Sampling constrained probability distributions using Spherical Augmentation

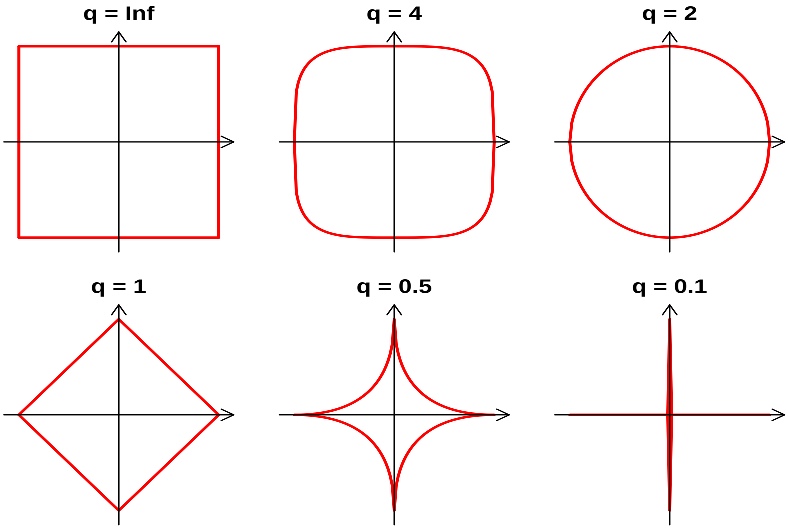
Shiwei Lan[1] Babak Shahbaba[2]

[1]Department of Statistics, University of Warwick, Coventry, UK

[2]Department of Statistics, University of California, Irvine, US

WCPM, June 4, 2015

Motivation



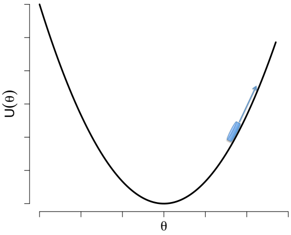
Background

- Sampling from probability distributions with constraints is common: Lasso, Bridge, probit, copula, and Latent Dirichlet Allocation, etc.
- Direct truncation is easily doable but computationally wasteful.
- Neal (2010) discusses a modified HMC algorithm for which the sampler bounces off the boundary once hitting it (Wall HMC).
- Brubaker et al (2012, constrained HMC on implicit manifolds), Pakman and Paninski (2012, exact HMC for truncated Gaussian), Byrne and Girolami (2013, Geodesic Monte Carlo on embedded manifolds), etc.

1. Review: from HMC to RHMC
2. Spherical Augmentation
   - Simple examples: ball and box
   - General q-norm constraints
   - Some functional constraints
3. Spherical Monte Carlo
   - Spherical HMC in the Cartesian coordinate
   - Spherical HMC in the spherical coordinate
   - Spherical LMC on the probability simplex
4. Experiments
5. Conclusion and future work

Warwick Statistics

Hamiltonian Monte Carlo



$$\dot{\boldsymbol{\theta}} = \frac{\partial H}{\partial \mathbf{p}}$$
$$\dot{\mathbf{p}} = -\frac{\partial H}{\partial \boldsymbol{\theta}}$$

- Position $\boldsymbol{\theta} \in \mathbb{R}^D \Longleftarrow$ variable of interest
- Momentum $\mathbf{p} \in \mathbb{R}^D \Longleftarrow$ fictitious, usually $\sim \mathcal{N}(\mathbf{0}, \mathbf{M})$
- Potential energy $U(\boldsymbol{\theta}) \Longleftarrow$ minus log of target density $f(\cdot)$
- Kinetic energy $K(\mathbf{p}) \Longleftarrow$ minus log of momentum density
- Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) \Longleftarrow$ constant.

Warwick Statistics

Hamiltonian Monte Carlo

Definition 1 (Hamiltonian dynamics)

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \frac{\partial}{\partial \mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) &=& \mathbf{M}^{-1}\mathbf{p} \\ \dot{\mathbf{p}} &= -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) &=& -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned}$$

Leapfrog: numerical integrator

$$\begin{aligned} \mathbf{p}(t+\varepsilon/2) &= \mathbf{p}(t) - (\varepsilon/2)\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t)) \\ \boldsymbol{\theta}(t+\varepsilon) &= \boldsymbol{\theta}(t) + \varepsilon \mathbf{M}^{-1}\mathbf{p}(t+\varepsilon/2) \\ \mathbf{p}(t+\varepsilon) &= \mathbf{p}(t+\varepsilon/2) - (\varepsilon/2)\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t+\varepsilon)) \end{aligned}$$

- Run for **L** steps and accept the joint proposal of $\mathbf{z} := (\boldsymbol{\theta}, \mathbf{p})$ with

$$\alpha = \min\{1, \exp(-H(\mathbf{z}^*) + H(\mathbf{z}))\}$$

Warwick Statistics

Riemannian Hamiltonian Monte Carlo

On the manifold $\{f(\cdot;\boldsymbol{\theta})\}$ with metric $G(\boldsymbol{\theta}) = -\mathrm{E}_{\mathbf{x}|\boldsymbol{\theta}}[\nabla^2_{\boldsymbol{\theta}} \log f(\mathbf{x};\boldsymbol{\theta})]$:

$$\begin{aligned} H(\boldsymbol{\theta}, \mathbf{p}) &= U(\boldsymbol{\theta}) + K(\mathbf{p}, \boldsymbol{\theta}) \\ &= -\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\log\det \mathbf{G}(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \\ &\equiv \phi(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \end{aligned}$$

where $\mathbf{p}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$. Girolami and Calderhead (2011) propose:

Definition 2 (Riemannian Hamiltonian dynamics)

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \frac{\partial}{\partial p} H(\boldsymbol{\theta}, \mathbf{p}) = \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \\ \dot{\mathbf{p}} &= -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) = -\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\partial\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \end{aligned}$$

Warwick Statistics

Lagrangian Monte Carlo

To resolve the implicitness of RHMC, Lan et al. (2012) propose

Definition 3 (Lagrangian Dynamics)

$$\dot{\boldsymbol{\theta}} = \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$$

$$\dot{\mathbf{p}} = -\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\partial\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$$

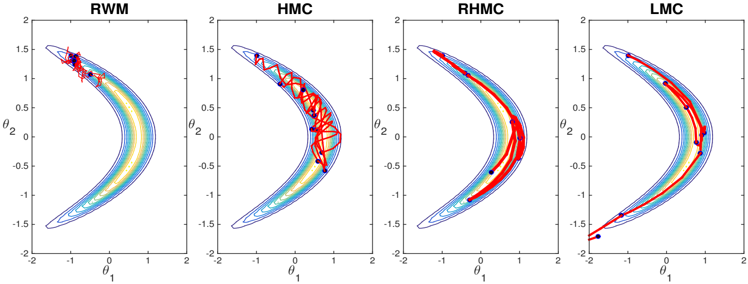
$\boxed{\mathbf{p} \to \mathbf{v}}$ $\Big\Downarrow$ **Lagrangian Dynamics**

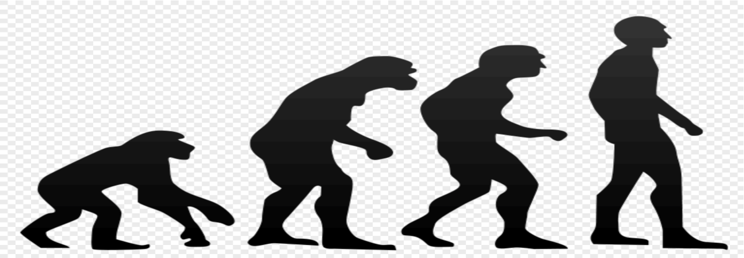
$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$

$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})$$

- Not Hamiltonian dynamics of $(\boldsymbol{\theta}, \mathbf{v})$!
- An *explicit* integrator can be found more efficient.

