Dimensionally Tight Bounds for Second-Order Hamiltonian Monte Carlo

Oren Mangoubi, Nisheeth K. Vishnoi
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

NeurIPS 2018

Sampling problem

Given a gradient oracle for $F: \mathbb{R}^d \to \mathbb{R}$, sample from the Gibbs distribution: $\boldsymbol{\pi(x) \propto e^{-F(x)}}$

Applications: Optimization (via annealing), computing integrals/volumes, Bayesian inference, molecular dynamics

Some Markov chains used for sampling

of gradient evaluations to sample from smooth, strongly logconcave π (for smoothness/convexity parameters $= \Theta(1)$):

- Random Walk Metropolis: d^2 (conj: d) [Gelman et al. '97]
- Unadjusted Langevin: d [Durmus, Moulines, '16]
- Underdamped Langevin: $d^{1/2}$ [Cheng et al. '17]

Hamilton's equations

Position x, velocity v, potential F

Invariant measure $\boldsymbol{e^{-F(x)} e^{-\frac{1}{2}\|v\|_2^2}}$

$$\frac{dx}{dt} = v, \frac{dv}{dt} = -\nabla F(x)$$

If $x(0) \sim \pi, v(0) \sim N(0, I_d)$, and solutions are computed with low error, can take long steps that (approximately) preserve π

2nd-order Hamiltonian Monte Carlo [Duane et al., '87]

Input: $X_0, \nabla F, T, \eta, s$

Output: X_s which is $\boldsymbol{\varepsilon}$**-close to** $\boldsymbol{\pi}$, for some $\varepsilon > 0$
(i.e., there is $Y \sim \pi$ s.t. $\|X_s - Y\|_2 < \varepsilon$ w.p. $1 - \varepsilon$)

For $\boldsymbol{i = 0, 1 \ldots, s-1}$, do

1. Generate $V_i \sim N(0, I_d)$
2. "Solve" Hamilton's eqs for $(x_0, v_0) = (X_i, V_i)$ for time T:

   For $\boldsymbol{j = 0, \ldots, \frac{T}{\eta} - 1}$, do

   $$x_{j+1} = x_j + \eta v_j - \tfrac{1}{2}\eta^2 \nabla F(x_j)$$
   $$v_{j+1} = v_j - \eta \nabla F(x_j) - \tfrac{1}{2}\eta^2 \frac{\nabla F(x_{j+1}) - \nabla F(x_j)}{\eta}$$

3. Set $X_{i+1} = x_{T/\eta}$

Previous conjectures and bounds for Hamiltonian Monte Carlo (HMC)

- **Informal conjecture:** $d^{1/4}$ gradient evaluations are sufficient for HMC with 2nd-order integrator if F is 1-smooth, 1-strongly convex, with additional bounds on higher-order derivatives [Creutz, '88]
- Metropolis 2nd-order leapfrog HMC requires $\Omega(d^{1/4})$ gradients for Gaussian and other replica product distributions -[Beskos et al. '10]
- $\tilde{O}(d^{1/2})$ gradients sufficient for first-order HMC [Mangoubi, Smith '17]

Main result

Assume: 1. F is m-strongly convex and M-smooth, and let $\kappa := M/m$
2. Lipschitz condition for $L_\infty, r > 0, \mathsf{X} := [\mathsf{X}_1, \ldots, \mathsf{X}_r] \in \mathbb{S}^{dr}$:

$$\left\|\left(\nabla^2 F(y) - \nabla^2 F(x)\right)v\right\|_2 \leq L_\infty \sqrt{r} \|\mathsf{X}^\top (y - x)\|_\infty \times \|\mathsf{X}^\top v\|_\infty$$

Then: $\tilde{O}\left(\max\left(d^{\frac{1}{4}}\kappa^{2.75}, r^{\frac{1}{4}}\kappa^{2.25}\sqrt{L_\infty}\right)\varepsilon^{-1/2}\right)$ gradients are sufficient for 2ndorder HMC to obtain a sample ε-close to π, from a warm start

(We obtain slightly weaker bounds from a cold start)

