L2 norm burst during BNN training (2)

-Summary-

23/09/07

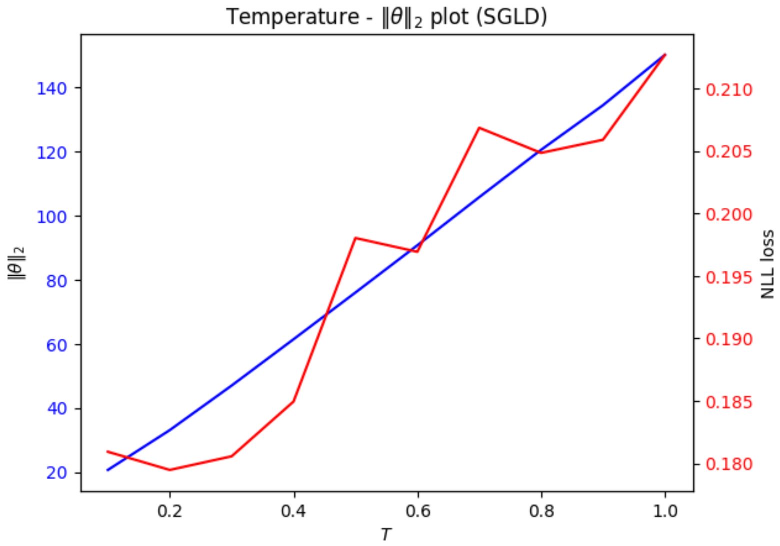
Observed problem (Review)

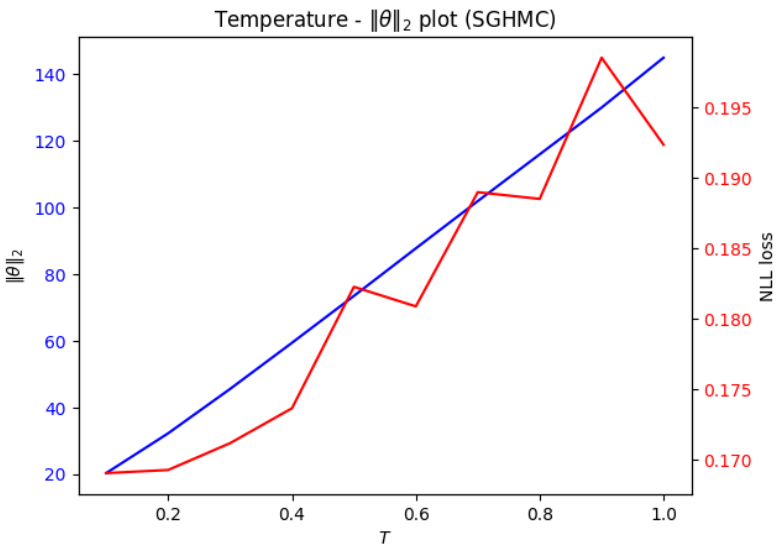
- During training of DNN, burst of weight L_2 norm is observed while the test accuracy is maintained high, or NLL is reasonably low.
- One strategy to avoid this issue is to adopt cold tempered posterior:

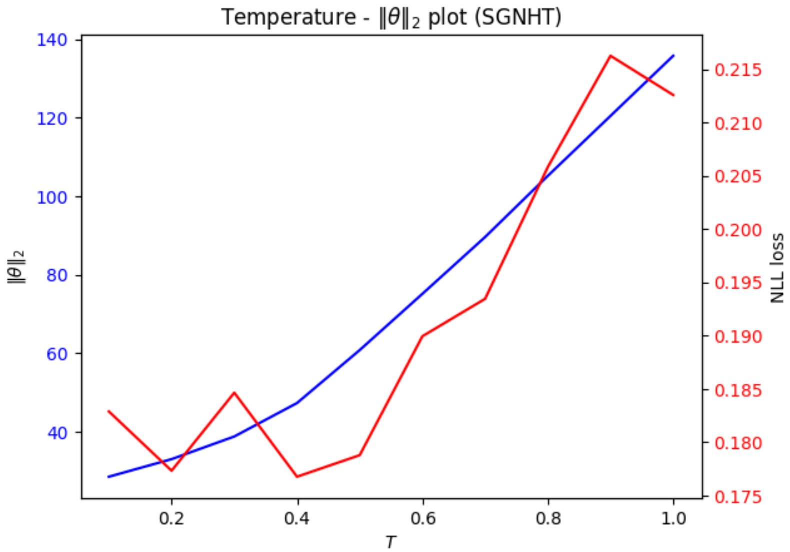
$$p_T(\theta|\mathcal{D}) \propto \exp\left(-\frac{U(\theta)}{T}\right)$$

- 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 $T \rightarrow 0$, the only highest mode of $p(\theta|\mathcal{D})$ survive, which results in MAP training (≅ SGD w/ weight decaying if prior is isotropic gaussian)

Observed problem (Review)







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

Observed problem (Review)

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)

Observed problem (Review)

- 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

Observed problem (Review)

- 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 :

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

Observed problem (Review)

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

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

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

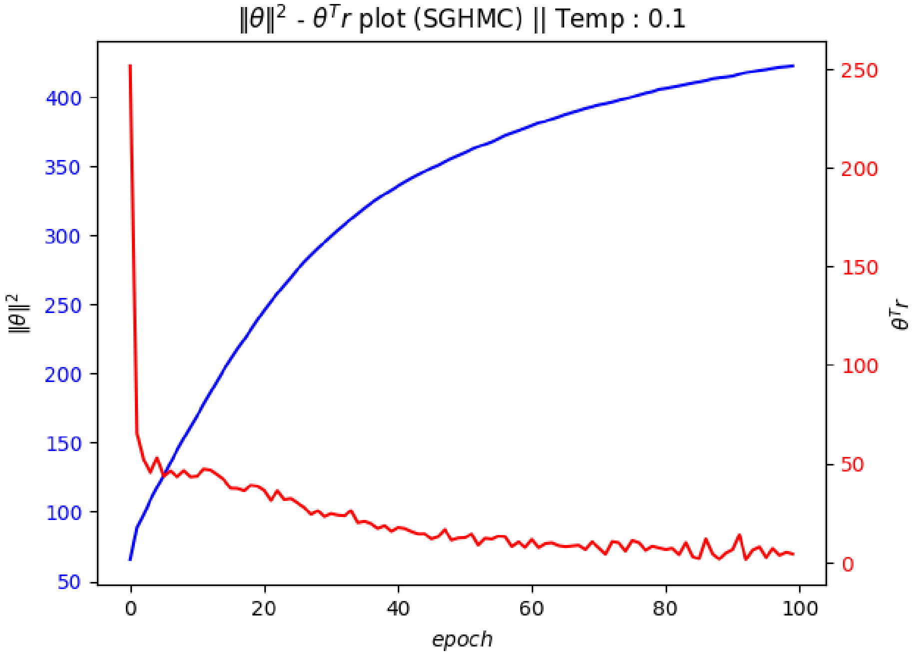
- By Fokker-Planck equation again, we have:

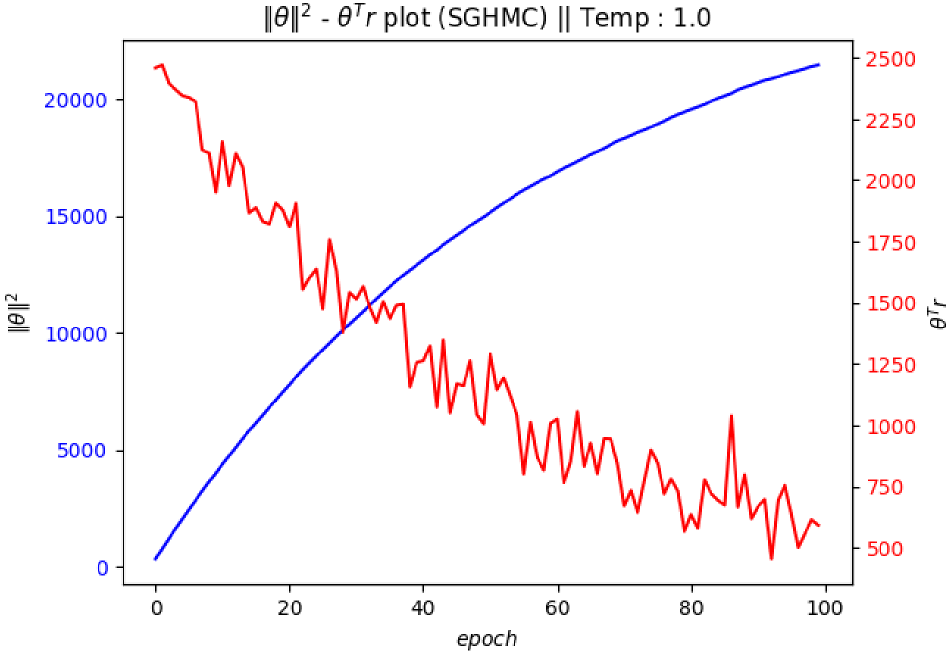
$$\frac{d}{dt}\mathbb{E}[\|\theta\|^2] = 2M^{-1}\mathbb{E}[\theta^T r], \qquad \frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$$

