{aligned}\pi(\mathbf{x})P(\mathbf{x},\mathbf{y}) &= \pi(\mathbf{x})q(\mathbf{x},\mathbf{y})\alpha(\mathbf{x},\mathbf{y})\\ &= \min\{\pi(\mathbf{x})q(\mathbf{y}),\pi(\mathbf{y})q(\mathbf{x})\}\\ &= \pi(\mathbf{y})P(\mathbf{y},\mathbf{x})\end{aligned} \tag{4}$$

Uniform Ergodicity

Lemma 2

In the independent MHK algorithm for lattice Gaussian sampling, there exists a constant $\delta > 0$ such that

$$\frac{q(\mathbf{x})}{\pi(\mathbf{x})} \geq \delta$$

for all $\mathbf{x} \in \mathbb{Z}^n$.

Proof:

$$\begin{aligned} \frac{q(\mathbf{x})}{\pi(\mathbf{x})} &= \frac{\rho_{\sigma,\mathbf{c}}(\mathbf{Bx})}{\prod_{i=1}^{n} \rho_{\sigma_i,\widetilde{x}_i}(\mathbb{Z})} \cdot \frac{\rho_{\sigma,\mathbf{c}}(\Lambda)}{\rho_{\sigma,\mathbf{c}}(\mathbf{Bx})} \\ &= \frac{\rho_{\sigma,\mathbf{c}}(\Lambda)}{\prod_{i=1}^{n} \rho_{\sigma_i,\widetilde{x}_i}(\mathbb{Z})} \\ &\geq \frac{\rho_{\sigma,\mathbf{c}}(\Lambda)}{\prod_{i=1}^{n} \rho_{\sigma_i}(\mathbb{Z})}, \end{aligned} \tag{5}$$

where the right-hand side (RHS) of (5) is completely independent of $\mathbf{x}$

Uniform Ergodicity

Theorem 1

Given the invariant lattice Gaussian distribution $D_{\Lambda,\sigma,\mathbf{c}}$, the Markov chain induced by the independent MHK algorithm is uniformly ergodic:

$$\|P^t(\mathbf{x},\cdot) - D_{\Lambda,\sigma,\mathbf{c}}(\cdot)\|_{TV} \leq (1-\delta)^t$$

for all $\mathbf{x} \in \mathbb{Z}^n$, where δ is given in Lemma 2.

Proof:

- Based on Lemma 2, we have
$$P(\mathbf{x},\mathbf{y}) \geq \delta\pi(\mathbf{y}). \tag{6}$$
According to coupling technique, every Markov move gives probability at least δ of making $\mathbf{X}$ and $\mathbf{X}'$ equal,
$$P(\mathbf{X} = \mathbf{X}') \geq \delta. \tag{7}$$
- Therefore, during t consecutive Markov moves, the probability of $\mathbf{X}$ and $\mathbf{X}'$ not equaling each other can be derived as
$$P(\mathbf{X}_t \neq \mathbf{X}'_t) = (1 - P(\mathbf{X} = \mathbf{X}'))^t \leq (1-\delta)^t \tag{8}$$
- By invoking the coupling inequality, we have
$$\|P^t(\mathbf{x},\cdot) - \pi(\cdot)\|_{TV} \leq P(\mathbf{X}_t \neq \mathbf{X}'_t) \leq (1-\delta)^t \tag{9}$$

Convergence Parameter δ (when $\mathbf{c} = \mathbf{0}$)

