Message-Passing Monte Carlo

Samuel Power

University of Cambridge, Dept. of Pure Mathematics and Mathematical Statistics

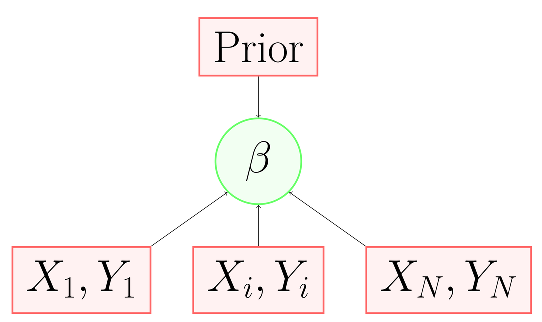
Factor Graphs

- Factor graphs[2] are a class of probabilistic model which represent probabilities as products of *local interaction* terms, called **factors**

$$\mathbb{P}(\{x_i\}_{i \in V}) \propto \prod_{a \in F} \Psi_a(x_{\partial a}) \quad (1)$$
$$= \prod_{a \in F} \exp(-U_a(x_{\partial a})) \quad (2)$$

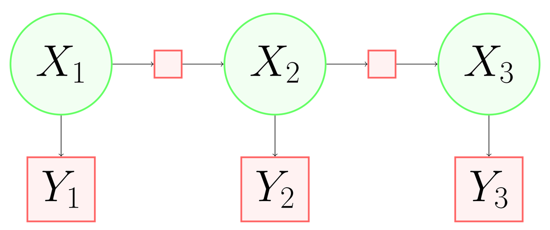
- Useful for high-dimensional models with simple (low-dimensional/sparse) interactions
- The graph consists of variable nodes, which represent the variables of the model, and factor nodes, which represent the factors in the product expansion (1).
- We connect the variable node ⓘ to the factor node $\boxed{a}$ if Ψ_a depends on x_i.
- Examples
  - Bayesian Linear Regression

$$\mathbb{P}(\beta|\mathcal{D}) \propto \exp\left(-\frac{1}{2}\tau\|\beta\|_2^2\right) \cdot \prod_i \exp\left(-\frac{(y_i - \langle x_i, \beta\rangle)^2}{2\sigma^2}\right) \quad (3)$$



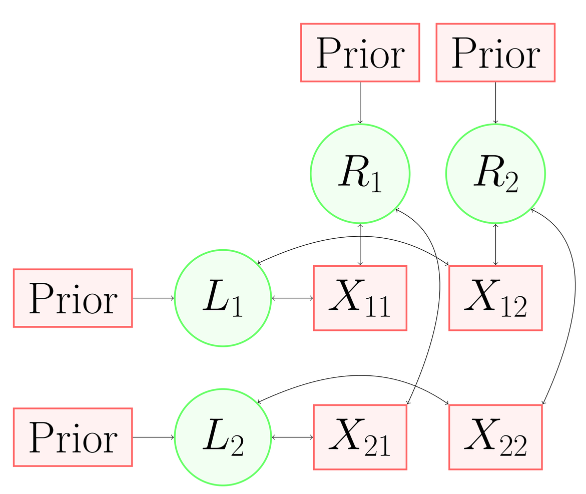
  - Hidden Markov Model

$$\mathbb{P}(X_{0:T}, Y_{1:T}) \propto \pi(X_0) \prod_{t=1}^{T} q(X_{t-1} \to X_t) \prod_{t=1}^{T} r(X_t \to Y_t) \quad (4)$$



  - Bayesian Matrix Factorisation

$$\mathbb{P}(L_{1:M}, R_{1:N}|X_{1:M,1:N}) \propto \prod_{i=1}^{M} \pi(L_i) \cdot \prod_{j=1}^{N} \pi(R_j) \quad (5)$$
$$\cdot \prod_{i,j} \exp\left(-\frac{(X_{i,j} - \langle L_i, R_j\rangle)^2}{2\sigma^2}\right) \quad (6)$$



- Computation on factor graphs should operate *locally*, propagate information *globally*.
- Gibbs sampling *fails* at the latter goal.

