L2 norm burst during BNN training (1)

-Summary-

23/08/29

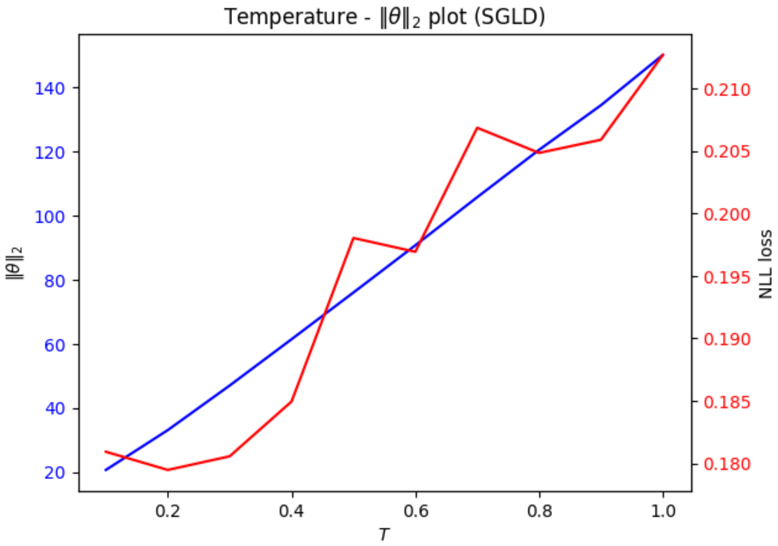
Observed problem (Review)

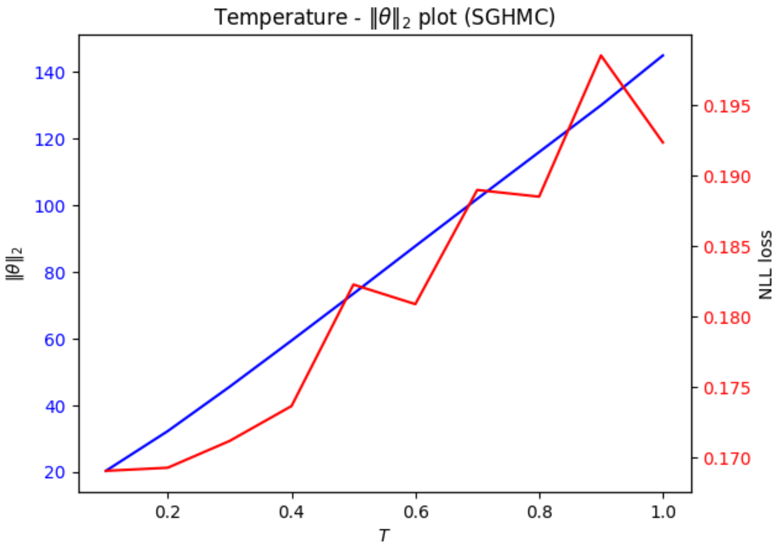
- During training of DNN, burst of weight L_2 norm is observed while the test accuracy is maintained high, or NLL is reasonably low.
- One strategy to avoid this issue is to adopt cold tempered posterior:

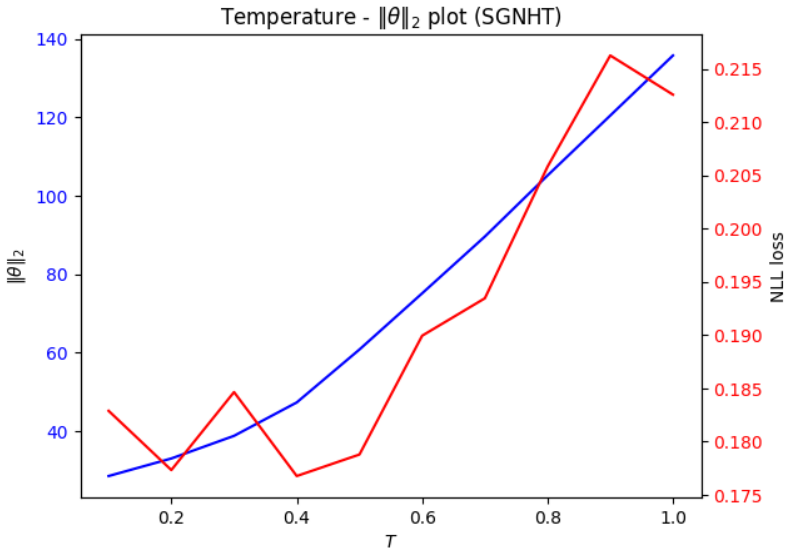
$$p_T(\theta|\mathcal{D}) \propto \exp\left(-\frac{U(\theta)}{T}\right)$$