Analysis of the Convergence Parameter δ

In the case of $\mathbf{c} = \mathbf{0}$, δ can be expressed by Theta series Θ_Λ and Jacobi theta function ϑ_3 as

$$
\begin{aligned}
\frac{q(\mathbf{x})}{\pi(\mathbf{x})} &= \frac{\rho_{\sigma,\mathbf{0}}(\Lambda)}{\prod_{i=1}^{n} \rho_{\sigma_i,\widetilde{x}_i}(\mathbb{Z})} \\
&\geq \frac{\sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{B}\mathbf{x}\|^2}}{\prod_{i=1}^{n} \rho_{\sigma_i}(\mathbb{Z})} \\
&= \frac{\Theta_\Lambda(\frac{1}{2\pi\sigma^2})}{\prod_{i=1}^{n} \vartheta_3(\frac{1}{2\pi\sigma_i^2})} = \delta \qquad (10)
\end{aligned}
$$

- Theta series Θ_Λ and Jacobi theta function ϑ_3 are

$$\Theta_\Lambda(\tau) = \sum_{\lambda\in\Lambda} e^{-\pi\tau\|\lambda\|^2}, \qquad (11)$$

$$\vartheta_3(\tau) = \sum_{n=-\infty}^{+\infty} e^{-\pi\tau n^2} \qquad (12)$$

with $\Theta_{\mathbb{Z}} = \vartheta_3$

Convergence Parameter δ (when $\mathbf{c} = \mathbf{0}$)

- Given the lattice basis $\mathbf{B}$, the value of δ can be calculated

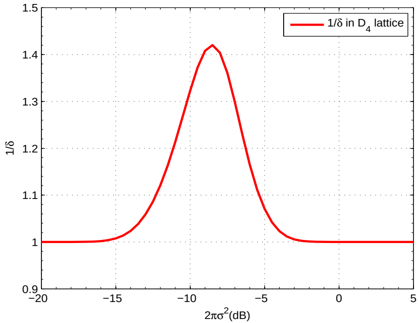


Fig. 2. The value of $\frac{1}{\delta}$ of the D_4 lattice in the case of $\mathbf{c} = \mathbf{0}$.

- Therefore, the mixing time of the Markov chain can be estimated by

$$t_{\text{mix}}(\epsilon) = \frac{\ln \epsilon}{\ln(1-\delta)} < (-\ln \epsilon) \cdot \left(\frac{1}{\delta}\right), \quad \epsilon < 1$$

Convergence Parameter δ (when $\mathbf{c} = \mathbf{0}$)

Lemma 3

In the case of $\mathbf{c} = \mathbf{0}$, the coefficient δ for an isodual lattice has a multiplicative symmetry point at $\sigma = \frac{1}{2\pi}$, and converges to 1 on both sides asymptotically when σ goes to 0 and ∞.

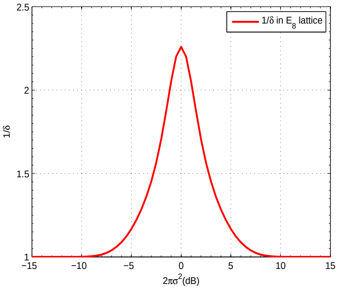


Fig. 3. The value of $\frac{1}{\delta}$ of the E_8 lattice in the case of $\mathbf{c} = \mathbf{0}$.

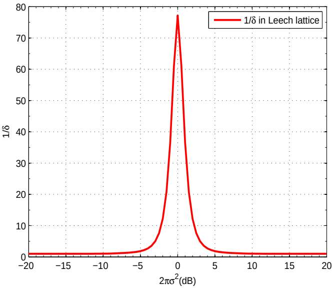


Fig. 4. The value of $\frac{1}{\delta}$ of the Leech lattice in the case of $\mathbf{c} = \mathbf{0}$.

Metropolis-Hastings Sampling

How MH algorithm works?

- Firstly, sample from the proposal density $T(\mathbf{x};\mathbf{y})$ to get a candidate.
- Then, make a decision based on the quantity α to decide whether to accept this candidate as $\mathbf{x}^{t+1}$ or not
$$\alpha = \min\left\{1, \frac{\pi(\mathbf{y})T(\mathbf{y};\mathbf{x})}{\pi(\mathbf{x})T(\mathbf{x};\mathbf{y})}\right\} \tag{13}$$
where $\pi(\cdot)$ denotes the target distribution.