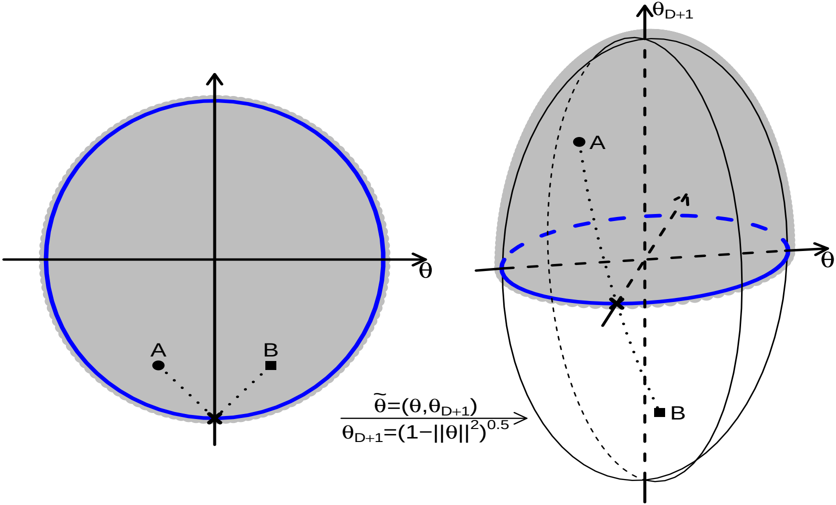
Geometry helps!



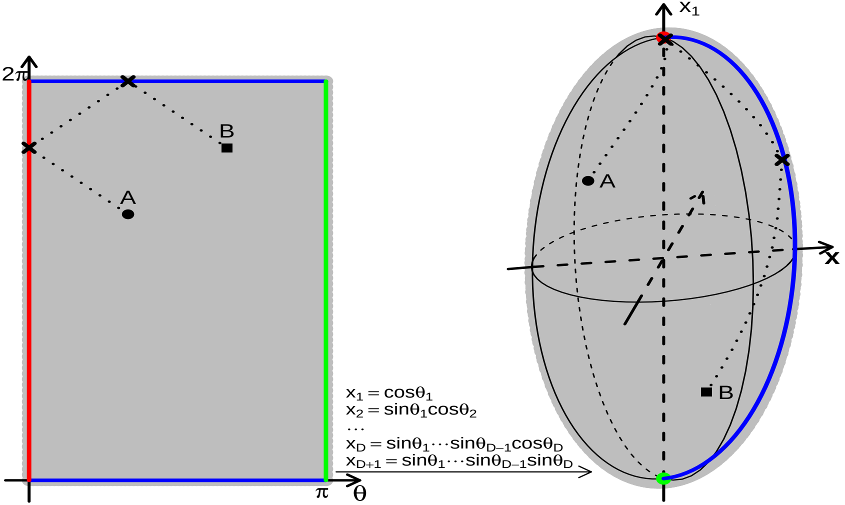


Warwick Statistics

Change of the domain: from unit ball $\mathcal{B}_{\mathbf{0}}^D(1)$ to sphere $\mathcal{S}^D$

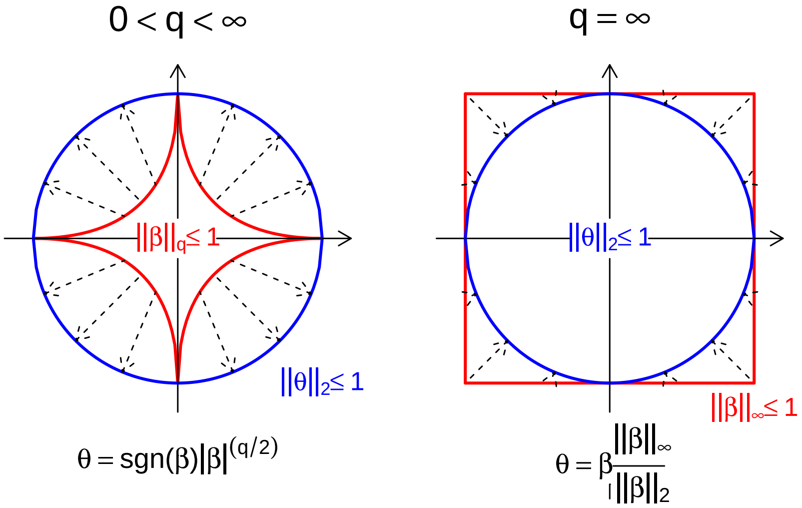


Change of the domain: from rectangle $\mathcal{R}_0^D$ to sphere $\mathcal{S}^D$



$x_1 = \cos\theta_1$
$x_2 = \sin\theta_1 \cos\theta_2$
$\cdots$
$x_D = \sin\theta_1 \cdots \sin\theta_{D-1} \cos\theta_D$
$x_{D+1} = \sin\theta_1 \cdots \sin\theta_{D-1} \sin\theta_D$

Mapping q-norm constrained domain to unit ball



Warwick Statistics

Some functional constraints

linear M linear constraints $\mathbf{l} \le \mathbf{A}\beta \le \mathbf{u}$, with $\mathbf{A}$ an $M \times D$ matrix, β a D-vector and $\mathbf{l}, \mathbf{u}$ both M-vectors.

- Assume $M = D$ and $\mathbf{A}_{D\times D}$ invertible. $\mathbf{A}^{-1}\mathbf{l} \le \beta \le \mathbf{A}^{-1}\mathbf{u}$ not true.
- Sample $\boldsymbol{\eta} := \mathbf{X}\beta$ with $\mathbf{l} \le \boldsymbol{\eta} \le \mathbf{u}$ and transform back to $\beta = \mathbf{A}^{-1}\boldsymbol{\eta}$.

dratic Quadratic constraints $l \le \beta^{\mathsf{T}}\mathbf{A}\beta + \mathbf{b}^{\mathsf{T}}\beta \le u$ with $l, u > 0$ scalars.

- Assume $\mathbf{A} = \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^{\mathsf{T}} > 0$. Use $\beta \mapsto \beta^* = \sqrt{\boldsymbol{\Sigma}}\mathbf{Q}^{\mathsf{T}}(\beta + \frac{1}{2}\mathbf{A}^{-1}\mathbf{b})$:

$$\circledcirc : l^* \le \|\beta^*\|_2^2 = (\beta^*)^{\mathsf{T}}\beta^* \le u^*,\ l^* = l + \frac{1}{4}\mathbf{b}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{b},\ u^* = u + \frac{1}{4}\mathbf{b}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{b}$$

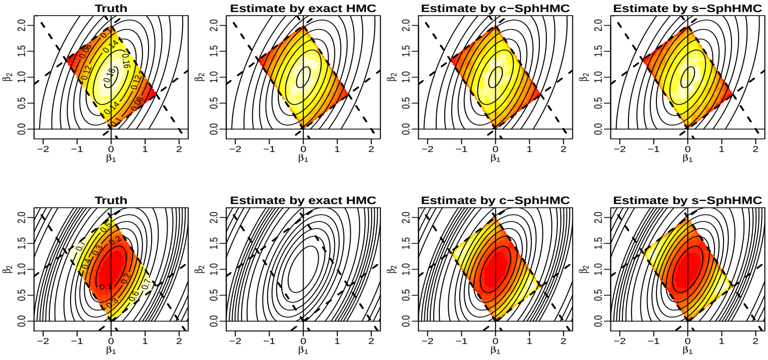
- It can further be mapped to unit ball:

$$T_{\circledcirc\to\mathcal{B}} : \mathcal{B}_{\mathbf{0}}^D(\sqrt{u^*})\backslash\mathcal{B}_{\mathbf{0}}^D(\sqrt{l^*}) \longrightarrow \mathcal{B}_{\mathbf{0}}^D(1),\ \beta^* \mapsto \boldsymbol{\theta} = \frac{\beta^*}{\|\beta^*\|_2}\frac{\|\beta^*\|_2 - \sqrt{l^*}}{\sqrt{u^*} - \sqrt{l^*}}$$

