Probabilistic Path Hamiltonian Monte Carlo

Cheng Zhang

Joint work with Vu Dinh, Arman Bilge and Erick Matsen

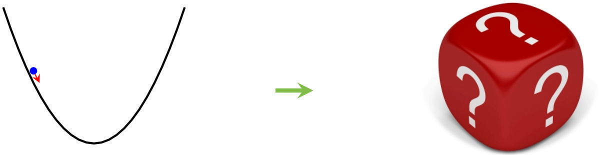
http://matsen.group

Fred Hutchinson Cancer Research Center

University of Washington

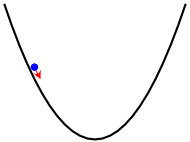
August 1, 2017

Sampling: From Continuous to Discrete Variables



- Advanced MCMCs, e.g. Hamiltonian Monte Carlo, can not handle discrete parameters in general.
- Simple MCMCs usually are not efficient at sampling continuous parameters.

Sampling: From Continuous to Discrete Variables



- Advanced MCMCs, e.g. Hamiltonian Monte Carlo, can not handle discrete parameters in general.
- Simple MCMCs usually are not efficient at sampling continuous parameters.

Question: How to sample from posteriors with both continuous and *structural* discrete parameters efficiently?

Bayesian Learning on Orthant Complexes

An *orthant complex* is a geometric object $\mathcal{X}$ obtained by gluing orthants of the same dimension that share certain boundaries together

$$\mathcal{X} = \{(\tau, q) : \tau \in \Gamma,\ q \in \mathbb{R}^n_{\geq 0}\}$$

where Γ is a countable set. Given observations D and a proper prior $\pi_0(\tau, q)$, the posterior is

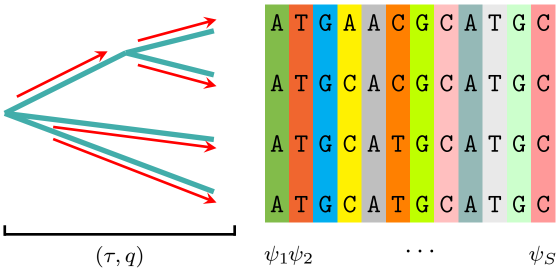
$$P(\tau, q|D) \propto L(D|\tau, q)\pi_0(\tau, q)$$

Assumptions:

- $(\tau, q_\tau) = (\tau', q_{\tau'}) \Rightarrow q_\tau = q_{\tau'},\ \tau' \in \mathcal{N}(\tau, q_\tau)$
- The adjacency graph of $\mathcal{X}$ has finite diameter k.
- $U(\tau, q) = -\log P(\tau, q)$ is continuous and smooth up to the boundary.

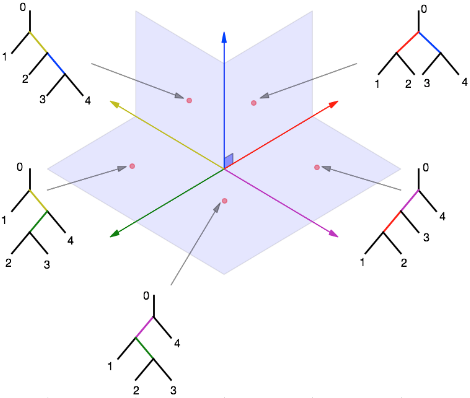
Example: Phylogenetic Inference

Let (τ, q) be a phylogenetic tree and $\psi = \{\psi_i\}_{i=1}^{S}$ be the observed sequences over the leaves.



Goal: reconstruct the evolution history (phylogenetic tree) based on observed sequences.

The Billera-Holmes-Vogtmann Space



The adjacent orthants are called **NNI neighbors**.

Challenges in Phylogenetic Inference

- A continuous-time Markov chain is used to model the evolution history which leads to the following likelihood

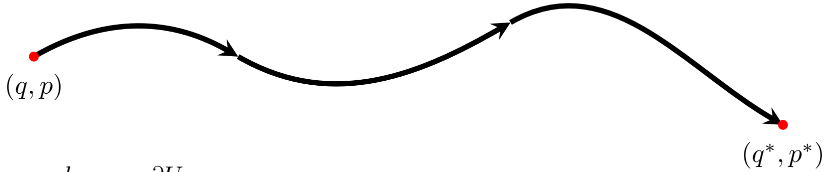
$$L(\psi|\tau, q) = \prod_{s=1}^{S} \sum_{a^s} \eta(a_\rho^s) \prod_{(u,v)\in E(\tau,q)} P_{a_u^s a_v^s}^{uv}(q_{uv})$$

