ICML @ Sydney

Thirty-fourth International Conference on Machine Learning

# Stochastic gradient monomial Gamma sampler

Yizhe Zhang[1,2], Changyou Chen[1], Zhe Gan[1], Ricardo Henao[1], Lawrence Carin[1]

Department of Electronic and Computer Engineering[1], Duke University, Durham, NC, 27708

Department of Statistical Science[2], Duke University, Durham, NC, 27708

## Motivation & Contribution

**1)** Improving stationary mixing efficiency in SGMCMC by leveraging a generalized (potentially heavy-tailed) kinetics.

**2)** Alleviating numerical issue and satisfying conditions for stationarity by leveraging smooth version of generalized kinetics.

**3)** Ameliorating convergence issue by introducing additional first order dynamics and stochastic resampling.

## MGHMC

*Hamiltonian Monte Carlo* (HMC) leverages Hamiltonian dynamics to propose new samples for x from $p(x) \propto \exp(-U(x))$, driven by the following partial differential equations (PDE):

$$\frac{d\boldsymbol{x}}{dt} = \nabla_p K(\boldsymbol{p}) \,, \qquad \frac{d\boldsymbol{p}}{dt} = -\nabla_x U(\boldsymbol{x}) \,.$$

- Consider the generalized kinetic $K(p; m, a) = \frac{|p|^{1/a}}{m}, a, m > 0$
- *monomial Gamma* (MG) distribution:

$$\pi(p; m, a) = \exp(-K(p; m, a)) = \frac{m^{-a}}{2\Gamma(a+1)} e^{-\frac{|p|^{1/a}}{m}} \,.$$

**Theorem**

*For univariate target distribution, the one time lag autocorrelation $\rho(x_t, x_{t+1})$ of the analytic MG-SS parameterized by $a$ asymptotically approaches zero when $a \to \infty$, under regularity condition of $U(x)$ and stationary assumption.*

- In addition to above, the MG-HMC with large $a$ is particularly advantageous for sampling multimodal distributions.
- Such a performance gain does not come in free.

## SGMCMC

- Sampling from $f(\theta) \propto \exp(-U(\theta))$ using minibatch data.
- SGHMC (stochastic gradient Hamiltonian Monte Carlo)

$$d\theta = \partial_p K(p) dt$$
$$dp = -\partial_\theta \tilde{U}(\theta) dt - B(\theta) \partial_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt).$$

- SGNHT (stochastic gradient Nosé-Hoover thermostat)

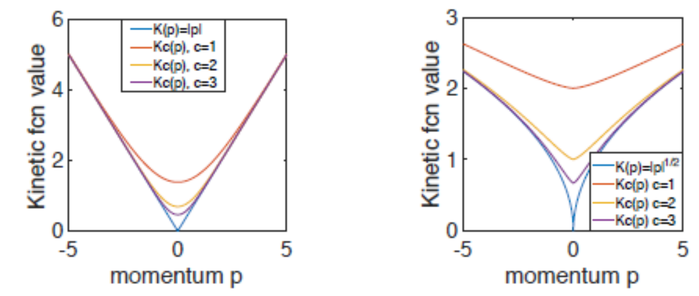
$$d\theta = \partial_p K(p) dt$$
$$dp = -\partial_\theta \tilde{U}(\theta) dt - \xi \partial_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt)$$
$$d\xi = (p^T p - 1) dt.$$

## SGMGS (Stochastic gradient MG sampler)

- Consider applying generalized kinetics $K(p)$ to SGNHT.

$$d\theta = \nabla K(p) dt \,,$$
$$dp = -[\nabla \tilde{U}(\theta) + \xi \odot \nabla K(p)] dt + \sqrt{2AI} dW \,,$$
$$d\xi = (\nabla K(p) \odot \nabla K(p) - \nabla^2 K(p)) dt \,.$$

- The existence and uniqueness of the solutions to the Fokker-Planck equation require Lipschitz continuity of drift and diffusion vectors.
- We propose a softened kinetics for $a = \{1, 2\}$