Application to Bayesian logistic "ridge" regression

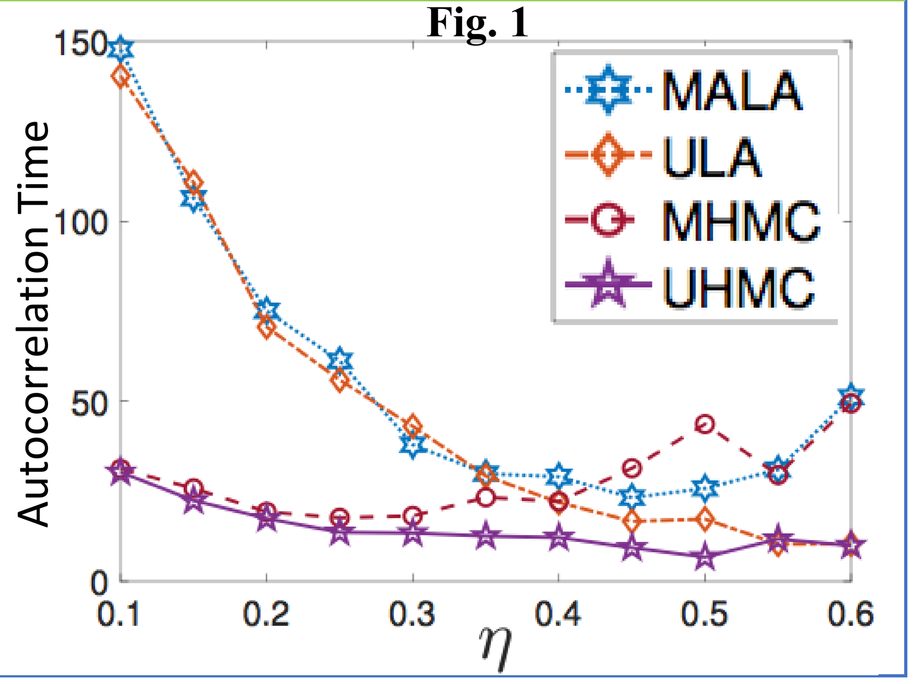
- Given data $(\mathsf{X}_i, \mathsf{Y}_i)$, sample from $\pi(x) \propto e^{-\sum_{i=0}^r f_i(x)}$,
  $f_i(x) = \mathsf{Y}_i \log(\sigma(x^\top \mathsf{X}_i)) + (1 - \mathsf{Y}_i)\log(\sigma(-x^\top \mathsf{X}_i))$,
  where $\sigma(s) = (1 + e^{-s})^{-1}$, prior: $f_0(x) = \|x\|_2^2$
- For logistic regression, $L_\infty = \sqrt{C}$, coherence $C := \max_{i \in [r]} \sum_{j=1}^r |\mathsf{X}_i^\top \mathsf{X}_j|$
- For example, if $r = d$ and $\mathsf{X}_1, \ldots, \mathsf{X}_r \sim \text{uniform}(\mathbb{S}^d)$, # of gradient calls is $\tilde{O}(d^{3/8}\varepsilon^{-1/2})$ from a warm start

Simulations

- Simulations performed on logistic regression, $\mathsf{X}_1, \ldots, \mathsf{X}_d \sim \text{uniform}(\mathbb{S}^d)$ suggest that 2nd order HMC (UHMC) has faster autocorrelation time[1] than Metropolis HMC and Langevin in this setting (Fig. 1)

(1) Autocorrelation is the correlation of points in the Markov chain with a delayed copy of themselves. Autocorrelation time can be estimated as $1 + 2\sum_{s=1}^{s_{\max}} \rho_s$ for some large $s_{\max}$, where ρ_s is autocorrelation with delay s

Fig. 1



Proof highlights

For simplicity, let $M, m = \Theta(1), \varepsilon \leq 1, r = d$. We couple our HMC chain X to an "idealized" HMC chain Y with exact solutions by giving their trajectories the same initial velocity (Fig. 2).