- Empirically, it is observed that as the $T \in (0,1]$ approaches to 0, the L_2 norm of weights (after convergence) becomes lower.
- If the $T \to 0$, the only highest mode of $p(\theta|\mathcal{D})$ survive, which results in MAP training (≅ SGD w/ weight decaying if prior is isotropic gaussian)

Reimplementation







- We observe that NLL loss decreases as the weight norm decreases, and this phenomenon can be addressed using the concept of Rademacher complexity.

Theoretical explanation

Theorem 1 [J. Wang, 2019]

Denote softmax function by SF. Let $\mathcal{H}$ be a family of functions for 3-layer NN with C outputs (identity activation on the output layer), and $\mathcal{H}_j$ be a family of functions for j-th output. For a loss function l with Lipschitz constant L_l, we have

$$\hat{\mathcal{R}}_n(l \circ SF \circ \mathcal{H} \circ S) \leq 2\sqrt{2}C \cdot L_l \sum_{j=1}^{m} \hat{\mathcal{R}}_{\mathrm{n}}\left(\mathcal{H}_{\mathrm{j}} \circ \mathrm{S}\right)$$

Theoretical explanation

Theorem 2 [Bartlett and Mendelson, 2003]

Let σ be Lipschitz with constant L_σ. Define class of functions $H_j = \left\{x \mapsto \sum_i w_{j,i}\sigma(v_i x) : \left\|w_j\right\|_2 \leq B_1, \|v_i\|_2 \leq B_0\right\}$.

Then, the following holds:

$$\hat{\mathcal{R}}_n\left(\mathcal{H}_j \circ S\right) \leq \frac{L_\sigma B_0 B_1}{\sqrt{n}} \max_{i\in[n]} \|x_i\|_2$$

Accordingly, by theorem 1, we get the following bounds of empirical Rademacher complexity:

$$\hat{\mathcal{R}}_n(l \circ SF \circ \mathcal{H} \circ S) \leq \frac{2\sqrt{2}C^2 L_l L_\sigma B_0 B_1}{\sqrt{n}} \max_{i\in[n]} \|x_i\|_2$$

(For a loss function l with Lipschitz constant L_l)

Phenomenon analysis

- First of all, why does the weight norm increases as temperature T increases?
  - During the derivation of Fokker-Planck equation:

$$\frac{d\mathbb{E}[\phi]}{dt} = \sum_i \mathbb{E}[\frac{\partial\phi}{\partial z_i} f_i(x)] + \frac{1}{2}\sum_{i,j} \mathbb{E}[\left(\frac{\partial^2\phi}{\partial z_i \partial z_j}\right) 2\left[\sqrt{D(z)}\sqrt{D(z)}^T\right]_{ij}]$$

  where the SDE is given by $dz = f(z)dt + \sqrt{2D(z)}dW$, and ϕ is twice differentiable.

  - According to the framework of [YA Ma, 2015], we pick followings to remove MH step:

$$f(z) = -[D(z) + Q(z)]\nabla H(z) + \Gamma(z), \qquad \Gamma_i(z) = \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right)$$

  where $Q(z)$ is skew-symmetric, $D(z)$ is P.S.D matrix

Phenomenon analysis

- For the SGHMC, we can pick:

$$Q = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 0 & 0 \\ 0 & C \end{bmatrix}$$

such that it gives the following update rule: (Assume $H(\theta, r) = U(\theta) + \frac{1}{2} r^T M^{-1} r$)

$$\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \begin{bmatrix} M^{-1} r \\ -\nabla U(\theta) - C M^{-1} r \end{bmatrix} dt + \begin{bmatrix} 0 \\ N(0, 2C \cdot dt) \end{bmatrix}$$