## Convergence issue and remedies

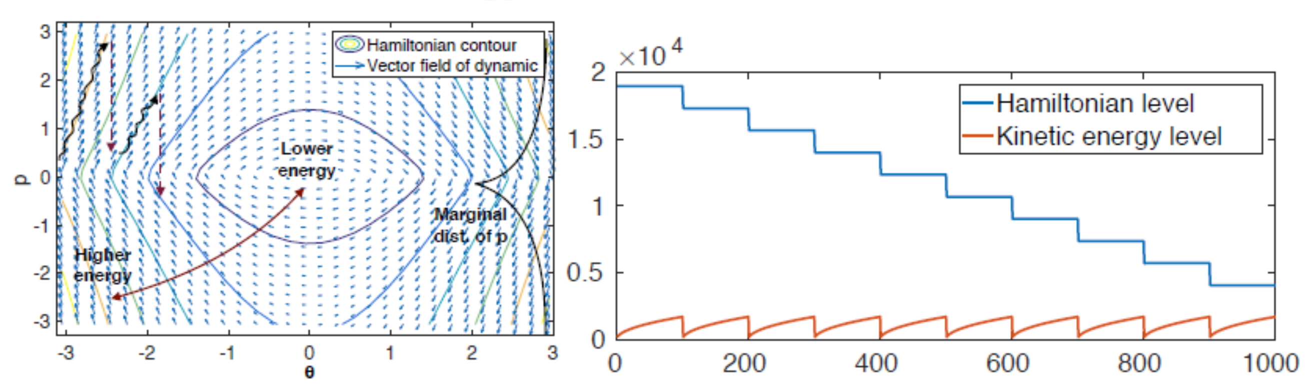
- Consider a Hamiltonian system defined in a more general form

$$H = K(p) + U(\theta) + F(\xi) \,,$$

- We consider adding Brownian motion to $\theta$ and $\xi$

$$d\theta = \nabla K_c(p) dt - \sigma_\theta \nabla U(\theta) dt + \sqrt{2\sigma_\theta} dW$$
$$dp = -(\sigma_p + \gamma \nabla F(\xi)) \odot \nabla K_c(p) dt - \nabla U(\theta) dt + \sqrt{2\sigma_p} dW,$$
$$d\xi = \gamma \left[ \nabla K_c(p) \odot \nabla K_c(p) - \nabla^2 K_c(p) \right] dt - \sigma_\xi \nabla F(\xi) dt + \sqrt{2\sigma_\xi} dW \,.$$

- $-\sigma_\theta \nabla U(\theta) dt + \sqrt{2\sigma_\theta} dW$, compensate for the weak updating signal from $\nabla K(p) = \frac{|p|^{1/a-1}}{am}$, by an immediate gradient
- The proposed SDE has better theoretic guarantee on the existence and convergence of bounded solutions for a
- Resampling makes the sampler to immediately move to a lower Hamiltonian energy level.




**Theorem**

$$\text{Bias: } \left| \mathbb{E}\hat{\phi}_T - \bar{\phi} \right| = O\left(1/(Th) + h\right) \,,$$

$$\text{MSE: } \mathbb{E}\left(\hat{\phi} - \bar{\phi}\right)^2 = O\left(1/(Th) + h^2\right) \,.$$

## Multiple-well Synthetic Potential

- Generate samples from a complex multimodal distribution.

$$U(\theta) \triangleq e^{\frac{3}{4}\theta^2 - \frac{3}{2}\sum_{i=1}^{10} c_i \sin\left(\frac{1}{4}\pi i(\theta+4)\right)}$$