- **The art of MH algorithm lies in choosing an appropriate proposal density**.
- One can use Klein's algorithm to generate the proposal density, which turns out to be symmetric: $T(\mathbf{x};\mathbf{y}) = T(\mathbf{y};\mathbf{x})$.

Here,
$$T(\mathbf{x}^t;\mathbf{x}^*) = \frac{1}{\prod_{i=1}^{n} \rho_{\sigma_i, x_i^t}(\mathbb{Z})} e^{-\frac{1}{\sigma^2}\|\mathbf{B}\mathbf{x}^t - \mathbf{B}\mathbf{x}^*\|^2}, \tag{14}$$
where $\rho_{\sigma,\mathbf{c}}(\mathbf{z}) = e^{-\frac{\|\mathbf{z}-\mathbf{c}\|^2}{2\sigma^2}}$ is a symmetrical Gaussian function. Then the acceptance ratio α
$$\alpha = \min\left\{1, \frac{\pi(\mathbf{x}^*)}{\pi(\mathbf{x}^t)}\right\} = \min\left\{1, e^{-\frac{1}{2\sigma^2}(\|\mathbf{c}-\mathbf{B}\mathbf{x}^*\|^2 - \|\mathbf{c}-\mathbf{B}\mathbf{x}^t\|^2)}\right\}. \tag{15}$$

Geometric Ergodicity

Definition

A Markov chain having stationary distribution $\pi(\cdot)$ is geometrically ergodic if there exists $0 < \delta < 1$ and $M(\mathbf{x}) < \infty$ such that for all $\mathbf{x}$

$$\|P^t(\mathbf{x}, \cdot) - \pi(\cdot)\|_{TV} \leq M(\mathbf{x})(1-\delta)^t.$$

- In MCMC, the drift condition is the usual way to prove the geometric ergodicity [Roberts and Tweedie 1996]

Definition

A Markov chain with discrete state space Ω satisfies the drift condition if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \Omega \to [1, \infty]$, such that

$$\sum_{\Omega} P(\mathbf{x}, \mathbf{y}) V(\mathbf{y}) \leq \lambda V(\mathbf{x}) + b\mathbf{1}_C(\mathbf{x})$$

for all $\mathbf{x} \in \Omega$, where C is a small set, $\mathbf{1}_C(\mathbf{x})$ equals to 1 when $\mathbf{x} \in C$ and 0 otherwise.

Geometric Ergodicity

Theorem

Given the invariant lattice Gaussian distribution $D_{\Lambda,\sigma,\mathbf{c}}$, the Markov chain established by the symmetric Metropolis-Hastings algorithm satisfies the drift condition. Therefore, it is geometrically ergodic.

- Overall, exponential convergence can be interpreted in two folds
  - when $x_i \notin C$, the Markov chain shrinks geometrically towards the small set C
  - when $x_i \in C$, the Markov chain converges exponentially fast to the stationary
  - there is a trade-off between these two convergence rates depending on the size of C
- However, as C is determined artificially, such a tradeoff is hard to analyze
- The small set C means that there exist $k > 0$, $1 > \delta > 0$ and a probability measure v on Ω such that

$$P^k(\mathbf{x}, \mathcal{B}) \geq \delta v(\mathcal{B}), \quad \forall \mathbf{x} \in C$$

  for all measurable subsets $\mathcal{B} \subseteq \Omega$
- For independent MHK, the entire state space Ω is small.

Open Questions

- Fast convergence requires strong lattice reduction. How to integrate MCMC and lattice reduction?
- Is the complexity of MCMC exponential or super-exponential? It would be a breakthrough even if it is exponential.
- What about other MCMC algorithms? How to design fast-mixing MCMC for discrete Gaussian sampling?
- What about quantum algorithms?


T H A N K Y O U