- Now, let $\phi(\theta, r) = \theta^T \theta = \|\theta\|^2$, then, by Fokker-Planck equation :

$$\frac{d}{dt} \mathbb{E}[\|\theta\|^2] = 2M^{-1} \mathbb{E}[\theta^T r]$$

Phenomenon analysis

- Now, if we impose cold posterior effect, we get: (Note: $p^S(\theta) \propto \exp\left(-\frac{1}{T^2}U(\theta)\right)$)

$$\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \begin{bmatrix} M^{-1}r \\ -\nabla U(\theta) - rCM^{-1} \end{bmatrix} dt + \begin{bmatrix} 0 \\ T \cdot N(0, 2Cdt) \end{bmatrix}$$

where $D(\theta, r) = \begin{bmatrix} 0 & 0 \\ 0 & CT^2 \end{bmatrix}$, $Q(\theta, r) = \begin{bmatrix} 0 & -T^2 \\ T^2 & 0 \end{bmatrix}$, and $H(\theta, r) = \frac{1}{T^2}\left(\nabla U(\theta) + \frac{1}{2} r^T M^{-1} r\right)$

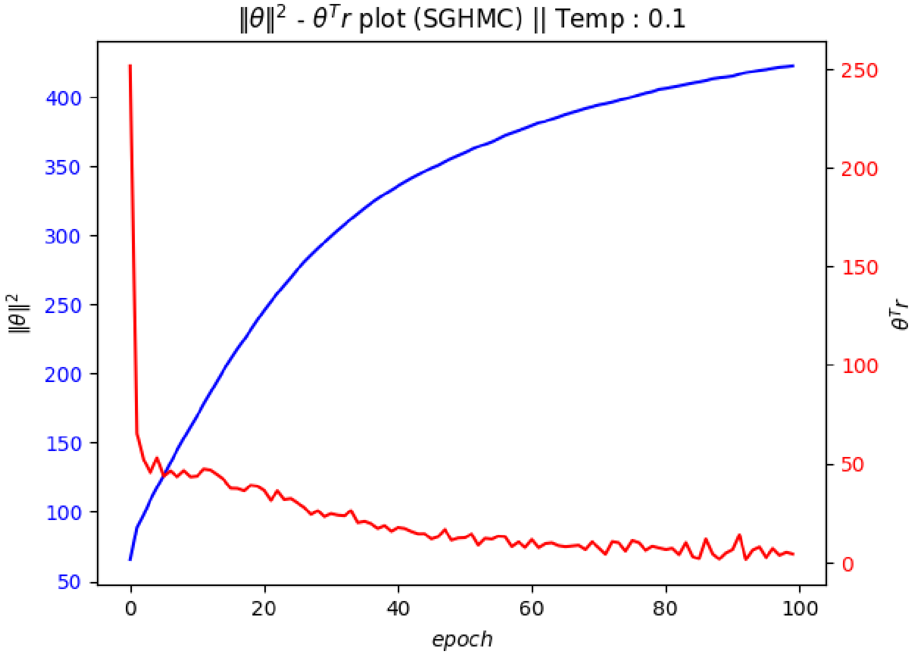
- By Fokker-Planck equation again, we have:

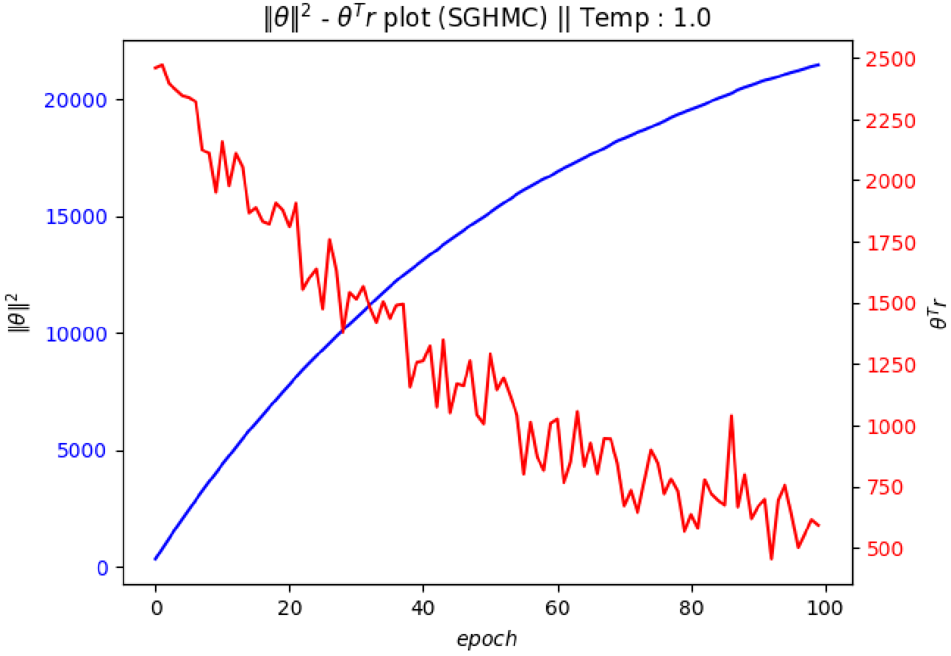
$$\frac{d}{dt}\mathbb{E}[\|\theta\|^2] = 2M^{-1}\mathbb{E}[\theta^T r], \qquad \frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$$