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

Observed problem (Review)

- 1st question : does the $\frac{d}{dt}\|\theta\|_2^2 \propto \theta^T r$ in practice? (No momentum sampling)

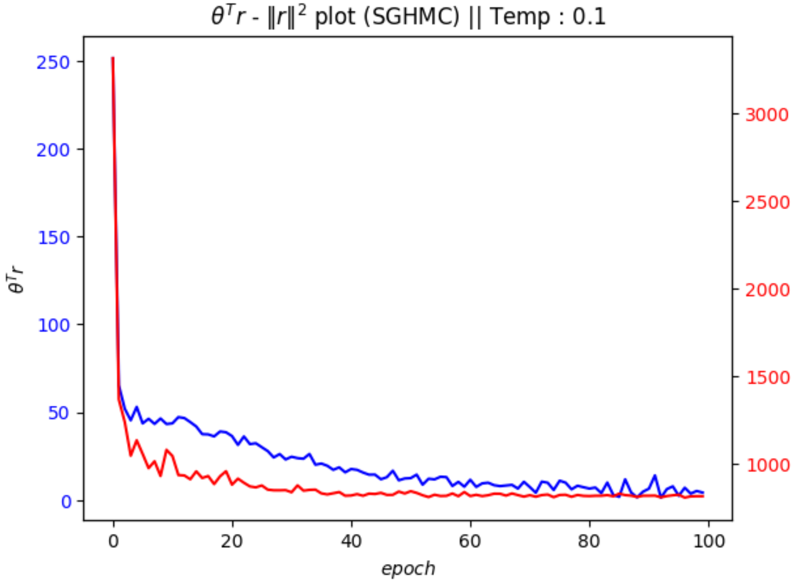


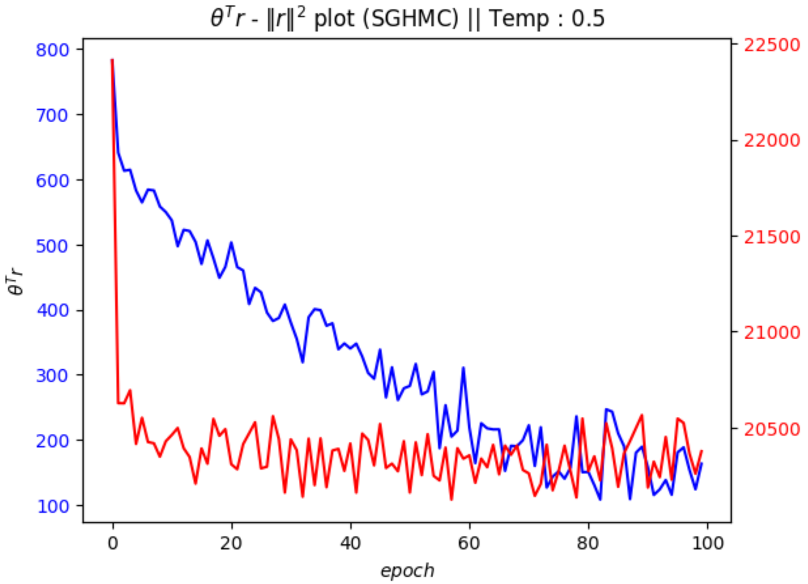


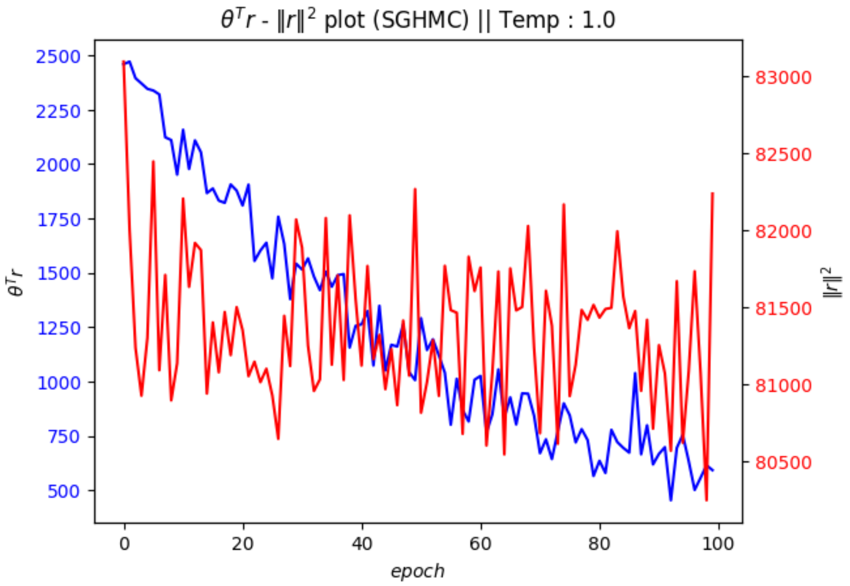
$\Rightarrow$ Yes, the behavior of $\theta^T r$ well represents the behavior of $\frac{d}{dt}\|\theta\|_2^2$.

Observed problem (Review)

- 2nd question : How much is the $\frac{d}{dt}\theta^T r$ dominated by $\|r\|^2$? (No momentum sampling)







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

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

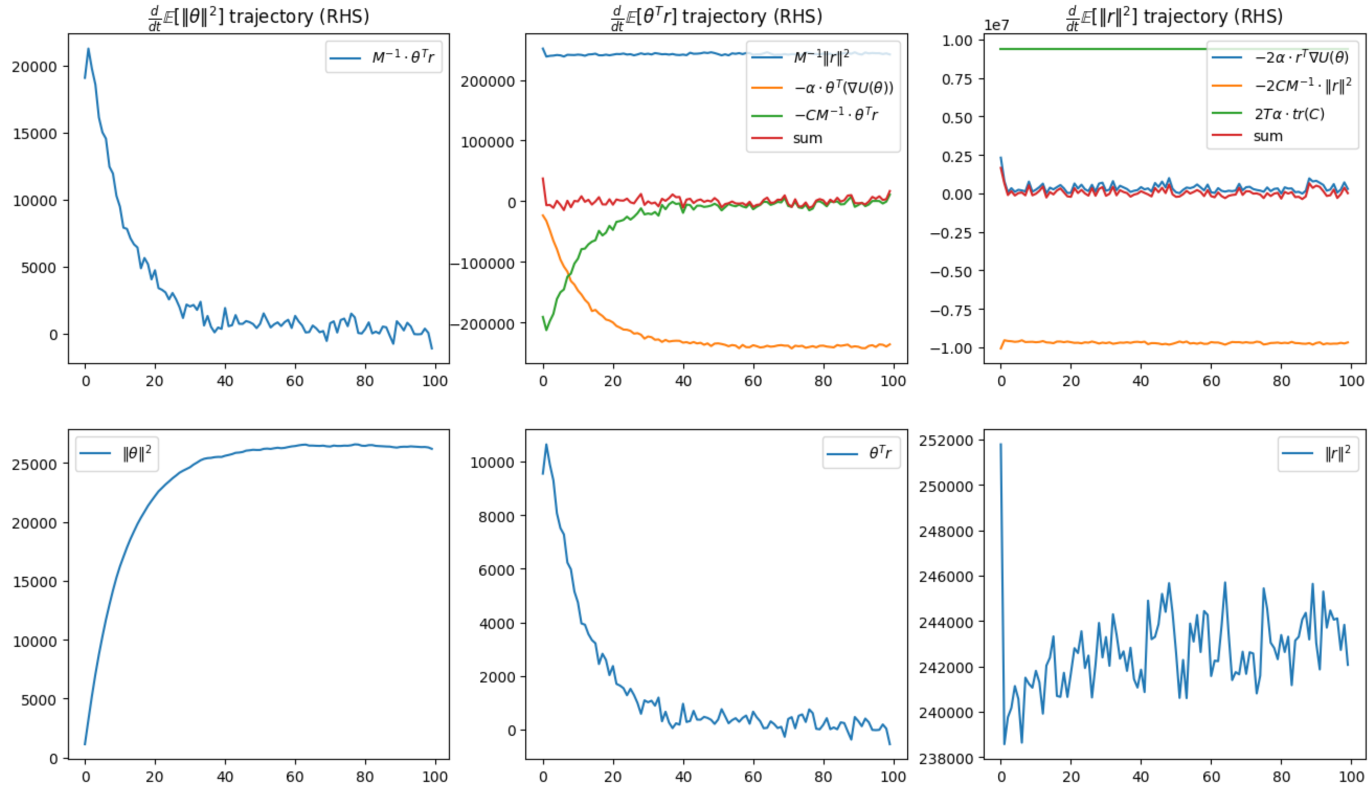
$$\frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)] \qquad \frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T \cdot tr(C)$$