- Efficient computation via Felsenstein's *pruning algorithm* (a.k.a. *belief propagation, sum-product message passing* etc.)
- **Challenging Topology Space:** The number of possible topologies $T(n)$ grows exponentially as the number of leaves n increases

$$T(n) = \frac{(2n-5)!}{(n-3)!\, 2^{n-3}} = e^{\mathcal{O}(n \log n)}$$

Hamiltonian Monte Carlo

$$H(q,p) = U(q) + K(p), \quad K(p) = \frac{1}{2}p^T p$$

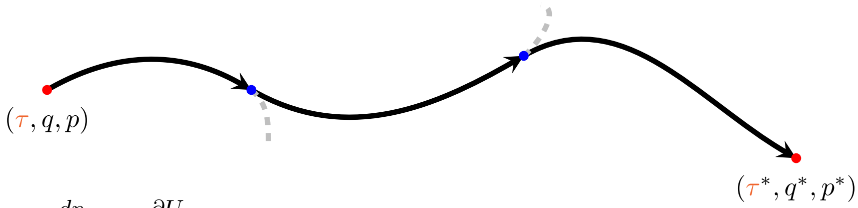


$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}(q)$$

$$\frac{dq_i}{dt} = p_i$$

Probabilistic Path Hamiltonian Monte Carlo

$$H(\tau, q, p) = U(\tau, q) + K(p), \quad K(p) = \frac{1}{2}p^T p$$

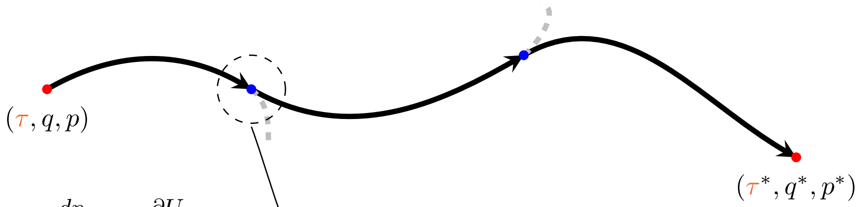


$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}(\tau, q)$$

$$\frac{dq_i}{dt} = p_i$$

Probabilistic Path Hamiltonian Monte Carlo

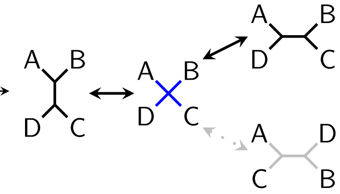
$H(\tau, q, p) = U(\tau, q) + K(p), \quad K(p) = \frac{1}{2}p^T p$



$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}(\tau, q)$$

$p_i = -p_i; \quad \tau \sim Z(\mathcal{N}(\tau, q))$

$$\frac{dq_i}{dt} = p_i$$



Theoretical Properties

Assume *symmetric transition*:

$$P(\tau'|\tau, q) = P(\tau|\tau', q), \quad \tau' \in \mathcal{N}(\tau, q)$$

Augmented state: $s = (\tau, q, p)$, a pair of measurable sets: $A,\ B$.

Theoretical Properties

Assume *symmetric transition*:

$$P(\tau'|\tau, q) = P(\tau|\tau', q), \quad \tau' \in \mathcal{N}(\tau, q)$$

Augmented state: $s = (\tau, q, p)$, a pair of measurable sets: A, B.

- **Probabilistic Reversibility.**

$$P((\tau, q, p), (\tau^*, q^*, p^*)) = P((\tau^*, q^*, -p^*), (\tau, q, -p))$$

Theoretical Properties

Assume *symmetric transition*:

$$P(\tau'|\tau, q) = P(\tau|\tau', q), \quad \tau' \in \mathcal{N}(\tau, q)$$

Augmented state: $s = (\tau, q, p)$, a pair of measurable sets: $A,\ B$.

- **Probabilistic Reversibility.**

$$P((\tau, q, p), (\tau^*, q^*, p^*)) = P((\tau^*, q^*, -p^*), (\tau, q, -p))$$

- **Stochastic Volume Preservation.**

$$\int_A \int_B P(s, s')\, ds' ds = \int_B \int_A P(s', s)\, ds ds'$$

Theoretical Properties

Assume *symmetric transition*:

$$P(\tau'|\tau, q) = P(\tau|\tau', q), \quad \tau' \in \mathcal{N}(\tau, q)$$

Augmented state: $s = (\tau, q, p)$, a pair of measurable sets: A, B.

- **Probabilistic Reversibility.**

$$P((\tau, q, p), (\tau^*, q^*, p^*)) = P((\tau^*, q^*, -p^*), (\tau, q, -p))$$

- **Stochastic Volume Preservation.**

$$\int_A \int_B P(s, s')\, ds' ds = \int_B \int_A P(s', s)\, ds ds'$$

- **k-accessibility.** Any two states in $\mathcal{X}$ can be connected by k iterations of probabilistic path HMC.