Geometric MCMC

- For high-dimensional targets, using the *geometry* of the measure to inform proposals is critical.
- *Hamiltonian Monte Carlo* (HMC) augments our position with a 'momentum' $p \sim \mathcal{N}(0, M)$, and then navigating the extended target

$$\pi(x, p) \propto \exp(-\mathcal{H}(x, p)) \quad (7)$$
$$\mathcal{H}(x, p) = U(x) + K(p) \quad (8)$$
$$K(p) = \frac{1}{2}p^T M^{-1} p \quad (9)$$

with Hamiltonian dynamics

$$\dot{x} = \frac{\partial \mathcal{H}}{\partial p} = M^{-1}p \quad (10)$$
$$\dot{p} = -\frac{\partial \mathcal{H}}{\partial x} = -\nabla U(x) \quad (11)$$

  - Widely-applicable, mixes well, uses gradient evaluations
  - Can be hard to choose effective M
- *Riemannian Manifold HMC* (RMHMC) allows the distribution of p to depend on x as

$$p \sim \mathcal{N}(0, M(x)), \quad (12)$$

leading to a **non-separable** Hamiltonian

$$\mathcal{H}(x, p) = U(x) + K(p|x) \quad (13)$$
$$K(p|x) = \frac{1}{2}p^T M(x)^{-1} p + \frac{1}{2}\log\det M(x). \quad (14)$$

  - Makes intimate use of geometry
  - Good at navigating complicated target measures.
  - Computationally expensive:
    - Requires matrix operations to sample from $\mathcal{N}(0, M(x))$
    - Requires implicit integrator: need fixed-point iterations to construct reversible symplectic integrator.

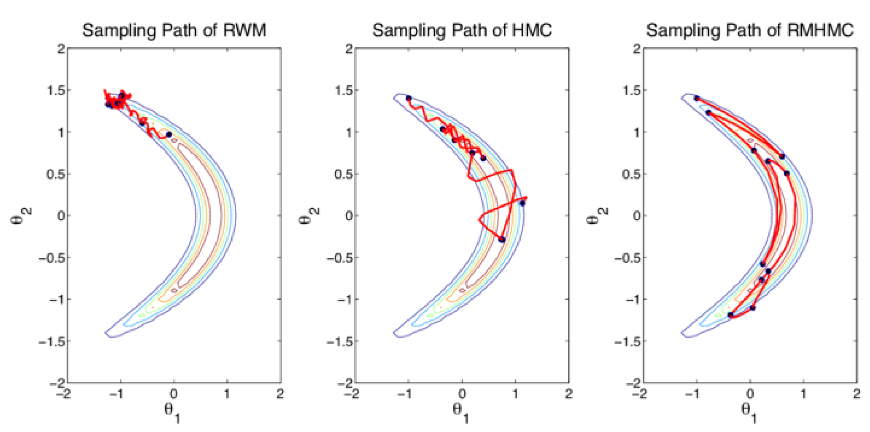


Figure 1: Comparing the paths of RWMH, HMC, RMHMC on a highly-curved target measure (figure from [1]).

Scalability

- Generally, methods which make heavier use of geometry have *faster mixing*, but incur greater *computational cost*.
- *Locality* allows for high-dimensional models to be treated in a modular way.
- Can we design a method which is *local, geometric,* and *tractable*?

Semi-Separable HMC

- Introduced in [3] for *hierarchical models.*
- Simplifies RMHMC by assuming
  - $M(x)$ is *block-diagonal* with blocks M_i, and
  - p_i is conditionally independent of x_i
    - Equivalently: M_i does not depend on x_i
- Instead of full-system RMHMC updates, do

1. Resample the momenta for **all** variables.
2. For each index i, fix $\{x_j, p_j\}_{j \neq i}$, and solve HD with respect to (x_i, p_i) only.

$$\mathcal{H}_i = U_i(x_i) + K_i(p_i) \quad (15)$$
$$U_i = U(x) + \sum_{j \neq i}\left(\frac{1}{2}\log\det M_j + \frac{1}{2}p_j^T M_j^{-1} p_j\right) \quad (16)$$
$$K_i = \frac{1}{2}p_i^T M_i^{-1} p_i \quad (17)$$

3. Repeat step 2 L times (*reversibly*).
4. Use the output as a Metropolis-Hastings proposal.

- Subsystems can be integrated efficiently.
- Better mixing due to '*auxiliary potentials*'

$$A_i = \frac{1}{2}p_i^T M(x_{\setminus i})^{-1} p_i^T, \quad (18)$$

allow for '*energy exchange*' between variables.