Observed problem

- 3rd Question : Our analysis can explain the behaviors of $\|\theta\|^2, \theta^T r, \|r\|^2$

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

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



Interpretation: (here, $\alpha = 1$, with no momentum resampling)

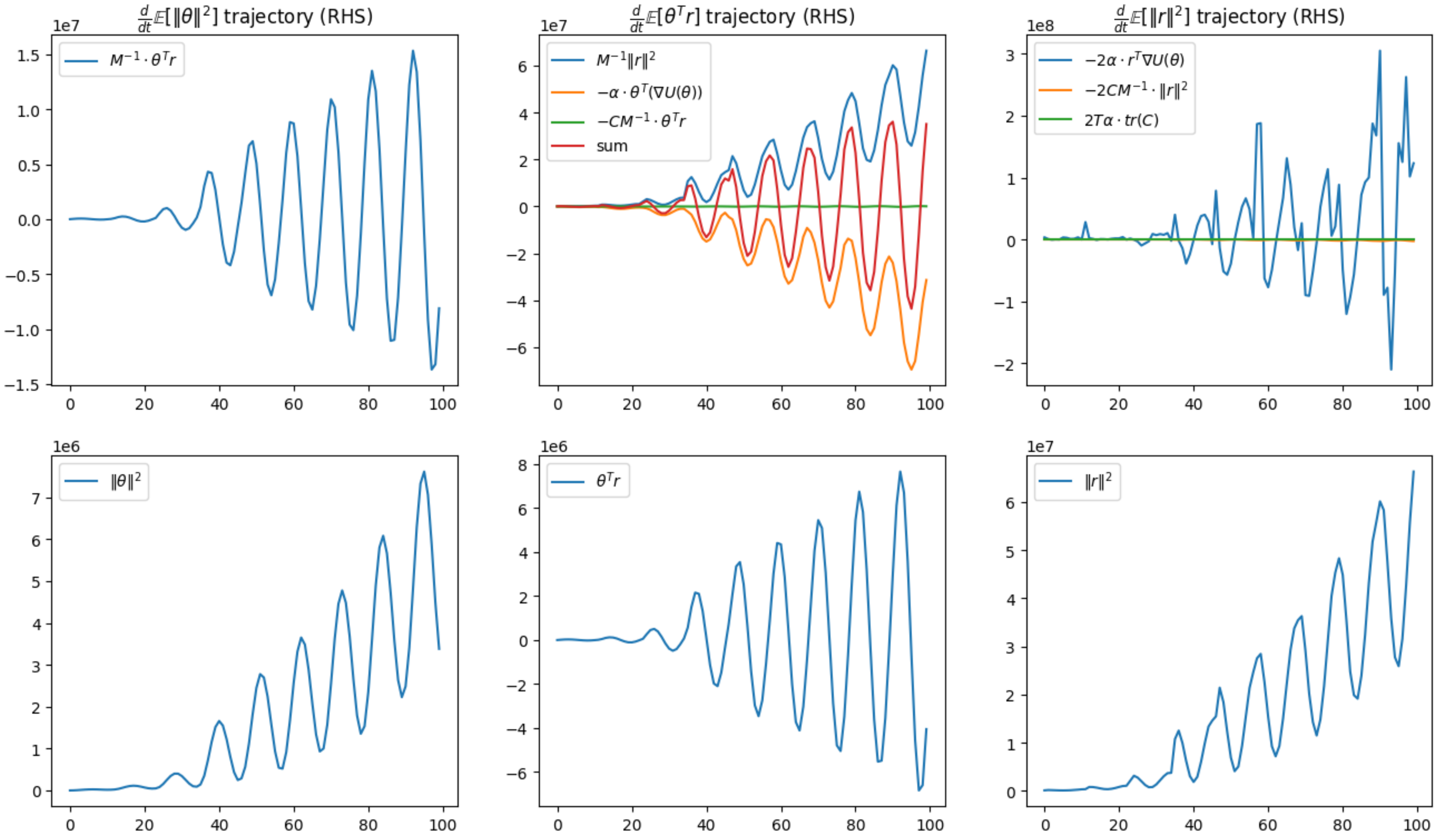
1. **There is a plateau of $\|\theta\|^2$, which implies the gradient actually does not burst exponentially.**
2. $\frac{d}{dt}\mathbb{E}[\|\theta\|^2] = 2M^{-1}\mathbb{E}[\theta^T r]$ **relation clearly holds.**
3. **However,** $\frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$**, and** $\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T \cdot tr(C)$ **relation is unclear in practice.**

   ∵ (1st) approximation of $\nabla U(\theta)$ / (2nd) Expectation over noise

Observed problem

- (Additional Question) : Is our analysis right in practice? → Not sure.

(For some parameters (w/o momentum resampling), the behavior cannot be well-explained by our analysis)



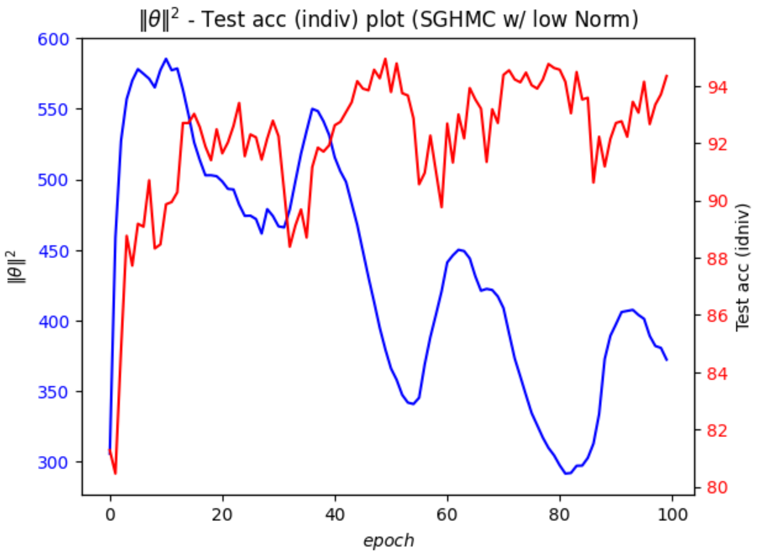
Why cold posterior is good?

- Back to the original problem, why does the cold posterior gives good performance?
  - There are several suspected reasons behind many studies : bad prior for θ, existence of data-augmentation, (or C.P is actually not effective), …. But, the conclusion is that the reason is unclear.
  - **Our hypothesis is that the sample drawn from SG-MCMC tends to have higher weight norm compared to the samples drawn from cold posterior**.
  - Then, does the samples drawn from SG-MCMC with regularized norm can give samples which gives higher test acc? $\rightarrow$ YES.

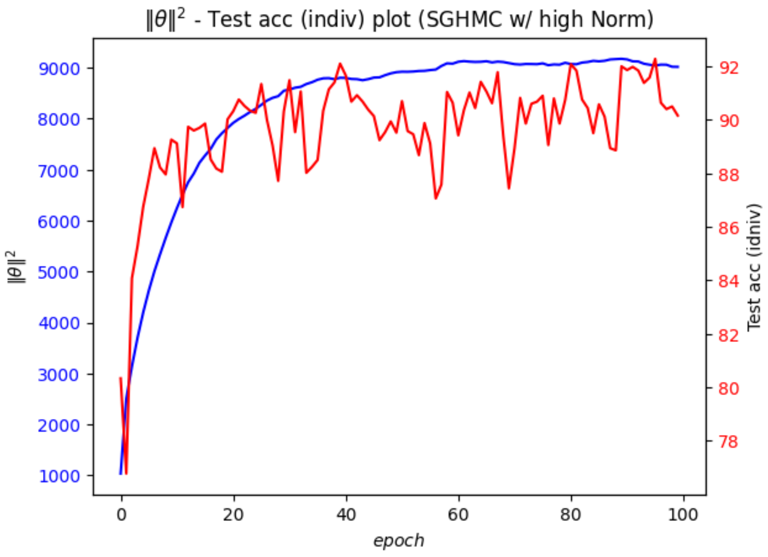
Why cold posterior is good?