Warwick Statistics

An example of linear constraints

$$0 \leq -0.5\beta_1 + \beta_2 \leq 2 \quad \text{and} \quad 0 \leq \beta_1 + \beta_2 \leq 2$$



- upper row: $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$
- lower row: $f(\boldsymbol{\beta}) \propto \frac{\sin^2 Q(\boldsymbol{\beta})}{Q(\boldsymbol{\beta})}, \quad Q(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})$

Warwick Statistics

Warwick Statistics

Change of variables

- Denote the original parameter as β and the constrained domain as $\mathcal{D}$. We use $\boldsymbol{\theta}$ to denote the coordinate of sphere $\mathcal{S}^D$. Change variables

Change of variables

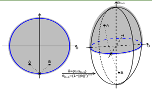
$$\int_{\mathcal{D}} f(\beta) d\boldsymbol{\beta}_{\mathcal{D}} = \int_{\mathcal{S}} f(\boldsymbol{\theta}) \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right| d\boldsymbol{\theta}_{\mathcal{S}} \tag{3.1}$$

- The energy functions will be changed to

$$\phi(\boldsymbol{\theta}) = -\log f(\boldsymbol{\theta}) - \log \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right| = U(\beta(\boldsymbol{\theta})) - \log \left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right|$$

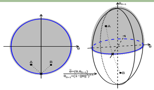
$$H(\boldsymbol{\theta}, \mathbf{v}) = \phi(\boldsymbol{\theta}) + \frac{1}{2} \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})}$$

- The Jacobian determinant $\left| \frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}} \right|$ can be used as weight afterwards. We then consider *partial* Hamiltonian $H^*(\boldsymbol{\theta}, \mathbf{v}) = U(\boldsymbol{\theta}) + \frac{1}{2} \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})}$



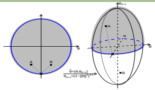
Spherical HMC
in the Cartesian coordinate

Spherical HMC for ball type constraints



$\mathcal{B}_{\mathbf{0}}^D(1) := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^D \theta_i^2} \leq 1\}$ $\xrightarrow[\theta_{D+1}=\pm\sqrt{1-\|\boldsymbol{\theta}\|_2^2}]{\boldsymbol{\theta}\mapsto\tilde{\boldsymbol{\theta}}=(\boldsymbol{\theta},\theta_{D+1})}$ $\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\}$

Spherical HMC for ball type constraints



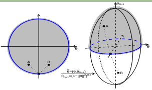
$\mathcal{B}_{\mathbf{0}}^D(1) := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^D \theta_i^2} \leq 1\}$ $\xrightarrow[\theta_{D+1}=\pm\sqrt{1-\|\boldsymbol{\theta}\|_2^2}]{\boldsymbol{\theta}\mapsto\tilde{\boldsymbol{\theta}}=(\boldsymbol{\theta},\theta_{D+1})}$ $\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\}$

Change of variables

$$\int_{\mathcal{B}_0^D(1)} f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{B}} = \int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| d\boldsymbol{\theta}_{\mathcal{S}_c} = \int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) |\theta_{D+1}| d\boldsymbol{\theta}_{\mathcal{S}_c}$$

where $f(\tilde{\boldsymbol{\theta}}) \equiv f(\boldsymbol{\theta})$.

Spherical HMC for ball type constraints



$\mathcal{B}_{\mathbf{0}}^D(1) := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^D \theta_i^2} \le 1\}$

$$\xrightarrow[\theta_{D+1}=\pm\sqrt{1-\|\boldsymbol{\theta}\|_2^2}]{\boldsymbol{\theta}\mapsto\tilde{\boldsymbol{\theta}}=(\boldsymbol{\theta},\theta_{D+1})}$$

$\mathcal{S}^D := \{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{D+1} : \|\tilde{\boldsymbol{\theta}}\|_2 = 1\}$

Change of variables

$$\int_{\mathcal{B}_0^D(1)} f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{B}} = \int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{B}}}{d\boldsymbol{\theta}_{\mathcal{S}_c}} \right| d\boldsymbol{\theta}_{\mathcal{S}_c} = \int_{\mathcal{S}_+^D} f(\tilde{\boldsymbol{\theta}}) |\theta_{D+1}| d\boldsymbol{\theta}_{\mathcal{S}_c}$$

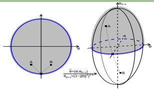
where $f(\tilde{\boldsymbol{\theta}}) \equiv f(\boldsymbol{\theta})$.

What We Want:
$\boldsymbol{\theta} \sim f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{B}}$

$$\xleftarrow[\text{weigh it by } |\theta_{D+1}|]{\text{drop } \theta_{D+1}}$$

What We Sample:
$\tilde{\boldsymbol{\theta}} \sim f(\tilde{\boldsymbol{\theta}}) d\boldsymbol{\theta}_{\mathcal{S}_c}$

Canonical spherical metric



- Here, the proper metric on $\mathcal{S}^D$ is called *canonical spherical metric*:

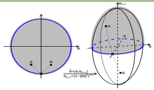
Definition 4 (canonical spherical metric)

$$\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta}) = \mathbf{I}_D + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}}{\theta_{D+1}^2} = \mathbf{I}_D + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}}{1 - \|\boldsymbol{\theta}\|_2^2} \tag{3.2}$$

- For any vector $\tilde{\mathbf{v}} = (\mathbf{v}, v_{D+1}) \in T_{\tilde{\boldsymbol{\theta}}}\mathcal{S}^D := \{\tilde{\mathbf{v}} \in \mathbb{R}^{D+1} : \tilde{\boldsymbol{\theta}}^\mathsf{T}\tilde{\mathbf{v}} = 0\}$, $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ can be viewed as a way to express the length of $\tilde{\mathbf{v}}$ in $\mathbf{v}$:

$$\mathbf{v}^\mathsf{T}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} = \|\mathbf{v}\|_2^2 + \frac{\mathbf{v}^\mathsf{T}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\mathbf{v}}{\theta_{D+1}^2} = \|\mathbf{v}\|_2^2 + \frac{(-\theta_{D+1}v_{D+1})^2}{\theta_{D+1}^2}$$
$$= \|\mathbf{v}\|_2^2 + v_{D+1}^2 = \|\tilde{\mathbf{v}}\|_2^2$$

Hamiltonian (Lagrangian) dynamics on sphere



On $\mathcal{B}_{\mathbf{0}}^{D}(1)$

$$H(\theta, \mathbf{v}) = U(\theta) + K(\mathbf{v}) = -\log f(\theta) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{I}\mathbf{v}$$

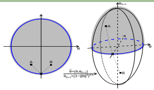
$$\xrightarrow{\theta \mapsto \tilde{\theta}}$$

