Stochastic Gradient MCMC with Repulsive Forces

Víctor Gallego, David Ríos Insua

ICMAT-CSIC (National Research Council)

Intro and Goals

- Recent developments in Bayesian techniques applied to large scale datasets or deep models include **variational approaches** such as *Automatic Differentiation Variational Inference* (ADVI) [1] and *Stein Variational Gradient Descent* (SVGD) [2], or **sampling approaches** such as *Stochastic Gradient Markov Chain Monte Carlo* (SG-MCMC) [3].
- Can we bridge the gap between variational and sampling methods?
  Yes, we propose an hybrid between SGLD and SVGD!

Background

- SG-MCMC [3]
  1. Choose state space $\boldsymbol{z} \in \mathbb{R}^d$ and target distribution $\pi \propto \exp(-H(\boldsymbol{z}))$.
  2. Choose suitable diffusion $\mathbf{D}(\boldsymbol{z})$ and curl $\mathbf{Q}(\boldsymbol{z})$) matrices.
  3. Discretize the generalized Langevin dynamics:

     $$\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t - \epsilon_t \left[(\mathbf{D}(\boldsymbol{z}_t) + \mathbf{Q}(\boldsymbol{z}_t)) \nabla H(\boldsymbol{z}_t) + \boldsymbol{\Gamma}(\boldsymbol{z}_t) \right] + \boldsymbol{\eta}_t,$$

     where $\boldsymbol{\eta}_t$ is some carefully chosen Gaussian noise.
  - Stochastic Gradient Langevin Dynamics (SGLD): $\mathbf{D} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$.
  - Hamiltonian variant (HMC): $\bar{\boldsymbol{z}} = (\boldsymbol{z}, \boldsymbol{p})$.
    $\mathbf{D} = \mathbf{0}$ and $\mathbf{Q} = \begin{pmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}$
- SVGD [2] frames posterior sampling as an optimization process, in which a set of K particles $\{\boldsymbol{z}_i\}_{i=1}^K$ is evolved iteratively via

  $$\boldsymbol{z}_{i,t+1} \leftarrow \boldsymbol{z}_{i,t} - \epsilon_t \frac{1}{K} \Sigma_{j=1}^K \left[k(\boldsymbol{z}_{j,t}, \boldsymbol{z}_{i,t}) \nabla H(\boldsymbol{z}_{j,t}) + \nabla_{\boldsymbol{z}_{j,t}} k(\boldsymbol{z}_{j,t}, \boldsymbol{z}_{i,t}) \right],$$

  where the RBF kernel $k(\boldsymbol{z}, \boldsymbol{z}') = \exp(-\frac{1}{h} \|\boldsymbol{z} - \boldsymbol{z}'\|^2)$ is typically adopted. This velocity field is chosen so as to to maximize the decreasing rate on the KL divergence between the particle distribution and the target.

Proposed scheme

Instead of using K parallel chains without interactions, we propose

$$\begin{array}{ccc} SGD & \xrightarrow{\text{noise}} & SGLD \\ \text{rep. b/w particles} \downarrow & & \downarrow \\ SVGD & \longrightarrow & \boldsymbol{SGLD+R} \end{array}$$

***Parallel SGLD plus repulsion* (SGLD+R)**:

$$\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t - \frac{\epsilon_t}{K} (\mathbf{K} \boldsymbol{\nabla} + \boldsymbol{\Gamma}) + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, 2\epsilon_t \mathbf{K}/K).$$

where $\mathbf{K}$ is a permuted block-diagonal matrix such that for each block $(\mathbf{K})_{i,j} = k(\boldsymbol{z}_{j,t}, \boldsymbol{z}_{i,t})$.
(i.e., instead of identity or diagonal diffusion matrix as in SGLD, we use a block-diagonal matrix accounting for distances between particles)

Since matrix $\mathbf{K}$ is definite positive (it was constructed from the RBF kernel), we may now use the key result from [3] (Theorem 1) to derive the following property:

Proposition

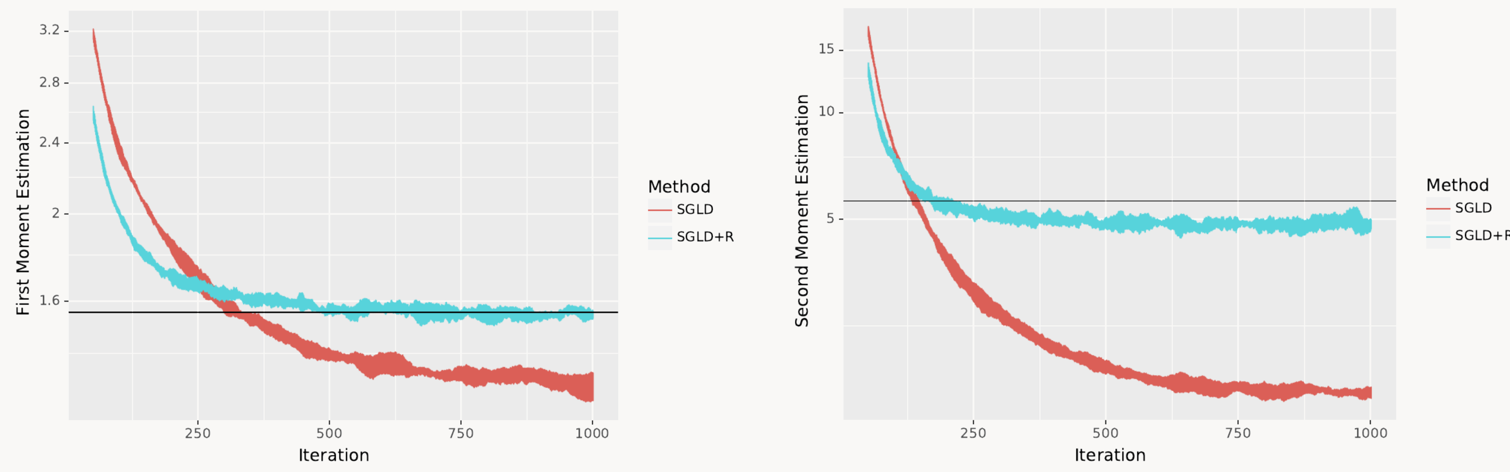
SGLD+R (or its general form, Eq. (1)) has $\pi(\boldsymbol{z}) = \Pi_{k=1}^K \pi(\boldsymbol{z}_k)$ as stationary distribution, and the proposed discretizations are asymptotically exact as $\epsilon_t \to 0$.