- Question : Does the samples drawn from SG-MCMC with regularized norm can give samples which tends to give higher test acc? → YES.

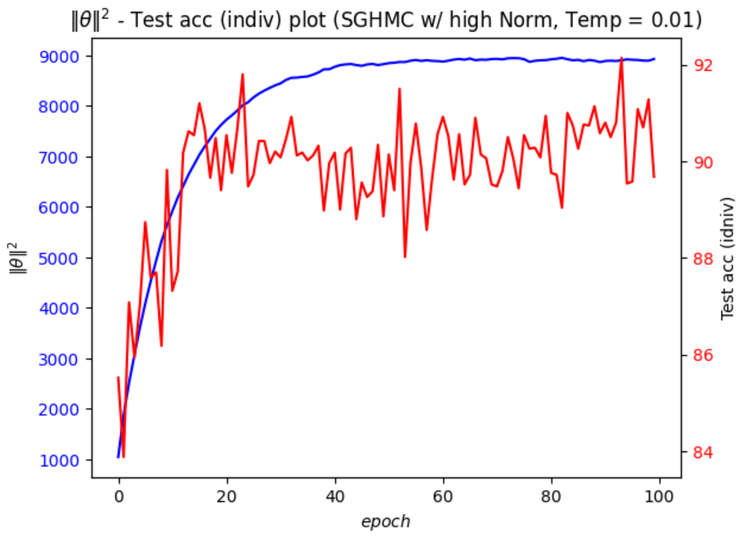
- Low norm : $||\theta||^2$ oscillates around 370 / High norm : $||\theta||^2$ stabilized around 9000



Averaged NLL : 0.16xx (Low norm)



Averaged NLL : 0.19xx (High norm)



Averaged NLL : 0.18xx (High norm /w T=0.01)

- **Furthermore, the performance of cold posterior degrades if the weight norm $||\theta||^2$ is high.**

Norm-adjusting SGHMC (heuristic)

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

(= equivalent to changing mass)

$$\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \begin{bmatrix} \alpha^{-1} \cdot M^{-1}r \\ -\nabla U(\theta) - \alpha^{-1}CM^{-1}r \end{bmatrix} d(t/\alpha) + \begin{bmatrix} 0 \\ N(0, 2Cd(t/\alpha)) \end{bmatrix}$$

- 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 $M\frac{d^2\theta}{dt^2} \cong -\alpha\nabla U(\theta) + \sqrt{2C\alpha}dW$

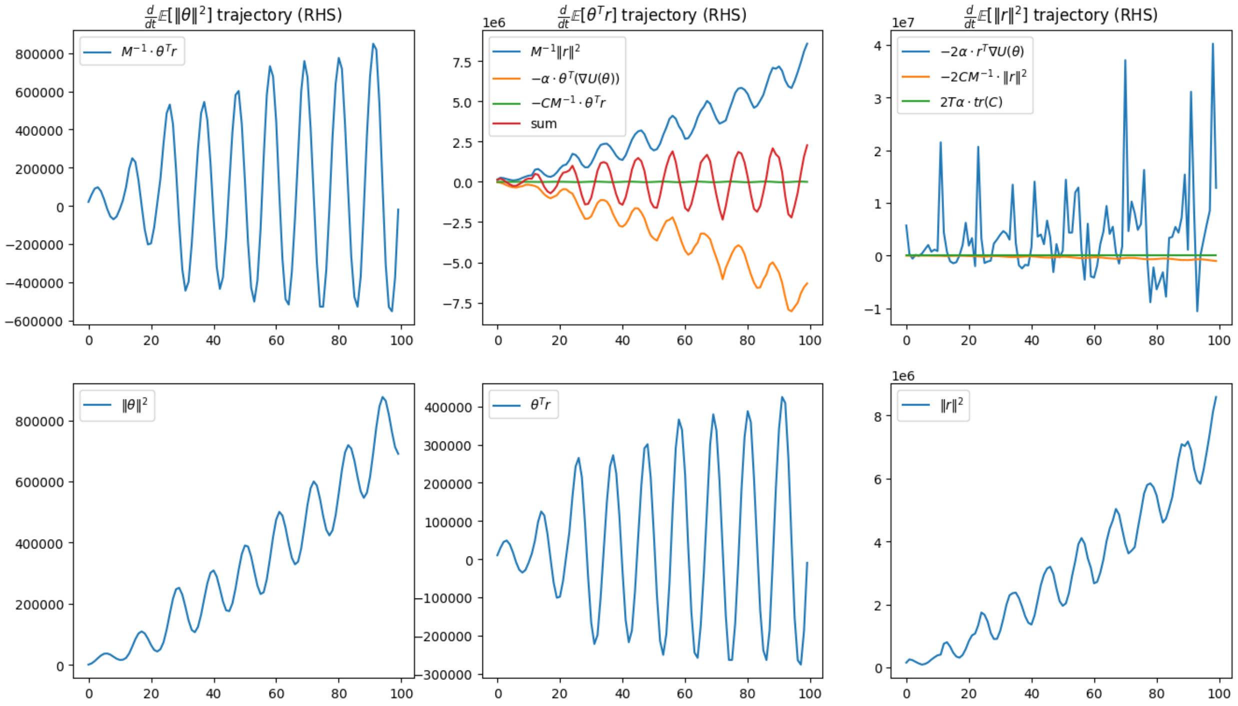
   (Note : when $C = 0$, $\frac{d^2\theta}{dt^2} \cong -M^{-1} \cdot \nabla U(\theta)$ (= exact dynamics driven by potential $U(\theta)$))

  - If $\alpha \cong 0$, it implies that θ gradually stops w/o being affected by $U(\theta)$.
  - If $\alpha \gg 1$, the sampling heavily relies on dyanmics driven by $U(\theta)$ (with some increased noise).

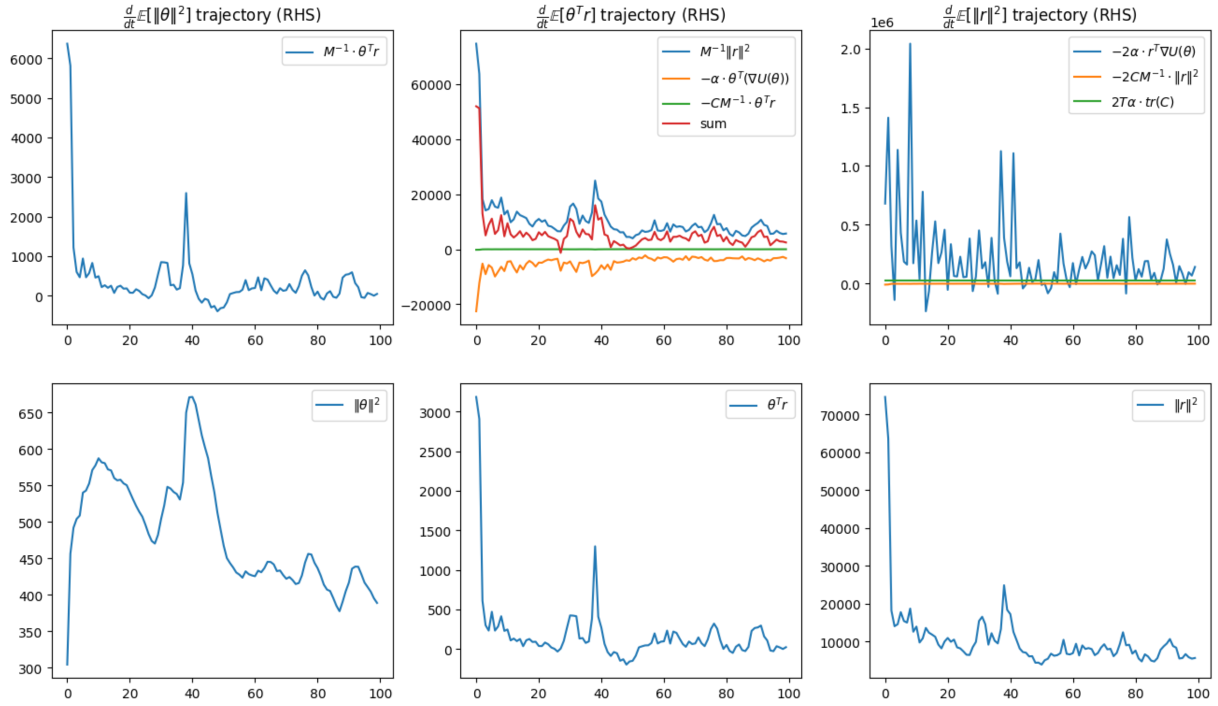
The increased noise is necessary to obtain stationary distribution $p^s(\theta) \propto \exp(-U(\theta))$

Norm-adjusting SGHMC (heuristic) observations

- The SGHMC method : $\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \begin{bmatrix} \alpha^{-1} \cdot M^{-1} r \\ -\nabla U(\theta) - \alpha^{-1}\gamma C M^{-1} r \end{bmatrix} d(t/\alpha) + \begin{bmatrix} 0 \\ N(0, 2C\gamma d(t/\alpha)) \end{bmatrix}$
    1. Boosting factor α ($\cong 3$), Adjusting factor $\gamma(\cong 0.001)$ gives an oscillating behaviors of $||\theta||^2, \theta^T r, ||r||^2$, which enables the $||\theta||^2$ to decrease regardless of temperature T.