On $\mathcal{S}^{D}$

$$H^{*}(\tilde{\theta}, \tilde{\mathbf{v}}) = U(\tilde{\theta}) + K(\tilde{\mathbf{v}}) = -\log f(\theta) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_c}(\theta)\mathbf{v}$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D) \xrightarrow{\mathbf{v} \mapsto \tilde{\mathbf{v}}} \tilde{\mathbf{v}} \sim (\mathbf{I}_{D+1} - \tilde{\theta}\tilde{\theta}^{\mathsf{T}})\mathcal{N}(\mathbf{0}, \mathbf{I}_{D+1})$$

Hamiltonian (Lagrangian) dynamics on sphere



On $\mathcal{B}_{\mathbf{0}}^D(1)$

On $\mathcal{S}^D$

$$H(\boldsymbol{\theta}, \mathbf{v}) = U(\boldsymbol{\theta}) + K(\mathbf{v}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{I}\mathbf{v}$$

$$\xrightarrow{\boldsymbol{\theta}\mapsto\tilde{\boldsymbol{\theta}}}$$

$$H^*(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{v}}) = U(\tilde{\boldsymbol{\theta}}) + K(\tilde{\mathbf{v}}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v}$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$$

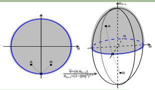
$$\xrightarrow{\mathbf{v}\mapsto\tilde{\mathbf{v}}}$$

$$\tilde{\mathbf{v}} \sim (\mathbf{I}_{D+1} - \tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\mathsf{T}})\mathcal{N}(\mathbf{0}, \mathbf{I}_{D+1})$$

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$
$$\|\boldsymbol{\theta}\|_2 \leq 1$$

$$\longrightarrow$$

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$
$$\theta_{D+1} = \sqrt{1 - \|\boldsymbol{\theta}\|_2^2}, \quad v_{D+1} = -\boldsymbol{\theta}^{\mathsf{T}}\mathbf{v}/\theta_{D+1}$$

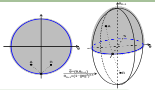


Split Lagrangian dynamics on sphere

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} - \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned} \tag{3.3}$$

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{0} \\ \dot{\mathbf{v}} &= -\frac{1}{2} \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned} \quad \longleftrightarrow \quad \begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} \end{aligned}$$

Split Lagrangian dynamics on sphere



$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} - \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned} \tag{3.3}$$

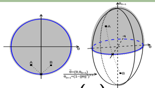
$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{0} \\ \dot{\mathbf{v}} &= -\frac{1}{2} \mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{aligned} \quad \longleftrightarrow \quad \begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_c}(\boldsymbol{\theta}) \mathbf{v} \end{aligned}$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(t) &= \tilde{\boldsymbol{\theta}}(0) \\ \tilde{\mathbf{v}}(t) &= \tilde{\mathbf{v}}(0) \\ &- \frac{t}{2} \left[\begin{bmatrix} \mathbf{I}_D \\ \mathbf{0}^{\mathsf{T}} \end{bmatrix} - \tilde{\boldsymbol{\theta}}(0) \boldsymbol{\theta}(0)^{\mathsf{T}} \right] \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(0)) \end{aligned}$$

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(t) =& \tilde{\boldsymbol{\theta}}(0) \cos(\|\tilde{\mathbf{v}}(0)\|_2 t) \\ &+ \frac{\tilde{\mathbf{v}}(0)}{\|\tilde{\mathbf{v}}(0)\|_2} \sin(\|\tilde{\mathbf{v}}(0)\|_2 t) \\ \tilde{\mathbf{v}}(t) =& - \tilde{\boldsymbol{\theta}}(0) \|\tilde{\mathbf{v}}(0)\|_2 \sin(\|\tilde{\mathbf{v}}(0)\|_2 t) \\ &+ \tilde{\mathbf{v}}(0) \cos(\|\tilde{\mathbf{v}}(0)\|_2 t) \end{aligned}$$

Error analysis



Denote $\mathbf{z} := (\boldsymbol{\theta}, \mathbf{v})$, $\mathbf{z}(t_n)$ as the true solution to (3.3) at time t_n and $\mathbf{z}^{(n)}$ the numerical solution at n-th step. We have the following bound of the error $e_n = \|\mathbf{z}(t_n) - \mathbf{z}^{(n)}\|$:

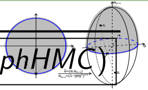
Proposition 1

Assume $\mathbf{f}(\boldsymbol{\theta}, \mathbf{v}) := \mathbf{v}^T \boldsymbol{\Gamma}_{\mathcal{S}}(\boldsymbol{\theta})\mathbf{v} + \mathbf{G}_{\mathcal{S}}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ *is smooth. Then*

$$e_{n+1} \leq (1 + M_1\varepsilon + M_2\varepsilon^2)e_n + \mathcal{O}(\varepsilon^3)$$

where $M_k = c_k \sup_{t\in[0,T]} \|\nabla^k \mathbf{f}(\boldsymbol{\theta}(t), \mathbf{v}(t))\|$, $k = 1, 2$ *for some constants* $c_k > 0$. $\varepsilon = t_{n+1} - t_n$ *is the discretization step size. Further accumulating the local errors for* $L = T/\varepsilon$ *steps yields the global error*

$$e_{L+1} \leq (e^{M_1 T} + T)\varepsilon^2$$



Algorithm 1 Spherical HMC in the Cartesian coordinate $(c - SphHMC)$

Initialize $\tilde{\boldsymbol{\theta}}^{(1)}$ at current $\tilde{\boldsymbol{\theta}}$ after transformation $T_{\mathcal{D}\to\mathcal{S}}$

Sample a new velocity value $\tilde{\mathbf{v}}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D+1})$

Set $\tilde{\mathbf{v}}^{(1)} \leftarrow \tilde{\mathbf{v}}^{(1)} - \tilde{\boldsymbol{\theta}}^{(1)}(\tilde{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\tilde{\mathbf{v}}^{(1)}$

Calculate $H(\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\mathbf{v}}^{(1)}) = U(\boldsymbol{\theta}^{(1)}) + K(\tilde{\mathbf{v}}^{(1)})$

for $\ell = 1$ to L **do**

$$\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})} = \tilde{\mathbf{v}}^{(\ell)} - \frac{\varepsilon}{2}\left(\begin{bmatrix}\mathbf{I}_D\\ \mathbf{0}^{\mathsf{T}}\end{bmatrix} - \tilde{\boldsymbol{\theta}}^{(\ell)}(\boldsymbol{\theta}^{(\ell)})^{\mathsf{T}}\right)\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(\ell)})$$

$$\tilde{\boldsymbol{\theta}}^{(\ell+1)} = \tilde{\boldsymbol{\theta}}^{(\ell)}\cos(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon) + \frac{\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}}{\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|}\sin(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon)$$

$$\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})} \leftarrow -\tilde{\boldsymbol{\theta}}^{(\ell)}\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\sin(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon) + \tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\cos(\|\tilde{\mathbf{v}}^{(\ell+\frac{1}{2})}\|\varepsilon)$$

$$\tilde{\mathbf{v}}^{(\ell+1)} = \tilde{\mathbf{v}}^{(\ell+\frac{1}{2})} - \frac{\varepsilon}{2}\left(\begin{bmatrix}\mathbf{I}_D\\ \mathbf{0}^{\mathsf{T}}\end{bmatrix} - \tilde{\boldsymbol{\theta}}^{(\ell+1)}(\boldsymbol{\theta}^{(\ell+1)})^{\mathsf{T}}\right)\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}^{(\ell+1)})$$