But, $\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T^2 \cdot tr(C)$ (= $2 \cdot tr(C)$ if w/o cold posterior)

Phenomenon analysis (Experiments)

- 1st question : does the $\frac{d}{dt}||\theta||_2^2 \propto \theta^T r$ in practice?



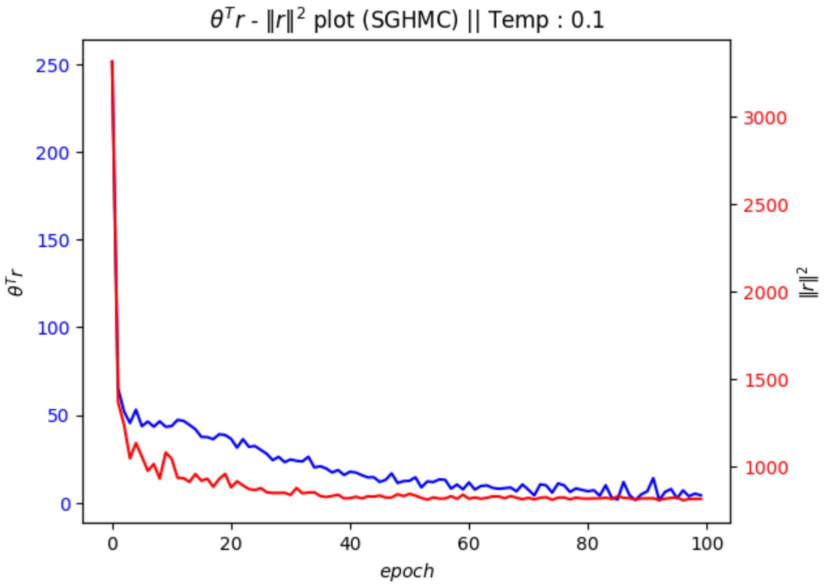


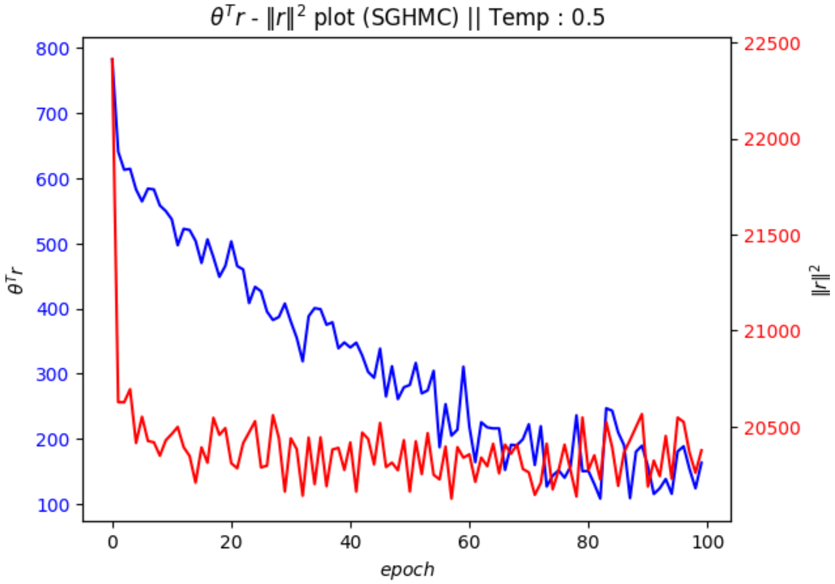
$\Rightarrow$ Yes, the behavior of $\theta^T r$ well represents the behavior of $\frac{d}{dt}||\theta||_2^2$.