Norm-adjusting SGHMC (w/o momentum resampling)



Norm-adjusting SGHMC (w/ momentum resampling)

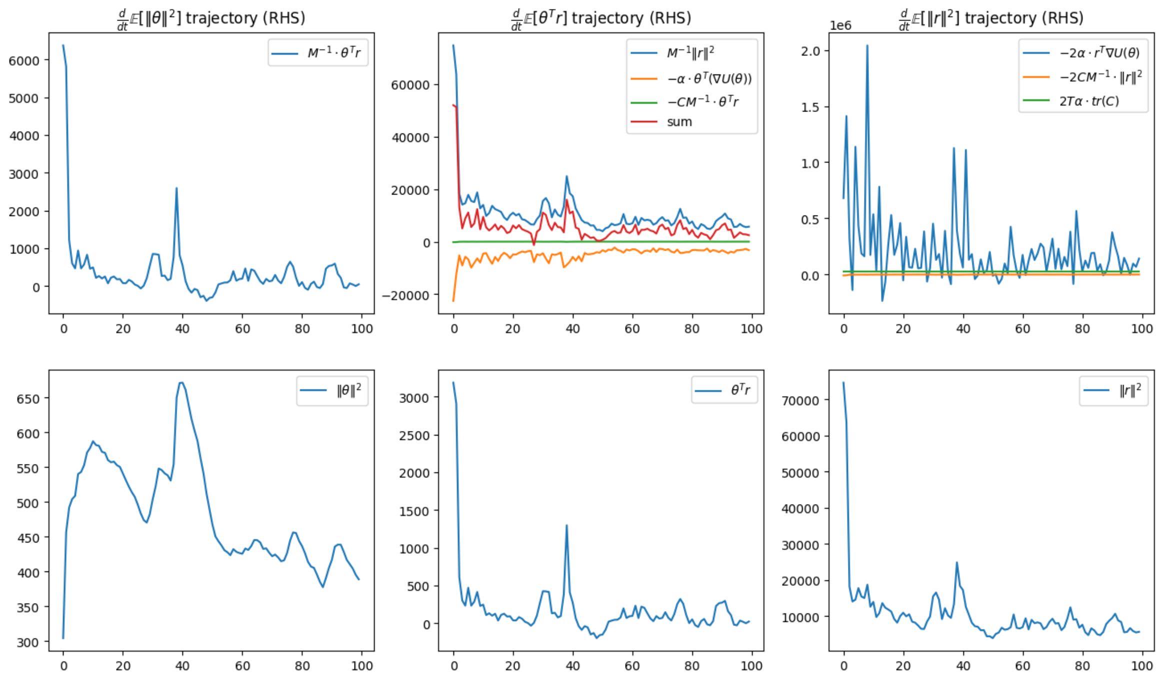
Norm-adjusting SGHMC (heuristic) observations

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

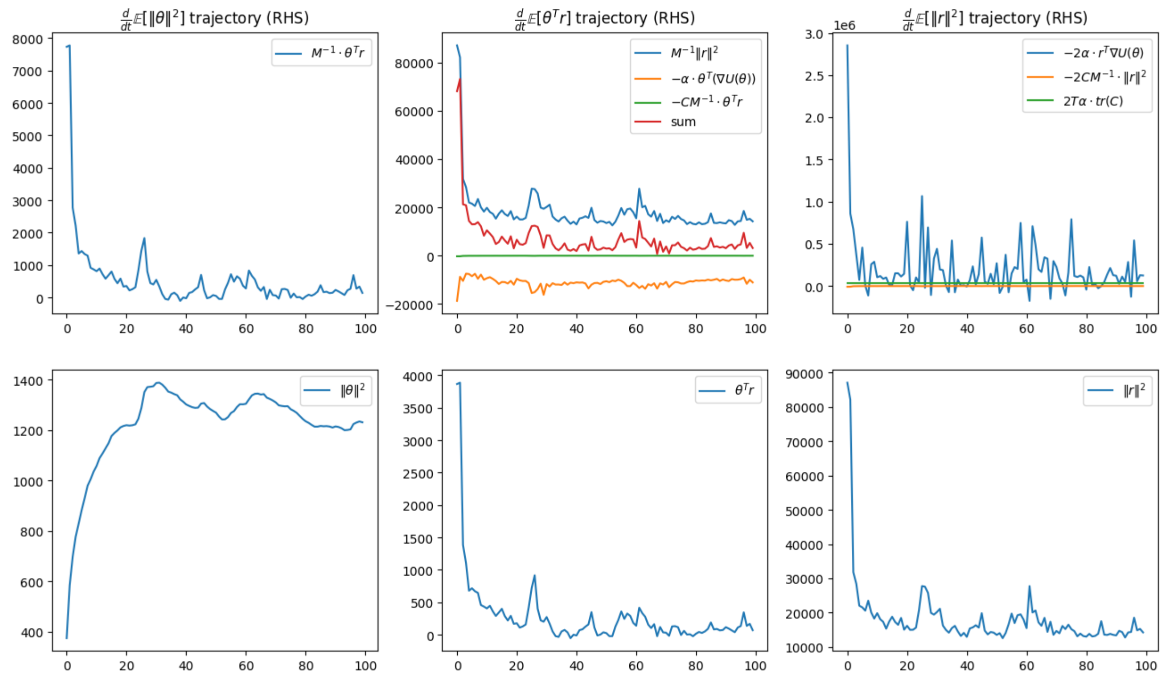
  2. Another **very crucial heuristic** is to adopt scaled momentum resampling:

$$r \sim N(0, \beta M)$$

where β **: momentum resampling scaler ($\cong \mathbf{0.001}$)**



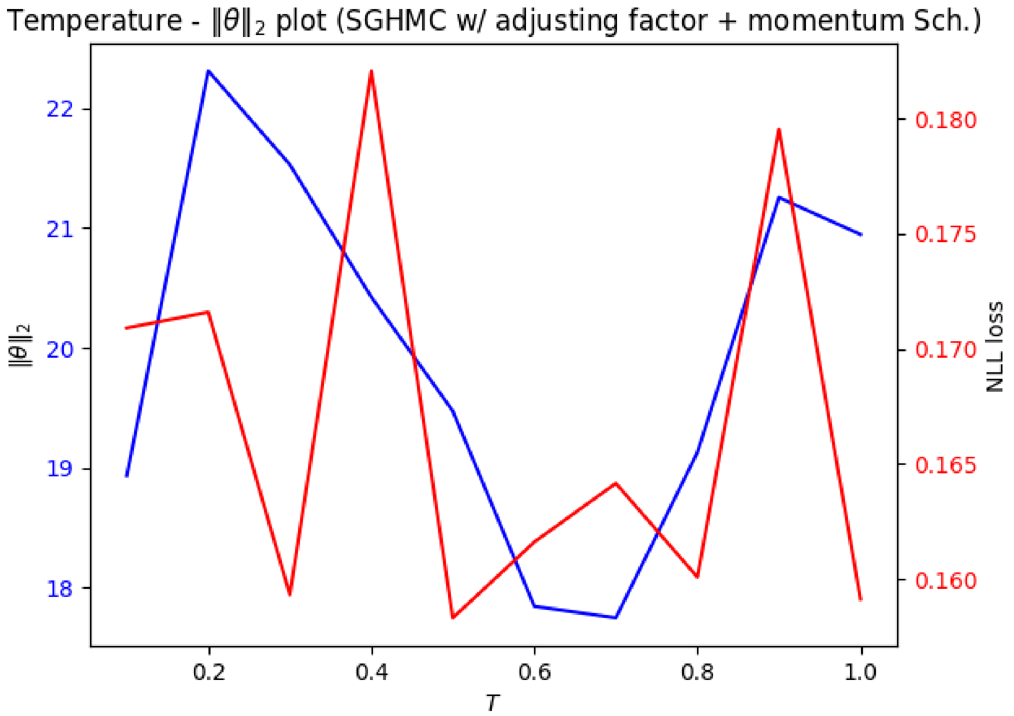
Norm-adjusting SGHMC (w/ $\beta = 0.001$)