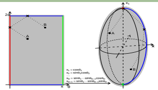
end for

Calculate $H(\tilde{\boldsymbol{\theta}}^{(L+1)}, \tilde{\mathbf{v}}^{(L+1)}) = U(\boldsymbol{\theta}^{(L+1)}) + K(\tilde{\mathbf{v}}^{(L+1)})$

Calculate the acceptance probability $\alpha = \min\{1, \exp[-H(\tilde{\boldsymbol{\theta}}^{(L+1)}, \tilde{\mathbf{v}}^{(L+1)}) + H(\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\mathbf{v}}^{(1)})]\}$

Accept or reject the proposal according to α for the next state $\tilde{\boldsymbol{\theta}}'$

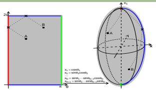
Calculate $T_{\mathcal{S}\to\mathcal{D}}(\tilde{\boldsymbol{\theta}}')$ and the corresponding weight $|dT_{\mathcal{S}\to\mathcal{D}}|$



Spherical HMC

in the spherical coordinate

Spherical HMC for box type constraints

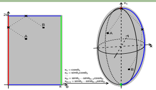


$\mathcal{R}_{\mathbf{0}}^D := [0,\pi]^{D-1} \times [0, 2\pi)$

$$\xrightarrow[x_d = \cos\theta_d \prod_{i=1}^{d-1} \sin\theta_i]{\boldsymbol{\theta} \mapsto \mathbf{x}}$$

$\mathcal{S}^D := \{\mathbf{x} \in \mathbb{R}^{D+1} : \|\mathbf{x}\|_2 = 1\}$

Spherical HMC for box type constraints



$\mathcal{R}_{\mathbf{0}}^D :=$
$[0,\pi]^{D-1} \times [0, 2\pi)$

$\xrightarrow[x_d = \cos\theta_d \prod_{i=1}^{d-1} \sin\theta_i]{\boldsymbol{\theta} \mapsto \mathbf{x}}$

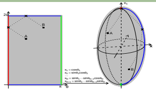
$\mathcal{S}^D := \{\mathbf{x} \in \mathbb{R}^{D+1} :$
$\|\mathbf{x}\|_2 = 1\}$

Change of measure

$$\int_{\mathcal{R}_{\mathbf{0}}^D} f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{R}_{\mathbf{0}}} = \int_{\mathcal{S}^D} f(\boldsymbol{\theta}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{R}_{\mathbf{0}}}}{d\boldsymbol{\theta}_{\mathcal{S}_r}} \right| d\boldsymbol{\theta}_{\mathcal{S}_r} = \int_{\mathcal{S}^D} f(\boldsymbol{\theta}) \prod_{d=1}^{D-1} \sin^{d-D} \theta_d d\boldsymbol{\theta}_{\mathcal{S}_r}$$

where $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}(\mathbf{x}))$ on $\mathcal{S}^D$.

Spherical HMC for box type constraints



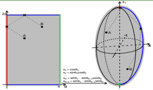
$\mathcal{R}_{\mathbf{0}}^D := [0,\pi]^{D-1} \times [0, 2\pi)$ $\xrightarrow[x_d = \cos\theta_d \prod_{i=1}^{d-1} \sin\theta_i]{\boldsymbol{\theta} \mapsto \mathbf{x}}$ $\mathcal{S}^D := \{\mathbf{x} \in \mathbb{R}^{D+1} : \|\mathbf{x}\|_2 = 1\}$

Change of measure

$$\int_{\mathcal{R}_{\mathbf{0}}^D} f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{R}_{\mathbf{0}}} = \int_{\mathcal{S}^D} f(\boldsymbol{\theta}) \left| \frac{d\boldsymbol{\theta}_{\mathcal{R}_{\mathbf{0}}}}{d\boldsymbol{\theta}_{\mathcal{S}_r}} \right| d\boldsymbol{\theta}_{\mathcal{S}_r} = \int_{\mathcal{S}^D} f(\boldsymbol{\theta}) \prod_{d=1}^{D-1} \sin^{d-D}\theta_d \, d\boldsymbol{\theta}_{\mathcal{S}_r}$$

where $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}(\mathbf{x}))$ on $\mathcal{S}^D$.

What We Want: $\boldsymbol{\theta} \sim f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{R}_{\mathbf{0}}}$ $\xleftarrow[\text{by } \prod_{d=1}^{D-1} \sin^{d-D}\theta_d]{\text{weigh sample } \boldsymbol{\theta}}$ What We Sample: $\boldsymbol{\theta} \sim f(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\mathcal{S}_r}$



Round spherical metric

- Here, the natural metric on $\mathcal{S}^D$ is called *round spherical metric*:

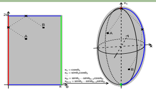
Definition 5 (round spherical metric)

$$\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta}) = \operatorname{diag}\left[1, \sin^2\theta_1, \cdots, \prod_{d=1}^{D-1}\sin^2\theta_d\right] \tag{3.4}$$

- For any vector $\mathbf{v} \in T_{\boldsymbol{\theta}}\mathcal{R}_{\mathbf{0}}^D$, we have

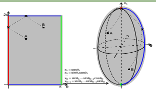
$$\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} \le \|\mathbf{v}\|_2^2 \le \|\tilde{\mathbf{v}}\|_2^2 = \mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})\mathbf{v}$$

Hamiltonian (Lagrangian) dynamics on sphere



On $\mathcal{R}_{\mathbf{0}}^D$		On $\mathcal{S}^D$
$H(\boldsymbol{\theta}, \mathbf{v}) = U(\boldsymbol{\theta}) + K(\mathbf{v}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{I}\mathbf{v}$	$\xrightarrow{\boldsymbol{\theta}\mapsto\mathbf{x}}$	$H^*(\mathbf{x}, \dot{\mathbf{x}}) = U(\mathbf{x}) + K(\dot{\mathbf{x}}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v}$
$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$	$\xrightarrow{\mathbf{v}\mapsto\dot{\mathbf{x}}}$	$\mathbf{v}(\mathbf{x}, \dot{\mathbf{x}}) \sim \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathcal{N}(\mathbf{0}, \mathbf{I}_D)$

Hamiltonian (Lagrangian) dynamics on sphere



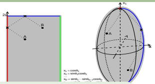
On $\mathcal{R}_{\mathbf{0}}^D$		On $\mathcal{S}^D$
$H(\boldsymbol{\theta}, \mathbf{v}) = U(\boldsymbol{\theta}) + K(\mathbf{v}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{I}\mathbf{v}$	$\xrightarrow{\boldsymbol{\theta}\mapsto\mathbf{x}}$	$H^*(\mathbf{x}, \dot{\mathbf{x}}) = U(\mathbf{x}) + K(\dot{\mathbf{x}}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v}$
$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$	$\xrightarrow{\mathbf{v}\mapsto\dot{\mathbf{x}}}$	$\mathbf{v}(\mathbf{x}, \dot{\mathbf{x}}) \sim \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathcal{N}(\mathbf{0}, \mathbf{I}_D)$

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$
$$\mathbf{l} \le \boldsymbol{\theta} \le \mathbf{u}$$

$\longrightarrow$

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^\mathsf{T}\boldsymbol{\Gamma}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$
$$\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{x}), \quad \mathbf{v} = \mathbf{v}(\mathbf{x}, \dot{\mathbf{x}})$$

Split Lagrangian dynamics on sphere

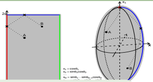


$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$

$$\dot{\boldsymbol{\theta}} = \mathbf{0}$$
$$\dot{\mathbf{v}} = -\frac{1}{2}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$

