My Research
Markov Chain Monte Carlo for High-Dimensional Problems

Jun Yang

January 28, 2025

Markov Chain Monte Carlo (MCMC) methods have become essential tools for solving complex problems in mathematics, statistics, and data science. These algorithms allow us to approximate solutions to problems where direct computation is infeasible, especially in high-dimensional settings. The core idea of MCMC is to simulate a Markov chain in which the stationary distribution is designed to be the target distribution you want to sample from. Under certain conditions, by running the Markov chain for a sufficiently long time, it converges to its stationary distribution. At this point, the output of the algorithm can be used as approximate samples from the target distribution. My research focuses on improving both the theory and methodology of MCMC, particularly for high-dimensional applications.

When we use an MCMC algorithm, the process can be divided into two main phases:

1. *Transient Phase (Burn-In):* This is the initial phase of the algorithm, where the Markov chain "warms up" and moves towards its target distribution. A key question here is: *How many iterations do we need to run before the chain reaches the target distribution?* This is related to the concept of *mixing time*, which measures how quickly the Markov chain converges.

2. *Stationary Phase:* Once the chain has reached its target distribution, we use the samples to approximate quantities of interest. In this phase, the focus shifts to understanding *how many samples we need to simulate*. This is related to the chain's ability to efficiently explore the state space, often measured using concepts like the *effective sample size* (ESS).

For developing new methods, an additional challenge arises from understanding and exploring the *concentration of measure* property of the target distribution in high dimensions. This property is often counter-intuitive and connects to surprising facts about high-dimensional geometry, such as the tendency for most of the probability mass to concentrate in a small region. My collaborators and I focus on addressing these challenges in our theoretical and methodological work.

Theory of high-dimensional MCMC

In our paper *"Complexity Results for MCMC Derived from Quantitative Bounds"* (Annals of Applied Probability, 2023), my collaborators and I propose a new approach to studying mixing times. Traditional methods often struggle to provide tight bounds in high-dimensional settings. We developed a technique called the "modified drift-and-minorization" approach, which focuses on identifying and handling challenging regions of the state space that slow down convergence. Using

this method, we analyzed several Gibbs samplers and showed that their mixing times can remain manageable even as the dimensionality increases.

Our second paper, *"Gaussian Approximation and Output Analysis for High-Dimensional MCMC"* (arXiv:2407.05492, under review), explores how to assess the quality of MCMC output in high dimensions. Specifically, we studied how well the sample averages of an MCMC algorithm can be approximated by a Gaussian distribution and how this approximation error depends on the dimensionality of the problem.

New high-dimensional MCMC algorithms

In the paper *"Dimension-Free Mixing for High-Dimensional Bayesian Variable Selection"* (Journal of the Royal Statistical Society: Series B, 2022), my collaborators and I developed an MCMC algorithm tailored for Bayesian variable selection. Traditional MCMC methods often struggle with high-dimensional variable selection problems because they rely on random local moves that can become inefficient as the number of variables increases. We proposed a novel algorithm that uses an informed proposal scheme to make more effective moves. This informed MCMC is designed using a new proof technique called the *two-stage drift condition*, which captures the concentration region of the posterior distribution.

Our most recent work, *"Stereographic Markov Chain Monte Carlo"* (Annals of Statistics, 2024), tackles a common challenge in high-dimensional MCMC: handling heavy-tailed target distributions. Standard MCMC algorithms often perform poorly in these settings because they struggle to explore unbounded state spaces efficiently. To address this, we introduced a new class of MCMC algorithms that map the original problem in Euclidean space onto a sphere. This transformation leverages the compactness and the concentration property of the high-dimensional sphere. By exploiting these properties, the proposed samplers achieve faster convergence and better mixing, even for challenging heavy-tailed distributions.

Broader Impact of My Research

My work contributes to both the theoretical understanding and practical implementation of MCMC algorithms. By addressing fundamental questions about mixing times and output analysis, my collaborators and I aim to provide researchers with tools to better understand and optimize their MCMC simulations. At the same time, the new methodologies we develop are designed to tackle real-world problems, from Bayesian variable selection to inference on heavy-tailed distributions.

Ultimately, we hope our research will make MCMC a more accessible and effective tool for statisticians, mathematicians, and data scientists working on high-dimensional problems. Whether you're analyzing complex datasets, developing new statistical models, or exploring uncharted mathematical territories, the advancements in MCMC theory and methodology offer exciting opportunities for discovery.