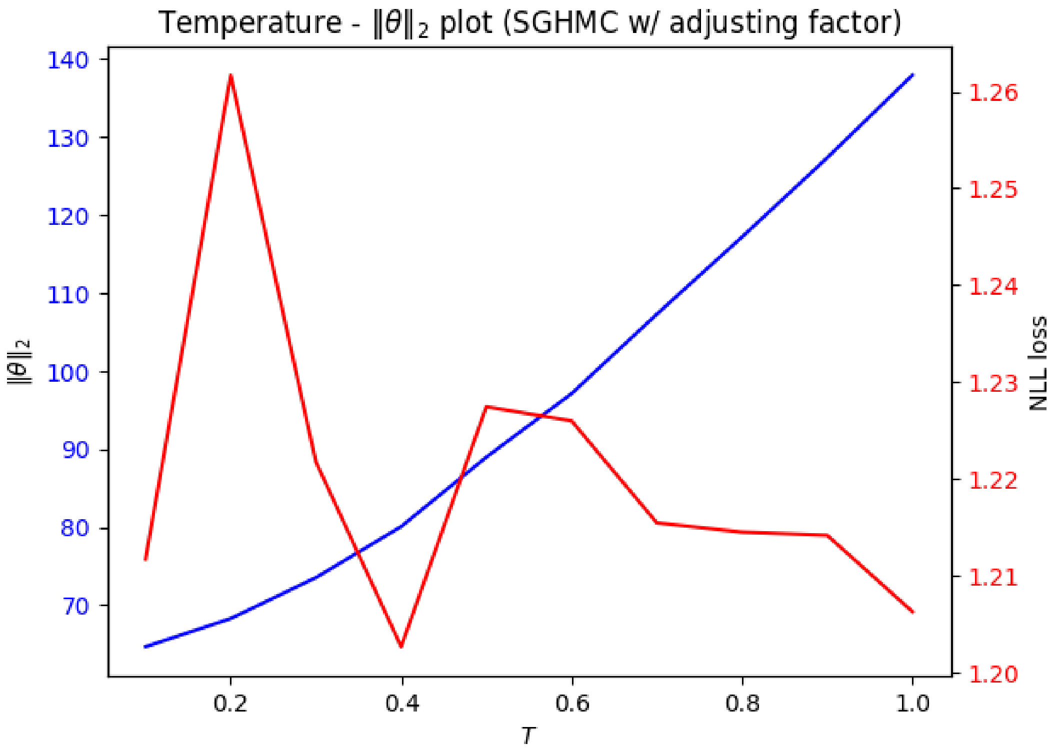
Norm-adjusting SGHMC (w/ $\beta = 0.1$)

Norm-adjusting SGHMC (heuristic) observations

- Results of revised method



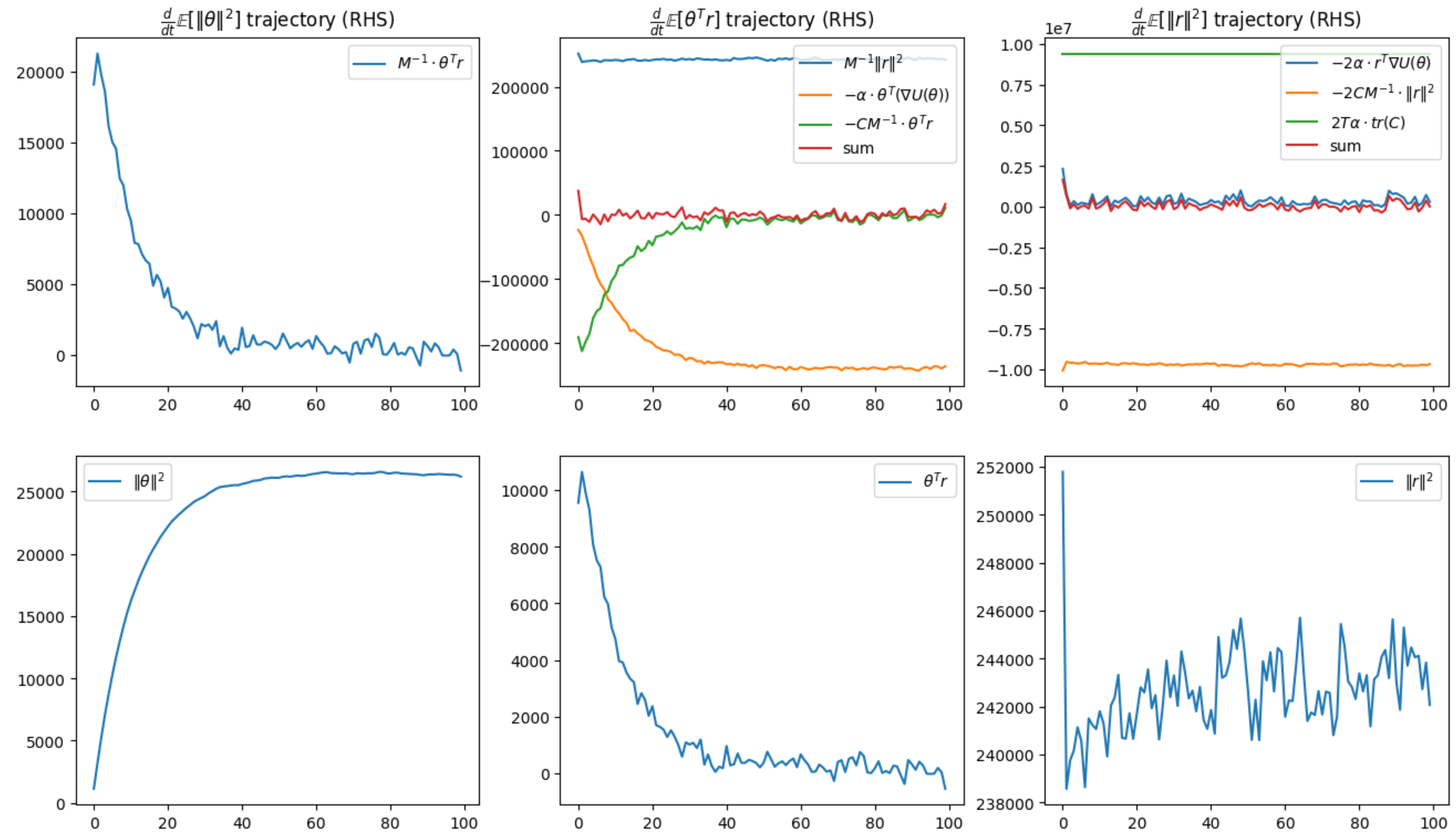
MNIST

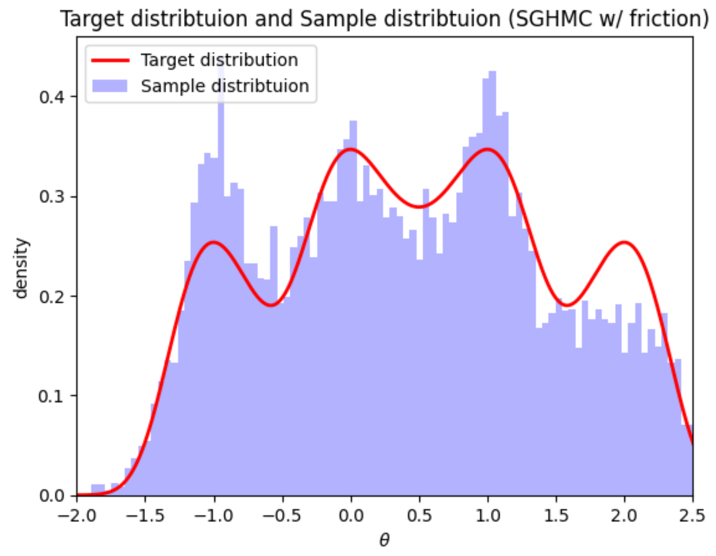


CIFAR-10 (w/o data augmentation)

- We can further regularize $\|\theta\|_2$ by controlling hyperparameter α, β, γ.