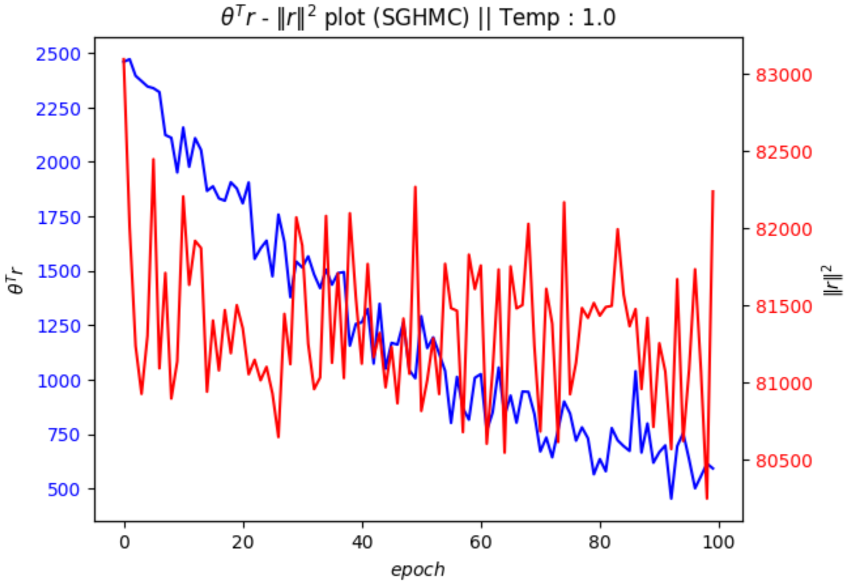
Phenomenon analysis (Experiments)

$$\frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$$

- 2^{nd} question : How much is the $\frac{d}{dt}\theta^T r$ dominated by $\|r\|^2$?







$\Rightarrow$ It seems that $\|r\|^2$ raise the starting point of $\theta^T r$, while $\|r\|^2$ remains almost constant

(if there is no momentum sampling) $\quad \frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T^2 \cdot tr(C)$

- Also, observe that colder T gives smaller $\|r\|^2$ in average, which corresponds to our analysis.

Phenomenon analysis (Experiments)

- When we observe :

$$\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T^2 \cdot tr(C)$$

since $2T^2 \cdot tr(C) \gg 0$, the cold temperature can effectively regularize $\|r\|^2$.

(or helps to form a smaller equilibrium point of $\|r\|^2$, which slow down the increasing speed of $\|\theta\|^2$)

Phenomenon analysis (Experiments)

$$\frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$$

- Then, how to avoid the $\|\theta\|^2$ burst?

1. make $\theta^T r$ suppressed → requires $\mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)] \cong 0$ or < 0.
2. Since $M^{-1}\|r\|^2$ can take a large portion in practice, we need to regularize $\|r\|^2$.
3. Recalling $\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T^2 \cdot tr(C)$,

① Set $T \ll 1$ (which leads to wrong stationary distribution)

② Lower the friction coefficient C (but, this also leads to conflicting behavior; $|r^T CM^{-1}r|$ ↘)

③ Frequent momentum resampling nearby $\mathbf{0}$, forcing to $\|r\|^2 \cong 0$ regularly.

④ **(Best)** devise new SGMCMC which adopts a parameter solely taking charge of T term.

Phenomenon analysis (Experiments)

$$\frac{d}{dt}\mathbb{E}[\|\theta\|^2] = 2M^{-1}\mathbb{E}[\theta^T r]$$

$$\frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$$

$$\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T^2 \cdot tr(C)$$

- One simple heuristic is to adopt adjusting factor:
  1. Use adjusting factor $\gamma \ll 1$ (ex: 10^{-3}) such that:

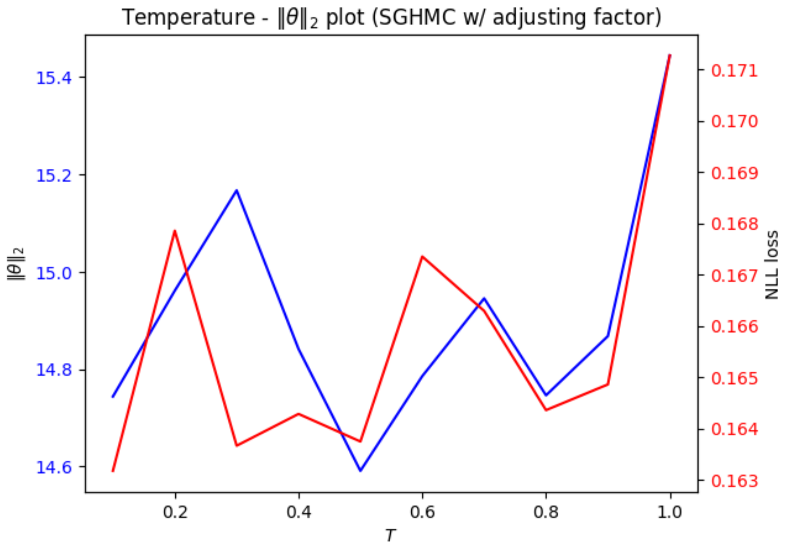
     $$C' = C\gamma, \qquad M' = M/\gamma, \qquad \epsilon' = \epsilon/\gamma \text{ (optional)}$$