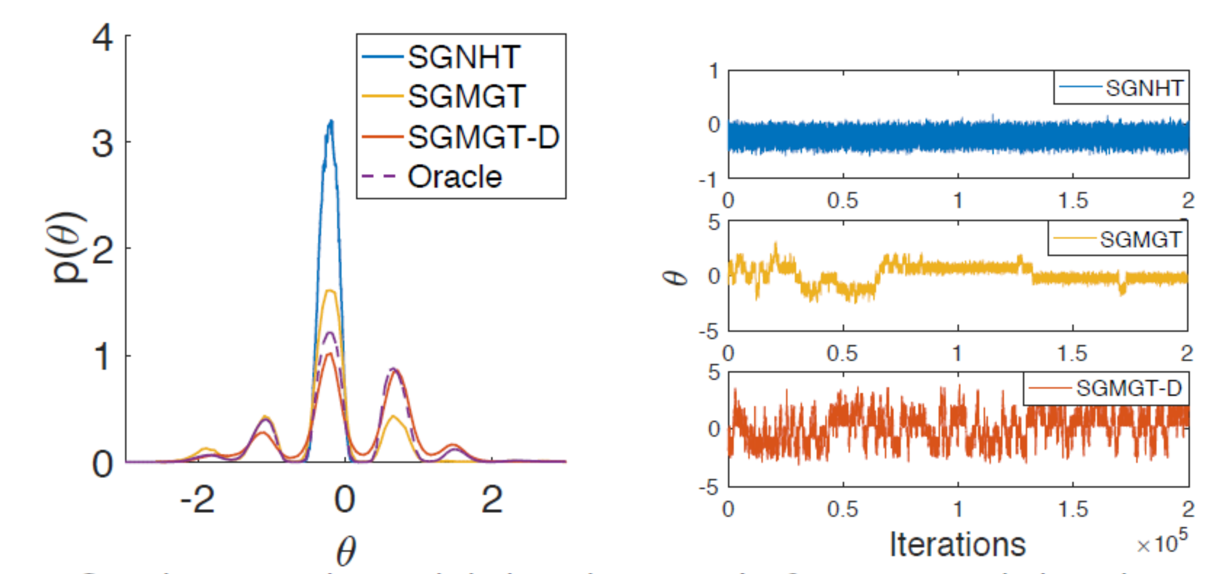



Figure: Synthetic multimodal distribution. Left: empirical distributions for different methods. Right: traceplot for each method.

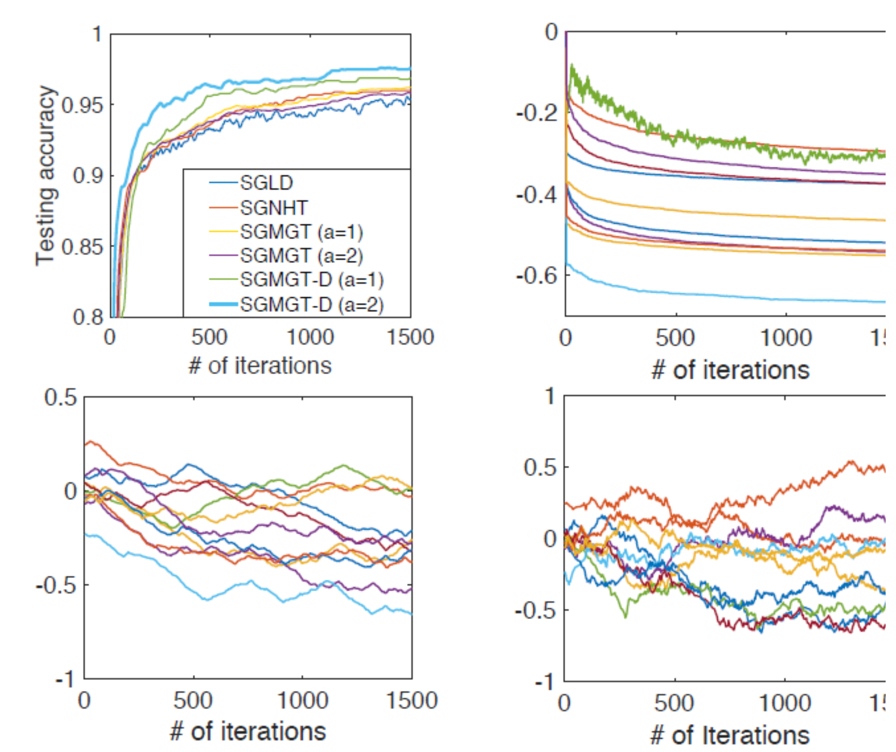
## Bayesian Logistic Regression

Table: Average AUROC and median ESS. Dataset dimensionality is indicated in parenthesis after the name of each dataset.