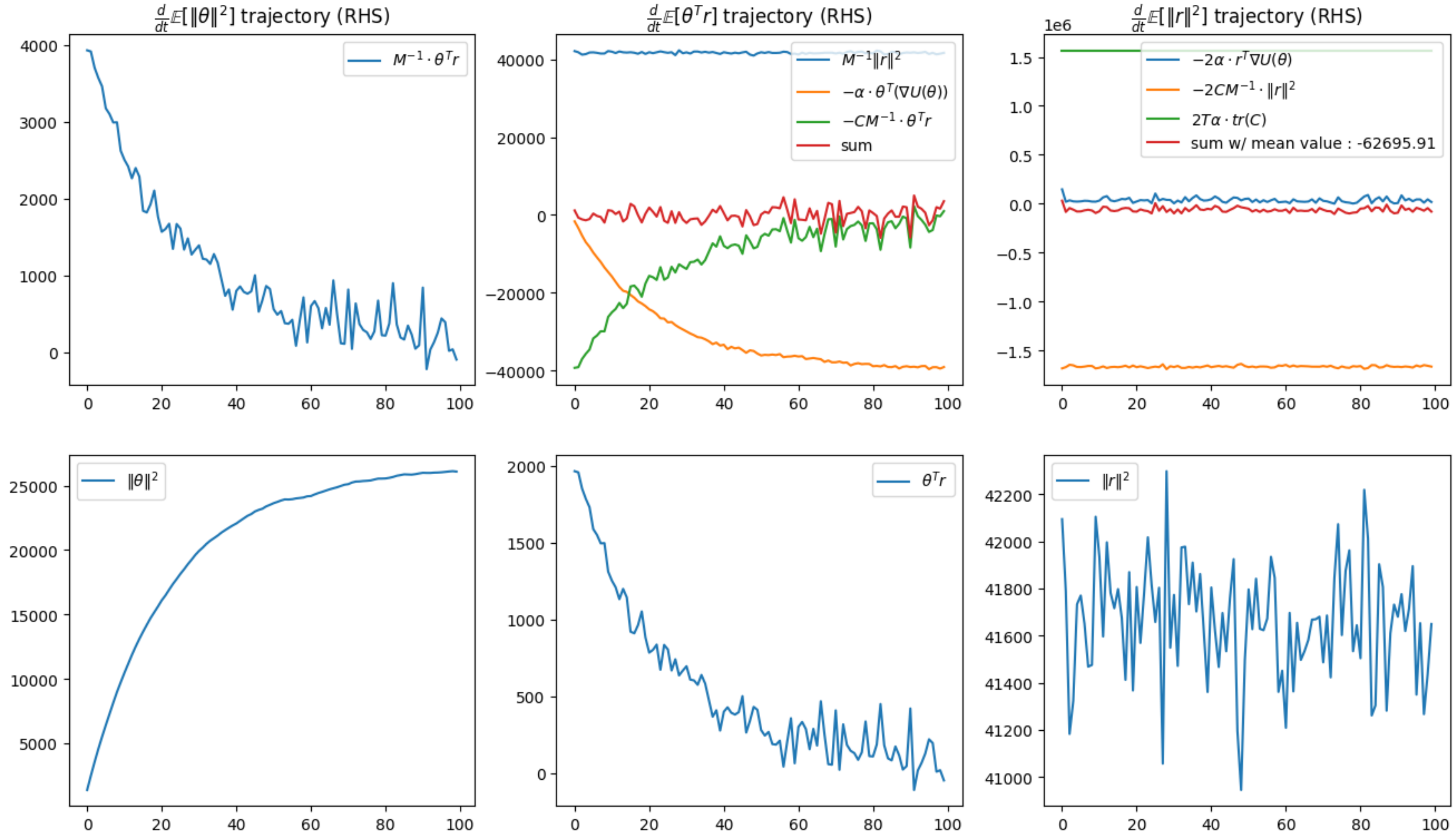
Phenomenon analysis (Experiments)

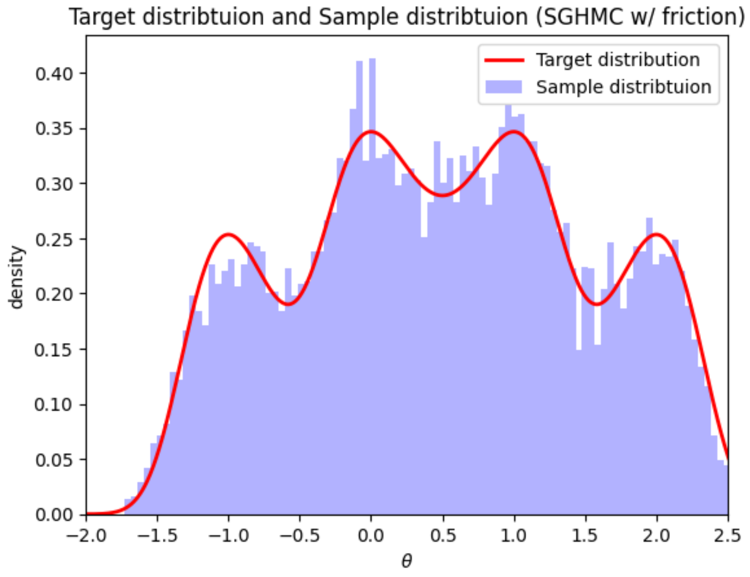




$\alpha = 1, \gamma = 1$

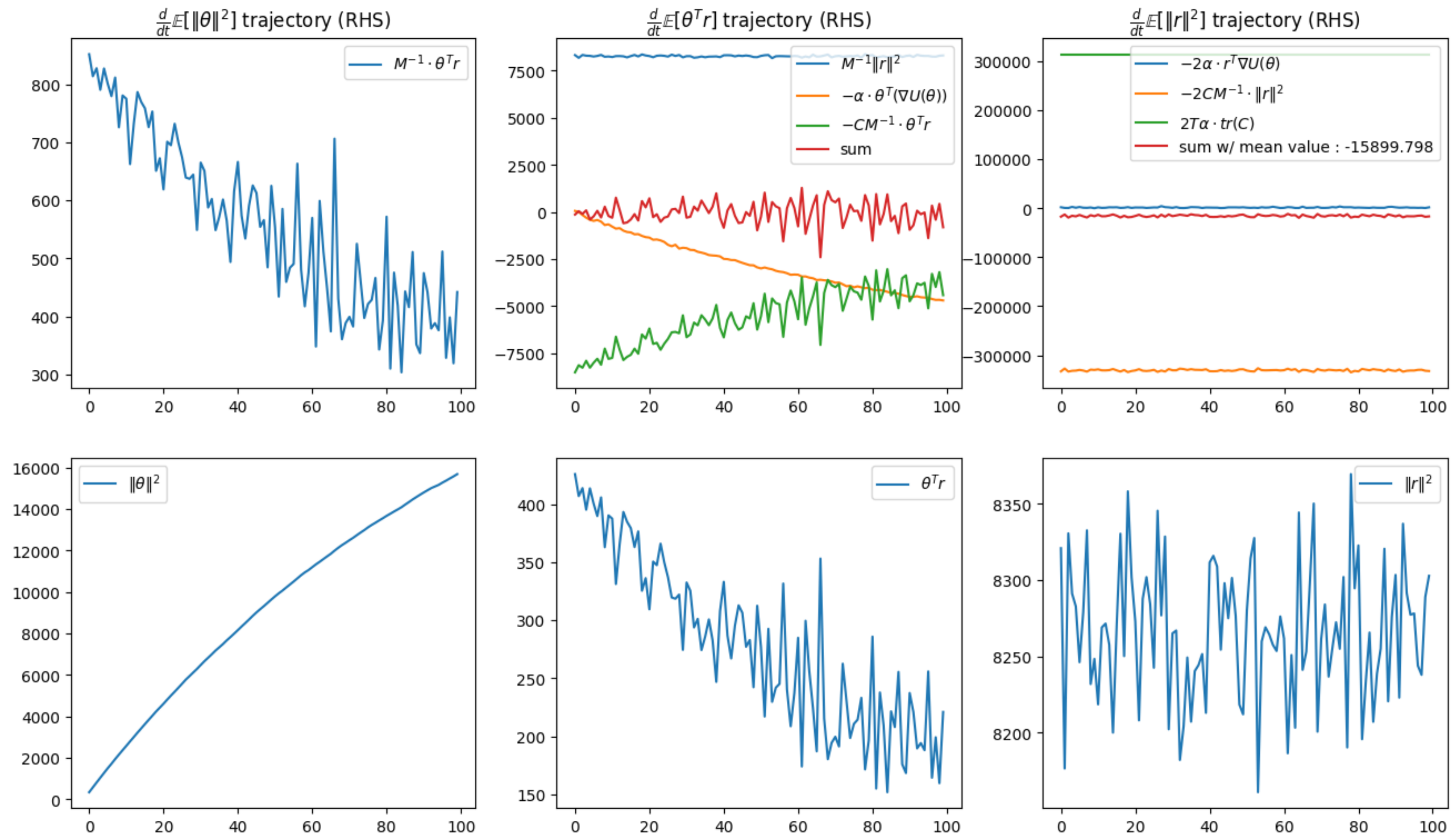
Phenomenon analysis (Experiments)