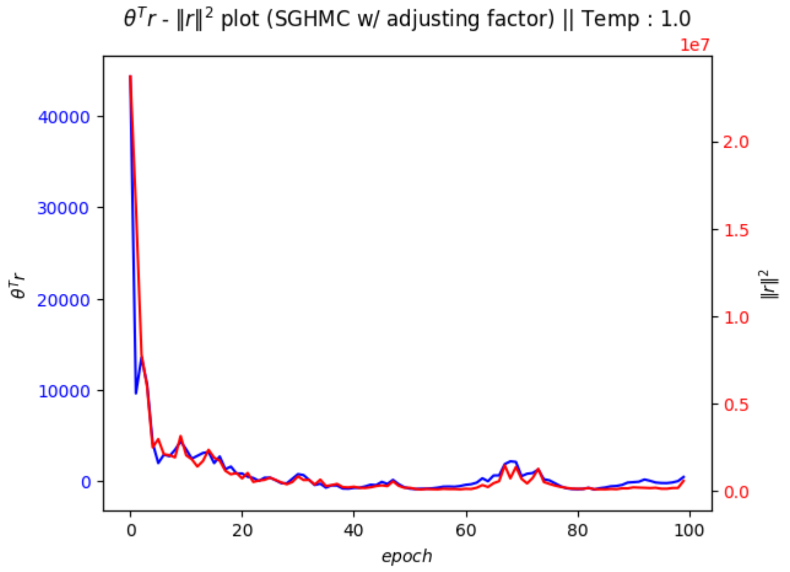
  2. Then, while the effect of term $2T^2 \cdot tr(C)$ can be minimized, the driving term $CM^{-1}r$ can be remained unaffected.

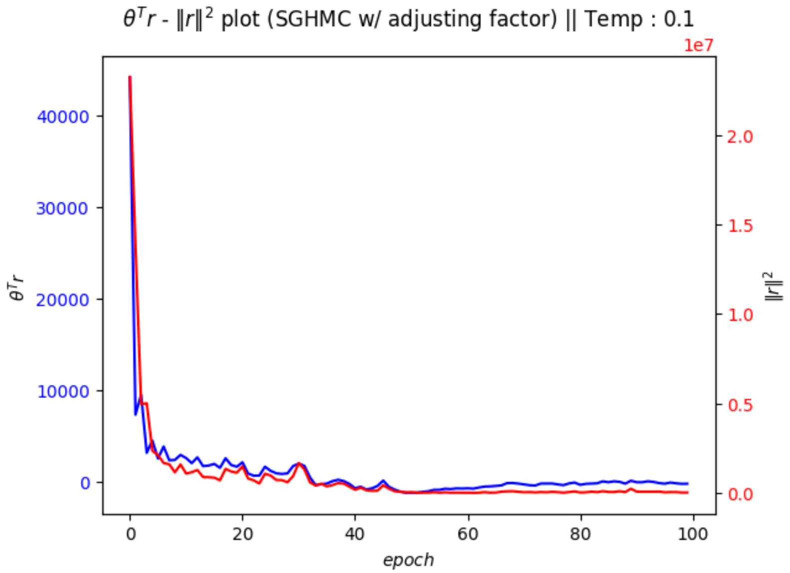
     (※ very low γ can lead to unstable optimization due to numerical errors.)