Factor Graph HMC

- Goal: extend SSHMC to **factor graphs**
- Design systems which make use of *locality.*
- We **split** the momentum p_i into terms corresponding to the factor nodes adjacent to i

$$p_i \mapsto \{p_{i,a}\}_{a \in \partial i} \quad (19)$$
$$p_{i,a} \sim \mathcal{N}(0, M_{i,a}) \quad (20)$$

- Stipulate that the mass matrix $M_{i,a}$ depend only on the variable nodes adjacent to a, **except** i:

$$M_{i,a} = M_{i,a}(x_{\partial a \setminus i}). \quad (21)$$

- $\rightsquigarrow$ Our Hamiltonian is

$$\mathcal{H}(\mathbf{x}, \mathbf{p}) = \sum_{a \in F} U_a(x_{\partial a}) + \sum_{(i,a) \in E} K_{i,a}(p_{i,a}|x_{\partial a \setminus i}) \quad (22)$$
$$K_{i,a} = \frac{1}{2}p_{i,a}^T M_{i,a}^{-1} p_{i,a} + \frac{1}{2}\log\det M_{i,a} \quad (23)$$

- We thus have a setup which *respects locality*, and where subsystems are *tractable.*
- *Can* define $M_{i,a}$ 'canonically' as

$$M_{i,a}(x_{\partial a \setminus i}) = \mathbf{E}\left[\nabla_{x_i}^2 U_a(x_i|x_{\partial a \setminus i})\right] \quad (24)$$

where

$$x_i \sim \exp(-U_a(x_i|x_{\partial a \setminus i})) \quad (25)$$

- Motivated by analogy with FIM (as in [1])
- $\cdots$ comes with some caveats.

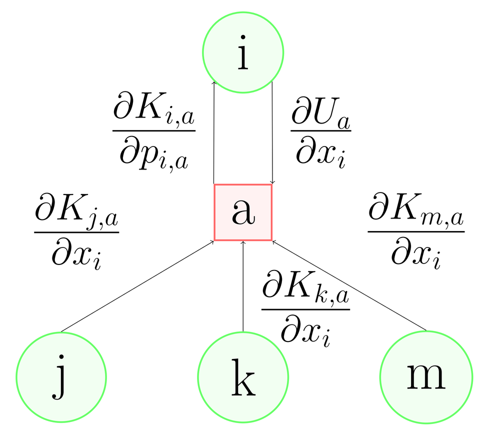
Implementation

- Writing $\mathbf{p}_i = \{p_{i,a}\}_{a \in \partial i}$, the component of the Hamiltonian which depends on $(x_i, \mathbf{p}_i)$ is

$$\mathcal{H}_i(x_i, \mathbf{p}_i) = \sum_{a \in \partial i}\left[U_a(x_{\partial a}) + \sum_{j \in \partial a \setminus i} K_{j,a}(p_{j,a}|x_{\partial a \setminus j})\right] \quad (26)$$

- This is preserved by the Hamiltonian-like *message-passing dynamics*

$$\dot{x}_i = \sum_{a \in \partial i} \frac{\partial K_{i,a}}{\partial p_{i,a}} \quad (27)$$
$$\dot{p}_{i,a} = -\frac{\partial U_a}{\partial x_i} - \sum_{j \in \partial a \setminus i} \frac{\partial K_{j,a}}{\partial x_i} \qquad a \in \partial i \quad (28)$$



- Retains many good features of HD
- Links to *belief propagation.*
- **Geometric, tractable, local**!

Future Work

- More extensive experimental testing
- Establish geometric ergodicity
- Does this *dominate* RMHMC-within-Gibbs?
- Extension to dense factor graphs (c.f. AMP).
- Stochastic gradient/'Big Data' versions?
- Impact of graph topology, update schedule.

References

[1] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[2] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

[3] Yichuan Zhang and Charles Sutton. Semi-separable hamiltonian monte carlo for inference in bayesian hierarchical models. In *Advances in neural information processing systems*, pages 10–18, 2014.