Log-concave sampling: Metropolis-Hastings algorithms are fast!

Raaz Dwivedi, Yuansi Chen, Martin Wainwright and Bin Yu
{raaz.rsk, yuansi.chen, wainwrig, binyu}@berkeley.edu

Sampling: A fundamental task!
Given a density, with unknown normalization constant, draw (approximate) samples from it
\[ x_1, \ldots, x_t \overset{i.i.d.}{\sim} \pi(x) \propto e^{-f(x)} \]

\begin{itemize}
  \item Integration: \( \int g(x)\pi(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i) \)
  \item Optimization: \( \min_{x \in \mathcal{X}} g(x) \approx \min \{ g(x_1), \ldots, g(x_n) \} \)
\end{itemize}

Set-up and Objectives
Goal: Given access to \( f(x) \) and \( \nabla f(x) \) for any \( x \in \mathbb{R}^d \), derive non-asymptotic mixing-time bound for \( k \)
\[ k_{\min}(\delta) = \min \{ k \mid \|P(x_k) - \pi^*\|_{TV} \leq \delta \} \]
with explicit dependence on \( d, \delta \) and other parameters.

Algorithms: Optimization vs Sampling
Discretization of gradient flow (ordinary differential equation) vs Langevin diffusion (stochastic differential equation):
\[ \begin{align*}
  & \text{Gradient Flow} \\
  & \quad \dot{x}_t = -\nabla f(x_t)dt \\
  & \quad x_{k+1} = x_k - h \nabla f(x_k) \\
  & \text{Gradient Descent} \\
  & \quad \text{minimize this} \\
  & \text{Start at } x_0 \\
  & \text{Find } x_t \\
  & \text{where } h = \text{step size} \quad k = \text{no. of steps} \\
  & \text{Langevin Diffusion} \\
  & \quad \dot{x}_t = -\nabla f(x_t)dt + \sqrt{2}dB_t \\
  & \quad x_{k+1} = x_k - h \nabla f(x_k) + \sqrt{2h}N(0,1) \\
  & \text{Unadjusted Langevin algorithm (ULA)}
\end{align*} \]

Unadjusted Langevin algorithm (ULA)
\begin{itemize}
  \item ULA is biased: \( \|P(x_t) - \pi^*\|_{TV} \to 0 \) but \( \|P(x_k) - \pi^*\|_{TV} \not\to 0 \).
  \item Large \( h \) \( \Rightarrow \) Faster convergence, large asymptotic bias, and, small \( h \) \( \Rightarrow \) slower convergence, small bias.
\end{itemize}

Metropolis adjusted Langevin algorithm
\begin{itemize}
  \item MALA is made up of two steps
    \[ Z = X_k - h\nabla f(X_k) + \sqrt{2h}N(0,1) \]
    (proposal step)
    \[ P(X_{k+1} \leftarrow Z) = \min \left\{ 1, \frac{\pi^*(Z)P(Z \to X_k)}{\pi^*(X_k)P(X_k \to Z)} \right\} \]
    (accept-reject step)
  \item If case of rejection, \( X_{k+1} \leftarrow X_k \).
  \item Accept-reject step \( \Rightarrow \) Detailed balance \( \Rightarrow \) Unbiasedness \( \Rightarrow \) larger step size \( \Rightarrow \) Faster mixing.
\end{itemize}

Numerical Experiments
\begin{itemize}
  \item (a) SLC: \( d \) dependency
  \item (b) SLC: \( \delta \) dependency
  \item (c) WLC: \( d \) dependency
  \item (d) WLC: \( \delta \) dependency
\end{itemize}

Proof Techniques

High conductance implies fast mixing:
\[ \Phi = \int_{\mathcal{A}_1} \frac{P(u \to A_1)\pi^*(u)du}{\pi^*(A_1)} \]
Mixing time: \( k_{\min}(\delta) \leq O(\log(1/\delta)/\Phi^2) \)

Proof Sketch
MALA: Bad conductance sets are far
Isoperimetry
\[ d(A_1 \cup A_2) \text{ is large} \]
Upper bound on mixing time: \( k \)

References