$\longleftrightarrow$

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v}$$



Split Lagrangian dynamics on sphere

$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v} - \mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$

$$\dot{\boldsymbol{\theta}} = \mathbf{0}$$
$$\dot{\mathbf{v}} = -\frac{1}{2}\mathbf{G}_{\mathcal{S}_r}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$

$$\longleftrightarrow$$

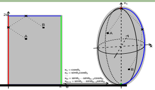
$$\dot{\boldsymbol{\theta}} = \mathbf{v}$$
$$\dot{\mathbf{v}} = -\mathbf{v}^{\mathsf{T}} \boldsymbol{\Gamma}_{\mathcal{S}_r}(\boldsymbol{\theta})\mathbf{v}$$

$$\downarrow$$

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(0)$$
$$\mathbf{v}(t) = \mathbf{v}(0) - \frac{t}{2} \cdot$$
$$\mathrm{diag}\left[1, \cdots, \prod_{d=1}^{D-1} \sin^{-2}\theta_d\right] \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(0)$$

$$\downarrow$$

$$(\boldsymbol{\theta}(0), \mathbf{v}(0)) \longrightarrow (\mathbf{x}(0), \dot{\mathbf{x}}(0))$$
$$\downarrow$$
$$(\mathbf{x}(t), \dot{\mathbf{x}}(t)) = g_r(\mathbf{x}(0), \dot{\mathbf{x}}(0))$$
$$\downarrow$$
$$(\boldsymbol{\theta}(0), \mathbf{v}(0)) \longleftarrow (\mathbf{x}(0), \dot{\mathbf{x}}(0))$$



Algorithm 2 Spherical HMC in the spherical coordinate (s-SphHMC)

Initialize $\boldsymbol{\theta}^{(1)}$ at current $\boldsymbol{\theta}$ after transformation $T_{\mathcal{D}\to\mathcal{S}}$
Sample a new velocity value $\mathbf{v}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$
Set $v_d^{(1)} \leftarrow v_d^{(1)} \prod_{i=1}^{d-1} \sin^{-1}(\theta_i^{(1)})$, $d = 1, \cdots, D$
Calculate $H(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}) = U(\boldsymbol{\theta}^{(1)}) + K(\mathbf{v}^{(1)})$
for $\ell = 1$ to L **do**
  $v_d^{(\ell+\frac{1}{2})} = v_d^{(\ell)} - \frac{\varepsilon^d}{2} \frac{\partial}{\partial \theta_d} U(\boldsymbol{\theta}^{(\ell)}) \prod_{i=1}^{d-1} \sin^{-2}(\theta_i^{(\ell)})$, $d = 1, \cdots, D$
  $(\boldsymbol{\theta}^{(\ell+1)}, \mathbf{v}^{(\ell+\frac{1}{2})}) \leftarrow \tilde{T}_{\mathcal{S}\to\mathcal{R}_\mathbf{0}} \circ g_\varepsilon \circ \tilde{T}_{\mathcal{R}_\mathbf{0}\to\mathcal{S}}(\boldsymbol{\theta}^{(\ell)}, \mathbf{v}^{(\ell+\frac{1}{2})})$
  $v_d^{(\ell+1)} = v_d^{(\ell+\frac{1}{2})} - \frac{\varepsilon^d}{2} \frac{\partial}{\partial \theta_d} U(\boldsymbol{\theta}^{(\ell+1)}) \prod_{i=1}^{d-1} \sin^{-2}(\theta_i^{(\ell+1)})$, $d = 1, \cdots, D$
end for
Calculate $H(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}) = U(\boldsymbol{\theta}^{(L+1)}) + K(\mathbf{v}^{(L+1)})$
Calculate the acceptance probability $\alpha = \min\{1, \exp[-H(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}) + H(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)})]\}$
Accept or reject the proposal according to α for the next state $\boldsymbol{\theta}'$
Calculate $T_{\mathcal{S}\to\mathcal{D}}(\boldsymbol{\theta}')$ and the corresponding weight $|dT_{\mathcal{S}\to\mathcal{D}}|$

Spherical LMC
on the probability simplex

Spherical LMC on the probability simplex

- A class of models having probability distributions defined on *simplex*

$$\Delta^K := \{\boldsymbol{\pi} \in \mathbb{R}^D | \, \pi_k \geq 0, \sum_{k=1}^{K} \pi_d = 1\}$$

- *Latent Dirichlet Allocation (LDA)* (Blei et al., 2003) is a hierarchical Bayesian model frequently used to model document topics.
- 1-norm constraint: identify the first (all positive) orthant with others.
- $T_{\Delta \to \sqrt{\Delta}} : \boldsymbol{\pi} \mapsto \boldsymbol{\theta} = \sqrt{\boldsymbol{\pi}}$ maps the simplex to the sphere

$$\sqrt{\Delta}^K := \{\boldsymbol{\theta} \in \mathcal{S}^{K-1} | \theta_k \geq 0, \, \forall \, k = 1, \cdots, K\} \subset \mathcal{S}^{K-1}$$

Spherical LMC on the probability simplex

- Prototype example: Dirichlet-Multinomial distribution

$$p(x_i = k|\boldsymbol{\pi}) = \pi_k, \quad k = 1, \cdots, K$$

$$p(\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$p(\boldsymbol{\pi}|\mathbf{x}) \propto \prod_{k=1}^{K} \pi_k^{n_k + \alpha_k - 1}, \quad n_k = \sum_{i=1}^{N} I(x_i = k), \; n = \sum_{k=1}^{K} n_k$$

- Fisher metric on $\sqrt{\Delta}$ coincides $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ on $\mathcal{S}^{K-1}$ up to a constant.

$$\mathbf{G}_\Delta(\boldsymbol{\pi}_{-K}) = n[\mathrm{diag}(1/\boldsymbol{\pi}_{-K}) + \mathbf{1}\mathbf{1}^\mathsf{T}/\pi_K]$$

$$\mathbf{G}_{\sqrt{\Delta}}(\boldsymbol{\theta}) = \frac{d\boldsymbol{\pi}_{-K}^\mathsf{T}}{d\boldsymbol{\theta}_{-K}} \mathbf{G}_\Delta(\boldsymbol{\pi}_{-K}) \frac{d\boldsymbol{\pi}_{-K}}{d\boldsymbol{\theta}_{-K}^\mathsf{T}} = 4n\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$$

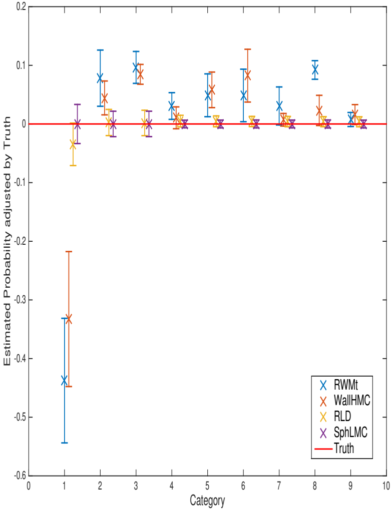
Spherical LMC on the probability simplex

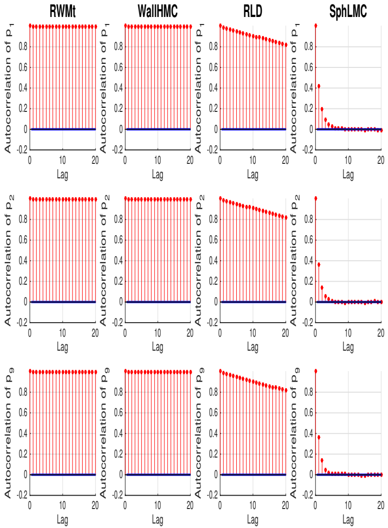
- Use $\mathbf{G}_{\sqrt{\Delta}}(\boldsymbol{\theta})$ instead of $\mathbf{G}_{\mathcal{S}_c}(\boldsymbol{\theta})$ in c-SphHMC.
- Include the volume adjustment term, $\left|\frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}}\right|$ in the Hamiltonian

$$H(\boldsymbol{\theta}, \mathbf{v}) = \phi(\boldsymbol{\theta}) + \frac{1}{2}\langle \mathbf{v}, \mathbf{v}\rangle_{\mathbf{G}_{\sqrt{\Delta}}(\boldsymbol{\theta})}, \quad \phi(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) - \log\left|\frac{d\boldsymbol{\beta}_{\mathcal{D}}}{d\boldsymbol{\theta}_{\mathcal{S}}}\right|$$

- No afterward re-weight: online learning
- c-SphHMC $\xrightarrow{\textit{above modifications}}$ Spherical Lagrangian Monte Carlo.
- SphLMC: stems from the Fisher metric on the simplex.