Theoretical Properties

Assume *symmetric transition*:

$$P(\tau'|\tau, q) = P(\tau|\tau', q), \quad \tau' \in \mathcal{N}(\tau, q)$$

Augmented state: $s = (\tau, q, p)$, a pair of measurable sets: A, B.

- **Probabilistic Reversibility.**

$$P((\tau, q, p), (\tau^*, q^*, p^*)) = P((\tau^*, q^*, -p^*), (\tau, q, -p))$$

- **Stochastic Volume Preservation.**

$$\int_A \int_B P(s, s')\, ds' ds = \int_B \int_A P(s', s)\, ds ds'$$

- **k-accessibility.** Any two states in $\mathcal{X}$ can be connected by k iterations of probabilistic path HMC.

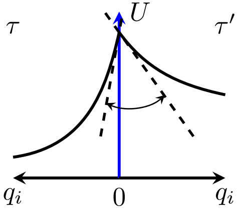
Theorem: *Probabilistic Path HMC preserves the posterior and is ergodic.*

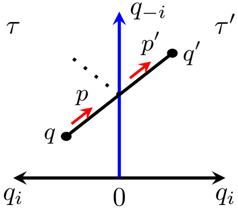
Reflection

(Afshar and Domke)

- $U(\tau, q)$ is continuous across boundary

$$\Delta E = U(\tau', q) - U(\tau, q) = 0, \quad q_i = 0$$





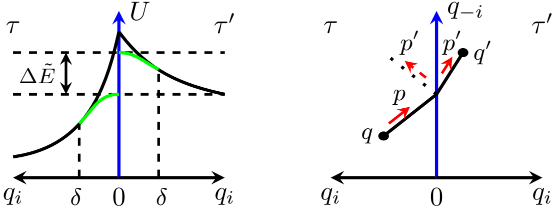
- Momentum and topology update: $p_i = -p_i, \quad \tau = \tau'$
- However, T steps leap-frog scheme with step size ϵ has the global numerical error $\mathcal{O}(C\epsilon + T\epsilon^3)$ if ∂U is discontinuous on the boundary, where C is the number of reflection events.

Surrogate Smoothing and Refraction

- Surrogate function. $\tilde{U}(\tau, q) = U(\tau, \tilde{q}), \quad \tilde{q}_i = g_\delta(q_i), \ \forall i$

$$g_\delta(x) = \begin{cases} x, & x \geq \delta \\ \frac{1}{2\delta}(x^2 + \delta^2), & 0 \leq x < \delta \end{cases}$$

- Surrogate makes the gradients equal. However, $\Delta\tilde{E} \neq 0$



- Momentum and topology update (refraction, Afshar and Domke)

$$(\tau, p_i) = \begin{cases} (\tau', \sqrt{\|p_i\|^2 - 2\Delta\tilde{E}}) & \|p_i\|^2 > 2\Delta\tilde{E} \\ (\tau, -p_i) & \text{otherwise} \end{cases}$$

Surrogate vs Exact PPHMC

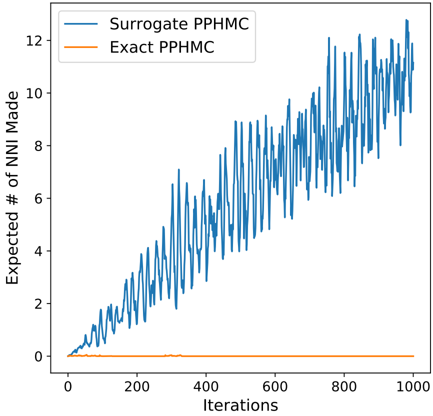


Figure: Expected number of NNI moves on a real data set.

Compared to MrBayes

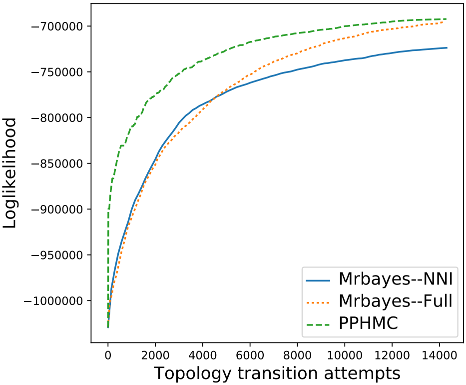


Figure: Loglikelihood vs topology transitions on a 1000 taxa simulated data set.

Conclusion

- Probabilistic path HMC extended HMC towards sampling both continuous and structural discrete parameters.
- The surrogate smoothing strategy enables long HMC paths with potential non-differentiable boundary transitions.
- Contribution in Bayesian phylogenetic inference: allowing several topology transitions in a single proposal with high acceptance rate and these transitions are all guided by the gradient and hence could be more “intelligent” than random choices.
- Future developments: enabling adaptive path length and extension to other classes of problems with similar continuous and discrete parameter structures.