Experiments

Synthetic distributions:
- **Mixture of Exponentials (MoE)**.
- **Mixture of 2D Gaussians (MoG)**.

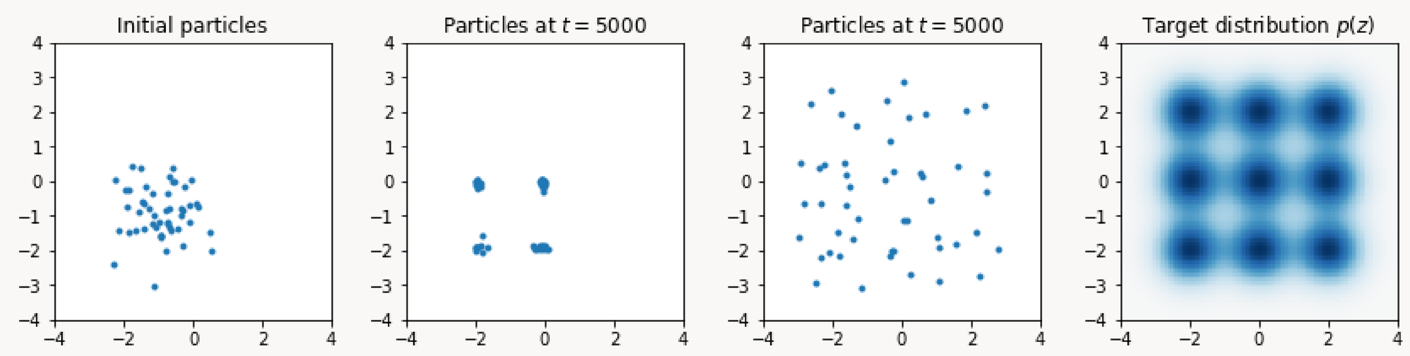
	ESS		ESS/s		Error of $\mathbb{E}[X]$	
Distribution	**SGLD**	**SGLD+R**	**SGLD**	**SGLD+R**	**SGLD**	**SGLD+R**
MoE	44.3	**59.1**	51.5	**61.0**	0.39	**0.14**
MoG	151.3	**169.5**	**36.3**	32.5	1.42	**1.19**

Table: Results for the two synthetic distributions task



(a) Estimation of $\mathbb{E}[X]$ (b) Estimation of $\mathbb{E}[X^2]$

Figure: Evolution of estimation during the MoE experiments (5 simulations). 10 particles are used for each sim. and black line depicts the exact value to be estimated



(a) Prior (b) SGLD (c) SGLD+R (d) MoG $\pi(z)$

Figure: Evolution of the particles during the MoG experiment

Bayesian Neural Network:
Feed-forward neural network over some regression tasks from the UCI datasets.

	Avg. Test RMSE		Avg. Test LL	
Dataset	**SGLD**	**SGLD+R**	**SGLD**	**SGLD+R**
Boston	2.392 ± 0.018	$\mathbf{2.295 \pm 0.017}$	-2.551 ± 0.018	-2.575 ± 0.007
Kin8nm	0.104 ± 0.001	0.104 ± 0.001	0.826 ± 0.005	0.831 ± 0.006
Naval	0.008 ± 0.000	0.008 ± 0.000	3.379 ± 0.011	$\mathbf{3.428 \pm 0.019}$
Protein	4.810 ± 0.003	$\mathbf{4.794 \pm 0.003}$	-2.991 ± 0.000	$\mathbf{-2.987 \pm 0.001}$
Wine	0.522 ± 0.004	$\mathbf{0.514 \pm 0.004}$	-0.765 ± 0.008	$\mathbf{-0.750 \pm 0.007}$
Yacht	0.942 ± 0.015	$\mathbf{0.894 \pm 0.029}$	-1.211 ± 0.020	-1.172 ± 0.026

Conclusions and Further Work

- We showed how to generate new SG-MCMC methods consisting in multiple chains plus repulsion between the particles.
- Repulsion between particles improves exploration of the space, avoiding particle collapse. Plus, we may collect much more samples than with SVGD.
- Explore different matrices $\mathbf{K}$ and $\mathbf{Q}$ in order to further accelerate the sampling process

$$\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t - \epsilon_t \left[(\mathbf{K} + \mathbf{Q}) \boldsymbol{\nabla} + \boldsymbol{\Gamma} \right] + \boldsymbol{\eta}_t. \qquad (1)$$

References

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe.
Variational Inference: A Review for Statisticians.
Journal of the American Statistical Association, 112(518):859–877, 2017.

[2] Qiang Liu and Dilin Wang.
Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm.
In *Advances In Neural Information Processing Systems*, 2016.

[3] Yi-An Ma, Tianqi Chen, and Emily Fox.
A Complete Recipe for Stochastic Gradient MCMC.
In *Advances in Neural Information Processing Systems*. 2015.

victor.gallego@icmat.es