Spherical LMC on the probability simplex





Warwick Statistics

Experiments

Definition 6 (Effective Sample Size)

For N samples, effective sample size is calculated as follows:

$$ESS = N[1 + 2\Sigma_{k=1}^{K}\rho(k)]^{-1}$$

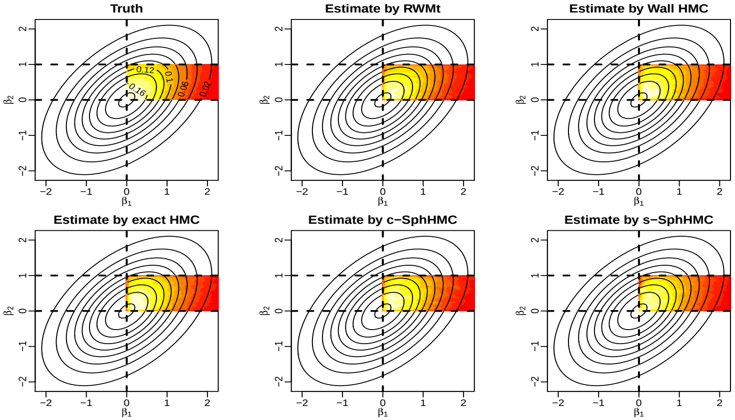
where $\rho(k)$ is the autocorrelation function with lag k, and $K \gg 1$.

- Performance measured by time-normalized ESS.
- Interpreted as number of nearly independent samples.
- Use the minimum ESS normalized by CPU time: min(ESS)/s.
- Compare RWMt, Wall HMC, exact HMC, c-SphHMC, s-SphHMC, RLD and SphLMC.

Warwick Statistics

Truncated Multivariate Gaussian

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}1 & .5\\ .5 & 1\end{bmatrix}\right), \qquad 0 \le \beta_1 \le 5, \quad 0 \le \beta_2 \le 1$$



Warwick Statistics

Truncated Multivariate Gaussian

- To evaluate efficiency, we increase the dimensionality for $D = 10,\ 100$

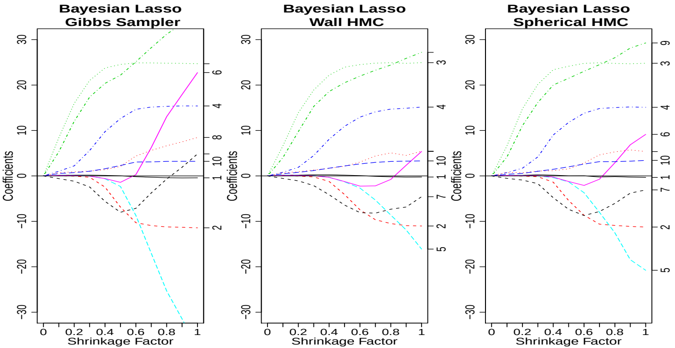
$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),\ \Sigma_{ij} = 1/(1+|i-j|); \quad 0 \le \beta_1 \le 5,\ 0 \le \beta_i \le 0.5,\ i \ne 1.$$

- RWM: $>$ 95% of times proposals rejected due to constraint violation.
- Wall HMC: average wall hits 3.81 (L=2, D=10), 6.19 (L=5, D=100).

Dim	Method	AP	s/iter	ESS(min,med,max)	Min(ESS)/s	spdup
D=10	RWMt	0.62	5.72E-05	(48,691,736)	7.58	1.00
	Wall HMC	0.83	1.19E-04	(31904,86275,87311)	2441.72	322.33
	exact HMC	1.00	7.60E-05	(1e+05,1e+05,1e+05)	11960.29	1578.87
	c-SphHMC	0.82	2.53E-04	(62658,85570,86295)	2253.32	297.46
	s-SphHMC	0.79	2.02E-04	(76088,1e+05,1e+05)	3429.56	452.73
D=100	RWMt	0.81	5.45E-04	(1,4,54)	0.01	1.00
	Wall HMC	0.74	2.23E-03	(17777,52909,55713)	72.45	5130.21
	exact HMC	1.00	4.65E-02	(97963,1e+05,1e+05)	19.16	1356.64
	c-SphHMC	0.73	3.45E-03	(55667,68585,72850)	146.75	10390.94
	s-SphHMC	0.87	2.30E-03	(74476,99670,1e+05)	294.31	20839.43

Warwick Statistics

Bayesian Lasso: regularized regression

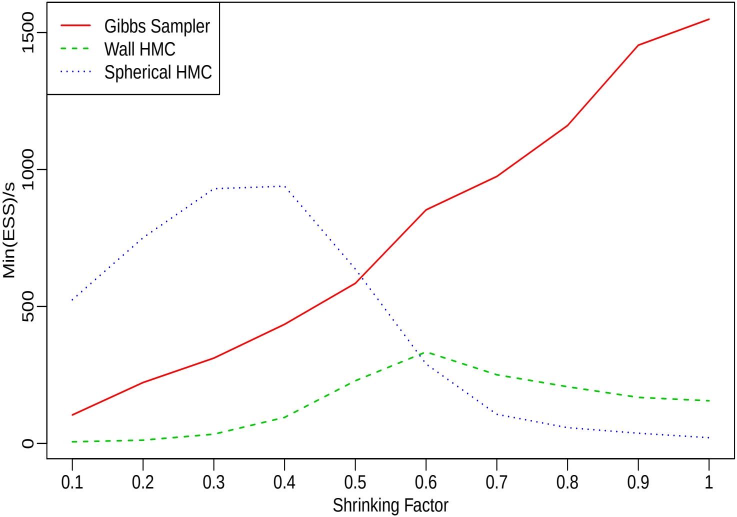


- Obtain the coefficients β by minimizing the residual sum of squares (RSS) subject to a constraint on the magnitude of β

$$\min_{\|\beta\|_1 \le t} \mathrm{RSS}(\beta), \qquad \mathrm{RSS}(\beta) := \sum_i (y_i - \beta_0 - x_i^{\mathsf{T}}\beta)^2$$

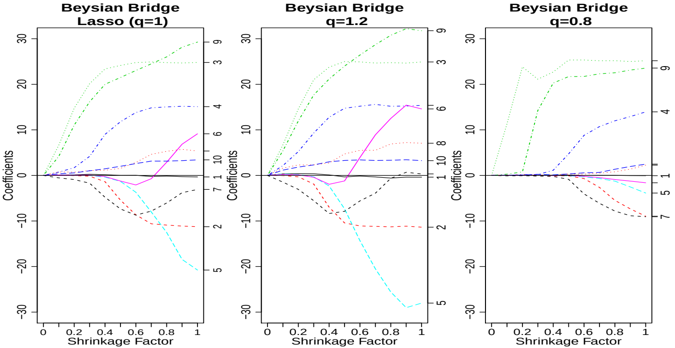
- Park and Casella (2008) use a Laplace prior: $P(\beta) \propto \exp(-\lambda|\beta|)$