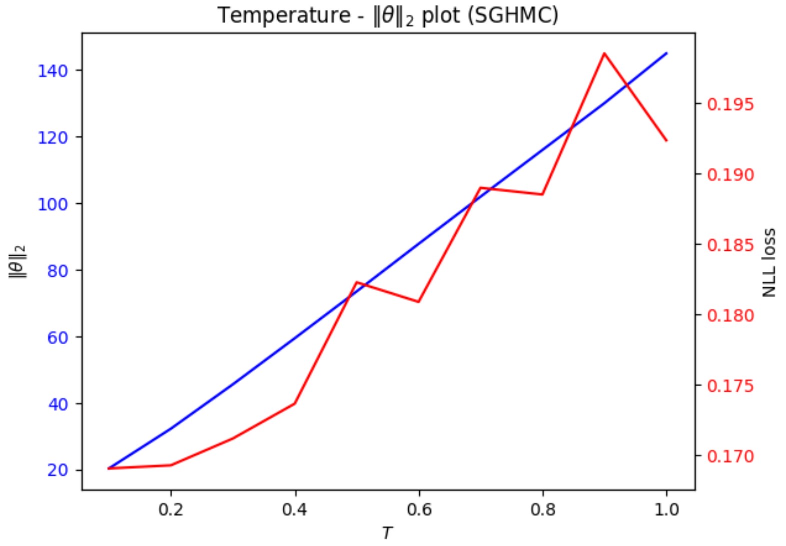
  3. Plus, the frequent momentum resampling (per one ensemble) can help to escape from the effect of temperature. (sometime $\theta^T r$ becomes negative)

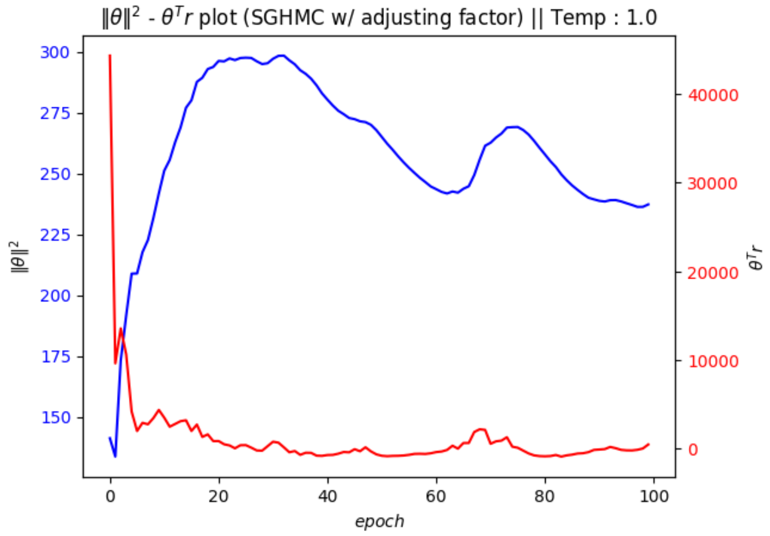
Phenomenon analysis (Experiments)

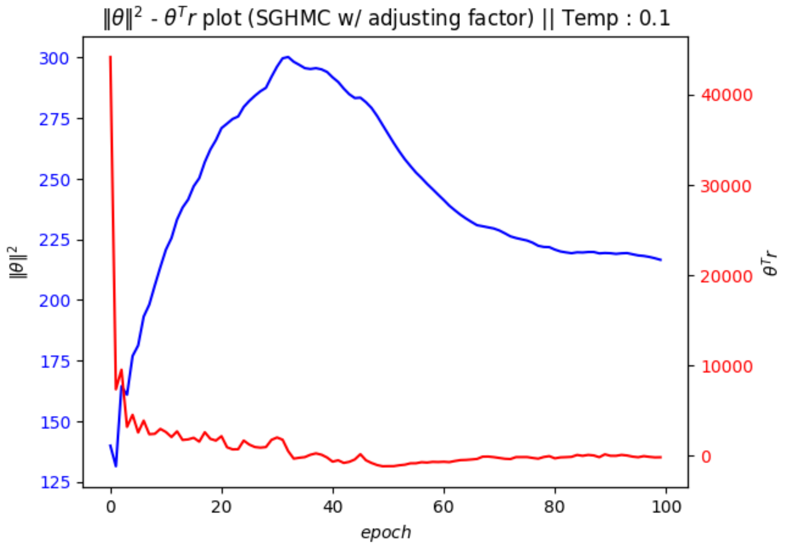












Phenomenon analysis (Experiments)

- New updating rule to boost the mixing: $\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \begin{bmatrix} M^{-1}r \\ -\alpha\nabla U(\theta) - CM^{-1}r \end{bmatrix} dt + \begin{bmatrix} 0 \\ N(0, 2C\alpha \cdot dt) \end{bmatrix}$

