Hamiltonian Monte Carlo

Aimé Fournier (after Michael Betancourt)

12.S592 Machine Learning with System Dynamics and Optimization
2020 February 28
Outline

● Computing expectations by exploring probability distributions
● Markov chain Monte Carlo
  o Ideal behavior
  o Metropolis-Hastings algorithm
● Foundations of Hamiltonian MC
  o Phase space and Hamilton’s equations
● Efficient HMC
  o Natural geometry of phase space
● Implementing HMC in practice

Goal: estimate probabilistic expectations $\mathbb{E}_\pi[f]$ of functions $f(q)$ on a $D$-dimensional sample space $q \in Q$, w.r.t. a probability distribution $\pi(q)$.

$$\mathbb{E}_\pi[f] = \int_Q f(q)\pi(q) dq$$

Typically $\pi(q)$ decreases quickly for large $|q|$ (assuming mode at $q = 0$), but $dq = \prod_{i=1}^D dq_i \propto |q|^{D-1}d|q|$. The typical set contributes most to $\mathbb{E}_\pi[f]$. 
Markov chain Monte Carlo

- Typical sets may have complicated geometry for high $D$.

- Markov chains $(q_1, \ldots, q_L)$ are the sequences created by Markov transitions $T(q, q')$ on $Q$. If $T$ preserves $\pi(q)$, then $q_L$ approaches the typical set.
Ideal behavior

- Initial chain yields biased estimator \( \hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(q_l) \).
- As chain explores typical set, the bias reduces quickly.
- Further bias reduction takes very long chains.
Metropolis-Hastings algorithm

- Given \( q \), to accomplish \( \mathbb{T} \), propose a random \( q' \) from a symmetric distribution on \( \mathcal{Q}(q, q') \), and accept \( q' \) with probability
\[
\alpha(q'|q) = \min \left(1, \frac{\pi(q')}{\pi(q)}\right)
\]
that rejects relatively improbable steps.

- If the \( \mathcal{Q} \) variance is large, then \( \pi(q') \) will often be small and \( q' \) will rarely be accepted.

- If the \( \mathcal{Q} \) variance is small, \( q' \) will often be accepted but it will take “forever” to explore the typical set.
Foundations of Hamiltonian MC

- There are many more directions obliquely off the typical set than strictly within it. We want the chain to stay in or close to the typical set.
- $\pi(q, p) = \pi(p|q)\pi(q)$ introduces momentum $p$ as an auxiliary parameter so that marginalization projects the phase-space chain down to the desired typical set.
- In physics, energy-conserving dynamics in a phase space of twice as many dimensions $(q, p)$ are constrained to a manifold $H^{-1}(E) = \{(q, p)|H(q, p) = -\log \pi(q, p) = E\}$. 
Phase space and Hamilton’s equations

- The 2D-phase-space trajectory \((q, p) \mapsto \phi_t(q, p)\) is given by
  \[
  \frac{d}{dt} (q, p) = \left( \frac{\partial}{\partial p}, -\frac{\partial}{\partial q} \right) H = \left( \frac{\partial}{\partial p}, -\frac{\partial}{\partial q} \right) K - \left( 0, \frac{\partial V}{\partial q} \right),
  \]
  where \(K = -\log \pi(p|q)\) and \(V = -\log \pi(q) = H - K\) are the effective kinetic and potential energies.
- Then by the chain rule
  \[
  \frac{dH}{dt} = \left( \frac{\partial}{\partial q}, \frac{\partial}{\partial p} \right) H \cdot \frac{d}{dt} (q, p) = 0,
  \]
  energy is conserved by \(\phi_t\) (also see Liouville's theorem).
- Now \(\mathbb{T}(q, q')\) has been decomposed using \(\pi(p|q), \phi_t\) and \(q_i = (q, p)_i\).
- Successive \(\mathbb{T}(q, q') \leftrightarrow \) phase-space Markov chain.
Efficient HMC

● The parameter $t$ and formulation of $K$ provide free parameters to be optimally tuned.

● Longer $t \implies$ more exploration of $H^{-1}(E)$, but also costs more computation and may become redundant after $H^{-1}(E)$ is explored.

● Generally $t = T(q, p)$ should be chosen around when the ESS starts to plateau.

● $\text{ESS} \approx \|\text{(corr matrix)}^{-1}\|_F$, according to Leinster.
The parameter $t$ and form of $K$ provide free parameters to be optimally tuned.

It often makes sense to measure distance in $Q$ using the Mahalanobis norm $(q - q') \cdot M \cdot (q - q')$, where $M = \mathbb{E}_\pi[\mathbf{\cdot} - \mu] \otimes (\mathbf{\cdot} - \mu)^{-1}$ is the precision (inverse covariance) matrix and $\mu = \mathbb{E}_\pi[\mathbf{\cdot}]$.

Then in order to conserve phase-space volume, one should measure momentum differences by $(p - p') \cdot M^{-1} \cdot (p - p')$.

By connecting with zero-mean Gaussian, we’re led to $K = \frac{1}{2} p \cdot M^{-1} \cdot p + \sqrt{\log \det 2\pi M}$.

Including $M = M(q) \approx \frac{\partial}{\partial q} \otimes \frac{\partial}{\partial q} V$, the Hessian can help with variability on $Q$. 
Implementing HMC in practice

- Most numerical integrators create accumulating deviation from $H^{-1}(E)$.
- Symplectic integrators still create error, but by their conservation in phase space, it cannot accumulate. The chain conserves a “shadow Hamiltonian” exactly.
- It can still happen that too coarse a time step $\varepsilon = T/L$ can cause sudden divergence; but it’s a useful indicator of strong $H(q,p)$ curvature.