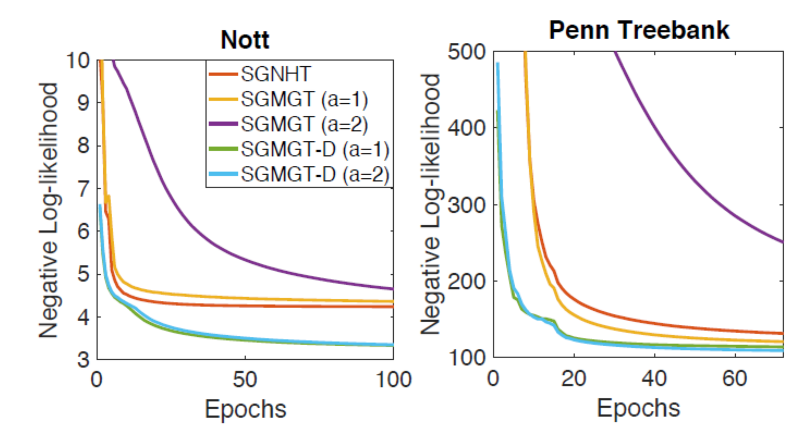
| AUROC ($D$) | A (15) | G (25) | H (14) | P(8) | R (7) | C (87) |
|---|---|---|---|---|---|---|
| SGNHT | 0.89 | 0.75 | 0.90 | 0.86 | 0.95 | 0.65 |
| SGMGT(a=1) | 0.92 | 0.78 | 0.91 | 0.86 | 0.87 | 0.70 |
| SGMGT-D(a=1) | 0.95 | 0.86 | 0.95 | **0.93** | **0.98** | **0.73** |
| SGMGT(a=2) | 0.93 | 0.79 | 0.93 | 0.88 | 0.86 | 0.62 |
| SGMGT-D(a=2) | **0.95** | **0.90** | **0.95** | 0.90 | 0.97 | 0.69 |
| ESS ($D$) | A (15) | G (25) | H (14) | P(8) | R (7) | C (87) |
| SGNHT | 869 | 941 | 1911 | 2077 | 1761 | 1873 |
| SGMGT-D(a=1) | **3147** | **2131** | 2448 | **4244** | 1494 | **3605** |
| SGMGT-D(a=2) | 2700 | 1989 | **2768** | 3430 | **2265** | 2969 |

## Deep models

- Discriminative RBM & Recurrent Neural Networks. Assuming flat prior. 5000 MC samples.




| Algorithms | Piano | Nott | Muse | JSB | PTB |
|---|---|---|---|---|---|
| SGLD | 11.37 | 6.07 | 10.83 | 11.25 | 127.47 |
| SGNHT | 9.00 | 4.24 | 7.85 | 9.27 | 131.3 |
| SGMGT (a=1) | 7.90 | 4.35 | 8.42 | 8.67 | 120.6 |
| SGMGT (a=2) | 10.17 | 4.64 | 8.51 | 8.84 | 250.5 |
| SGMGT-D (a=1) | **7.51** | **3.33** | 7.11 | 8.46 | 113.8 |
| SGMGT-D (a=2) | 7.53 | 3.35 | **7.09** | **8.43** | **109.0** |
| SGD | 11.13 | 5.26 | 10.08 | 10.81 | 120.44 |
| RMSprop | 7.70 | 3.48 | 7.22 | 8.52 | 120.45 |




## Conclusion

- Scalable MCMC inference with generalized HMC variants.

Future direction:

- Adaptive selection of monomial parameters
- Connection to optimization methods.