[Mangoubi, Smith '17] show that exact solutions with same initial velocity contract by a constant factor for $T = \Theta(1)$. We extend to 2nd order HMC by showing it approximates exact trajectories with error $O(\varepsilon)$:

- We bound (inductively on j) the errors $\|x_j - x(\eta j)\|_2$ and $\|v_j - v(\eta j)\|_2$ by $O(\eta j \varepsilon)$, where $(x(t), v(t))$ is the continuous solution to Hamilton's eqs with initial conditions (X_i, V_i):
- The error in the quadratic term of the velocity update is roughly $\left\|\left(\eta^2 \nabla^2 F(x + \eta v_j) - \eta^2 \nabla^2 F(x)\right)v_j\right\|_2 \overset{\text{Assumption 2}}{\leq} \eta^3 L_\infty \sqrt{d} \|\mathsf{X}^\top v_j\|_\infty^2$
- The invariance property of Hamiltonian mechanics implies v is roughly $N(0, I_d)$ at every point on the exact trajectory if HMC has a warm start (Fig. 3). Thus, $\|\mathsf{X}^\top v_j\|_\infty = O(\log(d))$ w.h.p., since by inductive assumption $\|v_j - v(\eta j)\|_2 = O(\eta j \varepsilon) = O(1)$
- After T/η iterations, the errors sum to $\tilde{O}(\eta^2 L_\infty \sqrt{r})$. Choosing η to have error ε, # of gradients is $T/\eta = \tilde{\Theta}(\varepsilon^{-1/2} d^{1/4} L_\infty^{1/2})$

Fig. 2: Second-order HMC trajectories approximate exact solutions which contract if given same initial velocity

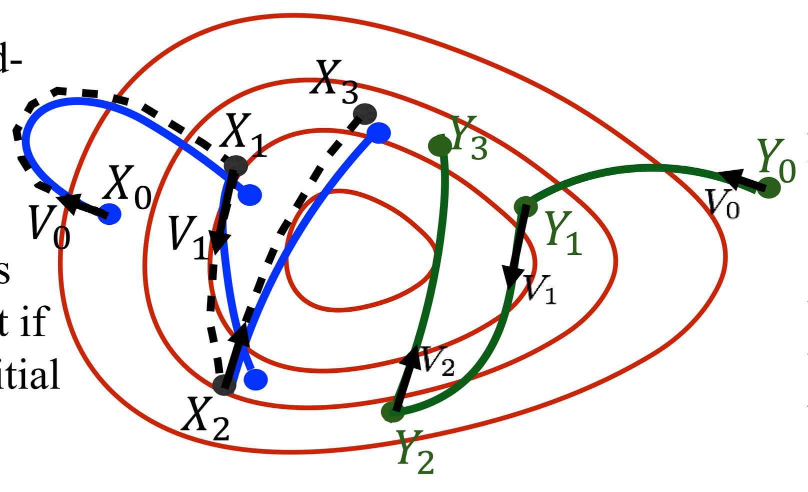
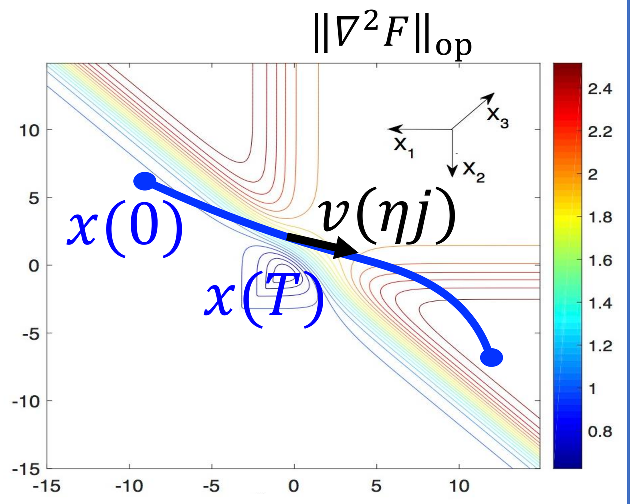


Fig. 3: Given a warm start, exact solutions have roughly $N(0, I_d)$ velocity at every point, meaning they are unlikely to travel in directions where Hessian changes most quickly



Conclusion and future directions

- First faster-than-$\sqrt{d}$ bound for sampling from a large class of logconcave distributions, including logistic regression posteriors
- Can we improve dependence on parameters C and κ?
- Can we generalize to nonconvex F and higher-order integrators?