- Observation :

1. when $\alpha \to 0$, it becomes $\begin{bmatrix} d\theta \\ dr \end{bmatrix} \cong \begin{bmatrix} M^{-1}r \\ -CM^{-1}r \end{bmatrix} dt \Rightarrow M\frac{d^2\theta}{dt^2} = -CM^{-1}r$ (= exact friction force)

2. when $\alpha \gg 1$, it becomes $\begin{bmatrix} d\theta \\ dr \end{bmatrix} \cong \begin{bmatrix} M^{-1}r \\ -\alpha\nabla U(\theta) \end{bmatrix} dt + \begin{bmatrix} 0 \\ N(0, 2C\alpha \cdot dt) \end{bmatrix}$, or $\frac{d^2\theta}{dt^2} \cong -M^{-1}\alpha\nabla U(\theta) + \sqrt{2C\alpha}dW$

   (Note : when $C = 0$ on naïve SGHMC, $\frac{d^2\theta}{dt^2} \cong -M^{-1} \cdot \nabla U(\theta)$)

- If $\alpha \cong 0$, it implies that θ gradually stops w/o being affected by $U(\theta)$.

- If $\alpha \gg 1$, then, the direction of driving force $\frac{d^2\theta}{dt^2}$ is aligned toward $-M^{-1}\nabla U(\theta)$ with some noise.

Phenomenon analysis (Experiments)

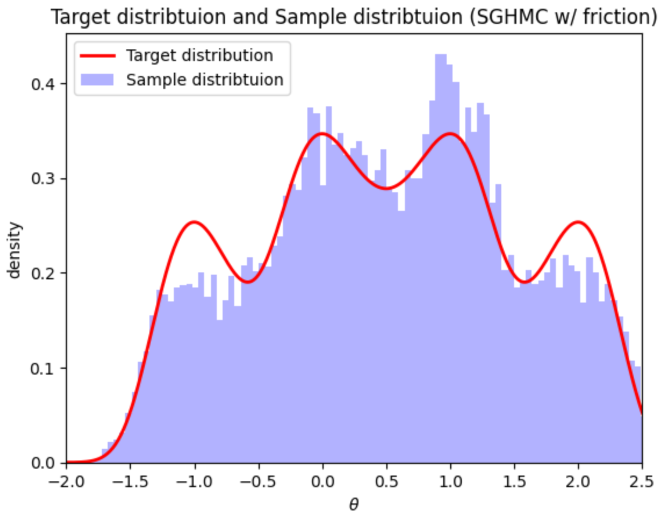
- Let's combine both method : γ : adjusting factor / α : boosting factor

$$\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \begin{bmatrix} \gamma M^{-1} r \\ -\alpha \nabla U(\theta) - \gamma^2 C M^{-1} r \end{bmatrix} dt + \begin{bmatrix} 0 \\ N(0, 2\gamma\alpha C \cdot dt) \end{bmatrix}$$

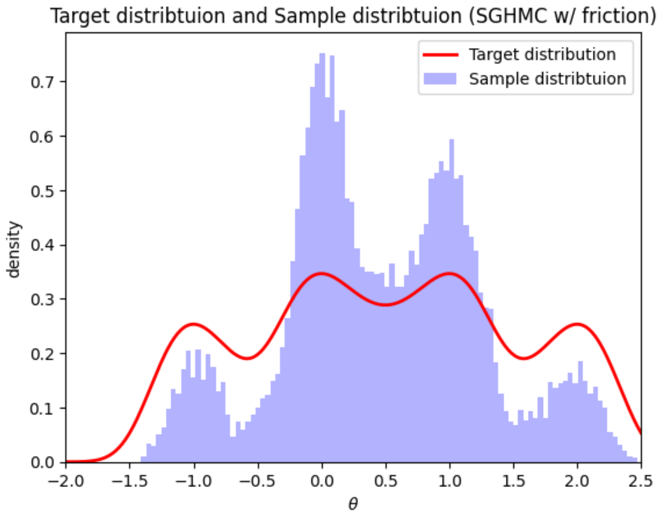
- Lower γ leads to reduce weight norm, and help to sample around local modes.

- Higher α resolves the sticky movement induced by low γ, providing fast mixing.

- During the process, momentum resampling is crucial to jump into another mode



Baseline : $\alpha = 1, \gamma = 1$



Baseline : $\alpha = 5, \gamma = 0.5$