Bayesian Lasso



Warwick Statistics

Bayesian Bridge: regularized regression

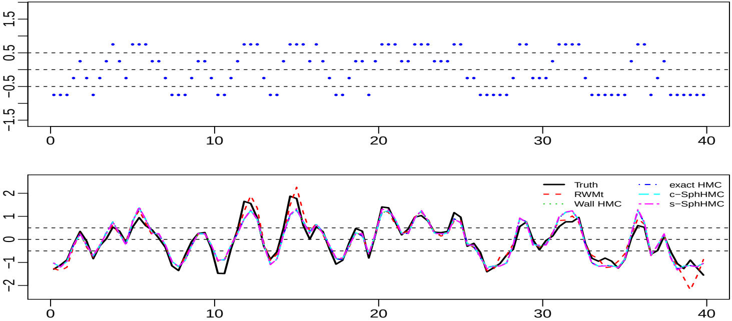


- Obtain the coefficients β by minimizing the residual sum of squares (RSS) subject to a constraint on the magnitude of β

$$\min_{\|\beta\|_q \le t} \mathrm{RSS}(\beta), \qquad \mathrm{RSS}(\beta) := \sum_i (y_i - \beta_0 - x_i^\mathsf{T} \beta)^2$$

- Polson et al (2013) have Bayesian Bridge with complicated priors

Reconstruction of quantized stationary Gaussian process



- Given N values of a function $\{f(x_i)\}_{i=1}^N$, taken values in a set $\{q_k\}_{k=1}^K$
- Assume this is a quantized projection of $y(x_i)$ from a stationary GP

$$f(x_i) = q_k, \quad \text{if } z_k \le y(x_i) < z_{k+1}$$

- The objective is to sample from the posterior distribution

$$p(\mathbf{y}|\mathbf{f}) \sim \mathcal{TN}(0, \Sigma), \quad \Sigma_{ij} = \sigma^2 \exp\left\{-\frac{|x_i - x_j|^2}{2\eta^2}\right\}, \sigma^2 = 0.6,\ \eta^2 = 0.2$$

Warwick Statistics

Reconstruction of quantized stationary Gaussian process

Method	AP	s/iter	ESS(min,med,max)	Min(ESS)/s	spdup
RWMt	0.70	7.11E-05	(2,9,35)	0.22	1.00
Wall HMC	0.69	9.94E-04	(12564,24317,43876)	114.92	534.48
exact HMC	1.00	1.00E-02	(72074,1e+05,1e+05)	65.31	303.76
c-SphHMC	0.72	1.73E-03	(13029,26021,56445)	68.44	318.32
s-SphHMC	0.80	1.09E-03	(14422,31182,81948)	120.59	560.86

Table: Comparing efficiency of RWMt, Wall HMC, exact HMC, c-SphHMC and s-SphHMC in reconstructing a quantized stationary Gaussian process. AP is acceptance probability, s/iter is seconds per iteration, ESS(min,med,max) is the (minimal,median,maximal) effective sample size, and Min(ESS)/s is the minimal ESS per second.

Warwick Statistics

LDA on Wikipedia corpus

- LDA(Blei et al. 2003) is a popular Bayesian model for topic modeling.
- The model consists of K topicks π_k, which are distributions over the words in the collection, drawn from a Dirichlet prior $\mathrm{Dir}(\beta)$.
- A document d is modeled by a mixture of topics, with mixing proportion $\eta_d \sim \mathrm{Dir}(\alpha)$.
- Documents are produced by drawing a topic assignment z_{di} i.i.d from η_d for each word w_{di} in document d, and then drawing the word w_{di} from the assigned topic $\pi_{z_{di}}$.

Warwick Statistics

LDA on Wikipedia corpus

- Conditioned on π, the documents are i.i.d, and the joint distribution can be factorized (Patterson and Teh, 2013)

$$p(w, z, \pi|\alpha, \beta) = p(\pi|\beta) \prod_{d=1}^{D} p(w_d, z_d|\alpha, \pi)$$

$$p(w_d, z_d|\alpha, \pi) = \prod_{k=1}^{K} \frac{\Gamma(\alpha + n_{dk\cdot})}{\Gamma(\alpha)} \prod_{w=1}^{W} \pi_{kw}^{n_{dkw}}$$

- To compare with sg-RLD (Patterson and Teh, 2013), apply SphLMC to update $\boldsymbol{\theta} = \sqrt{\boldsymbol{\pi}}$ with stochastic gradient for $L = 1$ decreasing ε

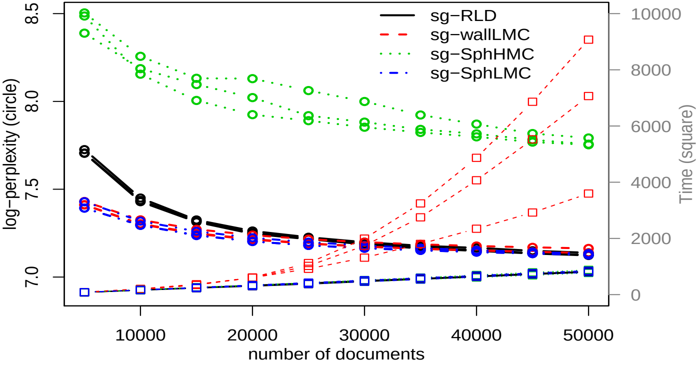
$$g_{kw} = [(n^*_{kw} + \beta - 1/2)/\theta_{kw} + \theta_{kw}(n^*_{k\cdot} + W(\beta - 1/2))]/(2 * n^*_{k\cdot})$$

where $n^*_{kw} = \frac{|D|}{|D_t|} \sum_{d \in D_t} \mathrm{E}_{z_d|w_d,\theta,\alpha}[n_{dkw}]$, and $|D_t| = 50$.

Warwick Statistics

LDA on Wikipedia corpus

- Online learn 50000 documents randomly downloaded from Wikipedia.
- Vocabulary consists of approx. 8000 words from Project Gutenburg.
- Evaluate the performance in perplexity on 1000 held-out documents.



Warwick Statistics

Conclusion

- *Spherical Augmentation (SA)* is a natural and efficient framework to handle norm related constraints in statistical inference.
- Spherical HMC and Spherical LMC demonstrate substantial advantage over existing methods. SA can have more extensions.
- Based on change of variables, SA defines the dynamics on sphere in 1 higher dimension by slack variable or embedding map. The resulting sampler moves on sphere freely while implicitly handling constraints.
- To account for the change of geometry, volume adjustment is needed to re-weight samples (SphHMC) or added to Hamiltonian (SphLMC).

Warwick Statistics

Future work

- Instead of Euclidean metric $\mathbf{I}$ on $\mathcal{B}_0^D(1)$, we can start from Fisher metric $\mathbf{G_F}(\boldsymbol{\theta})$, and consider metric like $\mathbf{G_F}(\boldsymbol{\theta}) + \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}/\theta_{D+1}^2$ for augmented space to facilitate exploring complicated structures.
- Derive an acceptance rule that does not drop quickly as dimension increases (Beskos et al., 2011).
- Develop tune-free algorithms for spherical HMC (Hoffman and Gelman, 2011).

Thank you !

Web: http://www.ics.uci.edu/~slan/SphHMC/Intro.html