Notes on MCMC convergence of HMC and General HMC ideas

Bradley Gram-Hansen

October 2, 2017

The preliminary basics for defining MCMC convergence. Consider a probability space $(Q, \mathcal{B}(Q), \pi)$ with an n-dimensional sample space Q, the Borel σ-algebra over Q, $\mathcal{B}(Q)$ and a distinguished probability measure π. Where, in a Bayesian setting, the distinguished probability measure π represents the posterior distribution.

Definition 1. *A Markov kernel τ is a map from an element of the sample space and the σ-algebra to a probability.*

$$\tau : Qx\mathcal{B}(Q) \to [0, 1]$$

such that the kernel is a measurable function in the first argument:

$$\tau(\cdot, A) : Q \to [0, 1] \ \forall A \in \mathcal{B}(Q)$$

and a probability measure in the second argument:

$$\tau(q, \cdot) : \mathcal{B}(Q) \to [0, 1] \forall q \in Q$$

and so by construction the Markov kernel defines a map:

$$\tau : Q \to \mathcal{P}(Q)$$

where $\mathcal{P}(Q)$ represents the space of probability measures over Q. Essentially, at each point on the sample space, the kernel defines a probability measure describing how to sample a new point. By averaging the Markov kernel over all initial points in the state space, we can construct a *Markov transition* from a probability measures, to probability measures:

$$\mathcal{T} : \mathcal{P}(Q) \to \mathcal{P}(Q)$$

by:

$$\pi'(A) = \pi\mathcal{T}(A) = \int \tau(q, A)\pi(dq) \forall q \in Q, A \in \mathcal{B}(Q)$$

when the transition has an eigenvalue equation of the form:

$$\pi\mathcal{T} = \pi$$

then the transition is aperiodic, irreducible, Harris recurrent, and preserves the target measure. The repeated application of a transition of this form constructs a Markov chain, that will eventually explore the entirety of π.

1 Reprameterization Trick

Definition 2. *Law of the Unconscious Statistician (LOTUS), states that one can compute the expectation of a measurable function g of a random variable r, by integrating $g(r)$ w.r.t the distribution of r:*

$$\mathbb{E}[g(r)] = \int g(r) dF_r$$

This means that to compute the expectation of $z = g(r)$ we only need to know g and the distribution of r. We do not need to know explicitly the distribution of z.

$$\mathbb{E}_{r\sim p(r)}[g(r)] = \mathbb{E}_{z\sim p(z)}[z]$$

add more stuff here

1.1 The Gumbel Distribution Trick

Definition 3. *The random variable G is said to have a standard Gumbel distribution if:*

$$G = \log(-\log(U))$$

where $U \sim Unif[0, 1]$.

Using the Gumbel distribution, we can parameterize any discrete distribution in terms of Gumbel random variables by using the follow fact:

Definition 4. *Let X be a discrete random variable with $P(X = k) \propto \alpha_k$ random variable and let $\{G_k\}_{k\leq K}$ be an i.i.d sequence of standard Gumbel random variables. Then:*

$$X = \arg\max_k(\log \alpha_k + G_k)$$

In a high level view this means that the recipe for sampling from a categorical distribution is:

- *Draw Gumbel noise by just transforming uniform samples*
- *Add it to $\log \alpha_k$ which only has to be known up to a normalising constant.*
- *Take the value k that produces the maximum.*

1.1.1 Relaxing the Discreteness

However, arg max that tries to embed our discrete parameter is not continuous. To circumvent this, we can relax the discrete set by considering random variables taking the values in a larger, unconstrained set. To construct this relaxation we recongnise that:

- Any discrete random variable can be expressed as a one-hot vector.
- The convex hull of the set of one-hot vector is the probability simplex

$$\Delta^{K-1} = \left\{ x \in \mathbb{R}_+^K, \sum_{k=1}^{K} x_k = 1 \right\}$$

Therefore, a natural way to extend a discrete random variable is by allowing it to take values in the probability simplex. This we can do via the partition function, indexed by a temperature parameter as follows:

$$f_\tau(x)_k = \frac{\exp(x_k \backslash \tau)}{\sum_{k=1}^{K} \exp(x_k \backslash \tau)} \tag{1}$$

this enables us to define the sequence of simplex-valued random variables:

$$X^\tau = (X_k^\tau)_k = f_\tau(\log \alpha + G) = \left(\frac{\exp(\log \alpha_k + G_k) \backslash \tau}{\sum_{k=1}^{K} \exp((\log \alpha_k + G_k) \backslash \tau)} \right)$$

where X^τ is the "concrete" " distribution, that is a mixture of **con**tinuous and dis**crete**, denoted $x^\tau \sim Concrete(\alpha, \tau)$

Definition 5. *Let $X \sim Concrete(\alpha, \lambda)$ with location parameters $\alpha \in (0, \inf)^n$ and temperature $\lambda \in (0, \inf)$*

- *(Reparameterization) If $G_k \sim$ Gumbel i.i.d, then $X_k = (X_k^\tau)_k = f_\tau(\log \alpha + G) = \left(\frac{\exp(\log \alpha_k + G_k) \backslash \tau}{\sum_{k=1}^{K} \exp((\log \alpha_k + G_k) \backslash \tau)} \right)$*
- *(Rounding) $\mathcal{P}(X_k > X_i fori \neq k) = \frac{\alpha_k}{(\sum_{i=1}^{n} \alpha_i)} 4$*
- *(Zero temperature) $\mathcal{P}(lim_{\lambda \to 0} X_k = 1) = \alpha_k \backslash (\sum_{i=1}^{n} \alpha_i)$*

The pdf of the concrete distribution is given by:

$$p_{\alpha,\tau}(x) = (n-1)! \tau^{n-1} \Pi_{k=1}^{K} ($$