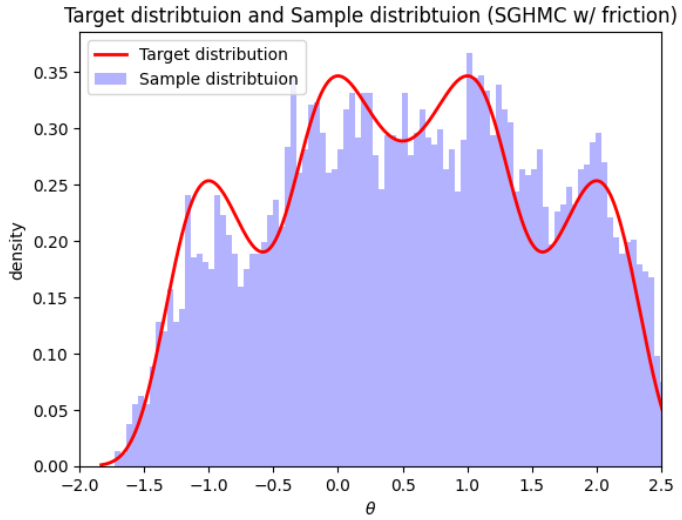




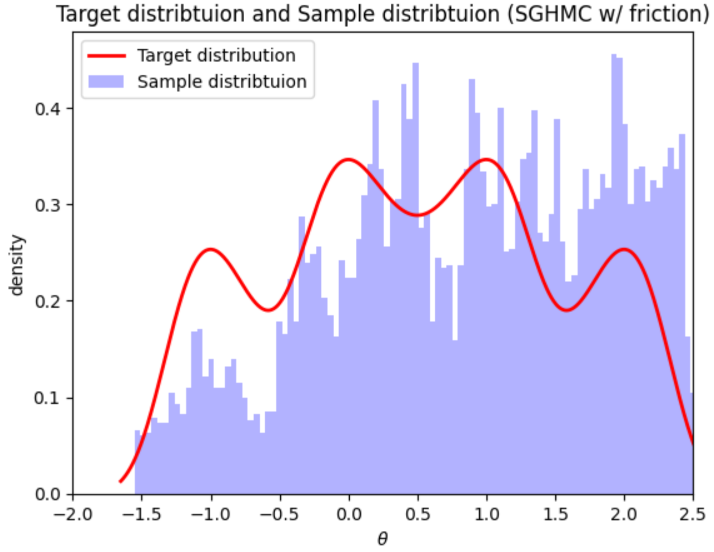
$\boldsymbol{\alpha = 0.5, \gamma = 1}$

Phenomenon analysis (Experiments)





When $\alpha = 0.1$

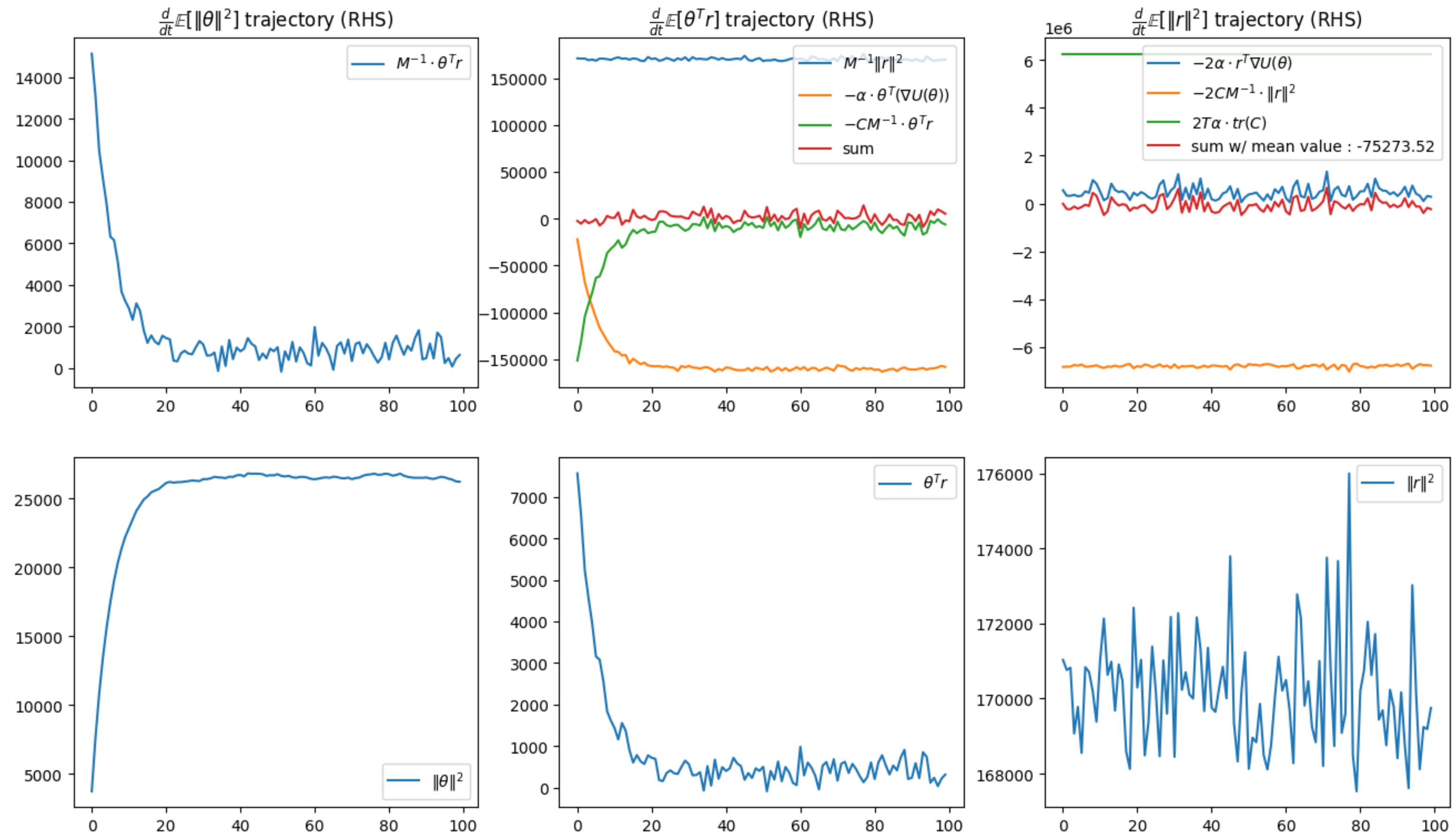


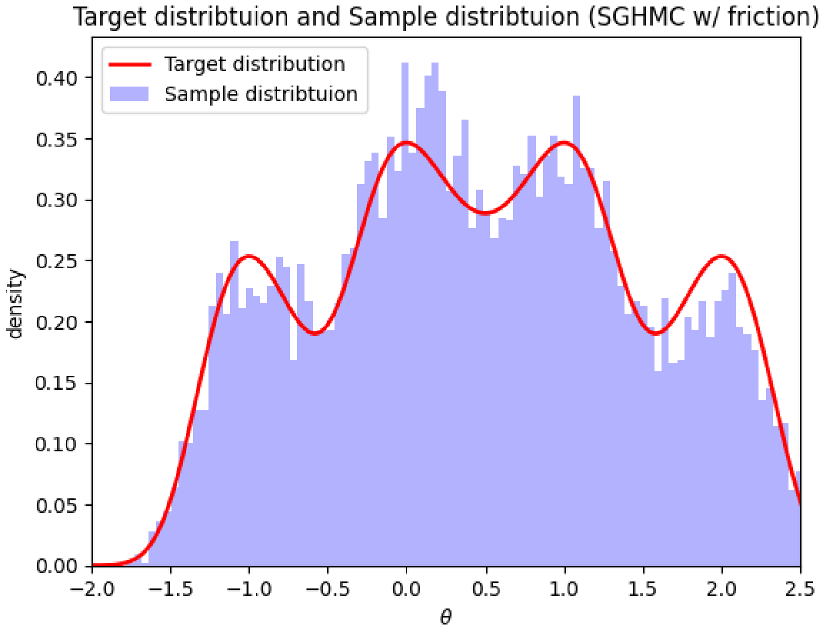
$\alpha = 0.1, \gamma = 1$

Lower $\alpha \rightarrow$ small focus on ∇U
$\therefore$ slower mixing (bad effect)

When $\alpha = 0.001$ (poor mixing)

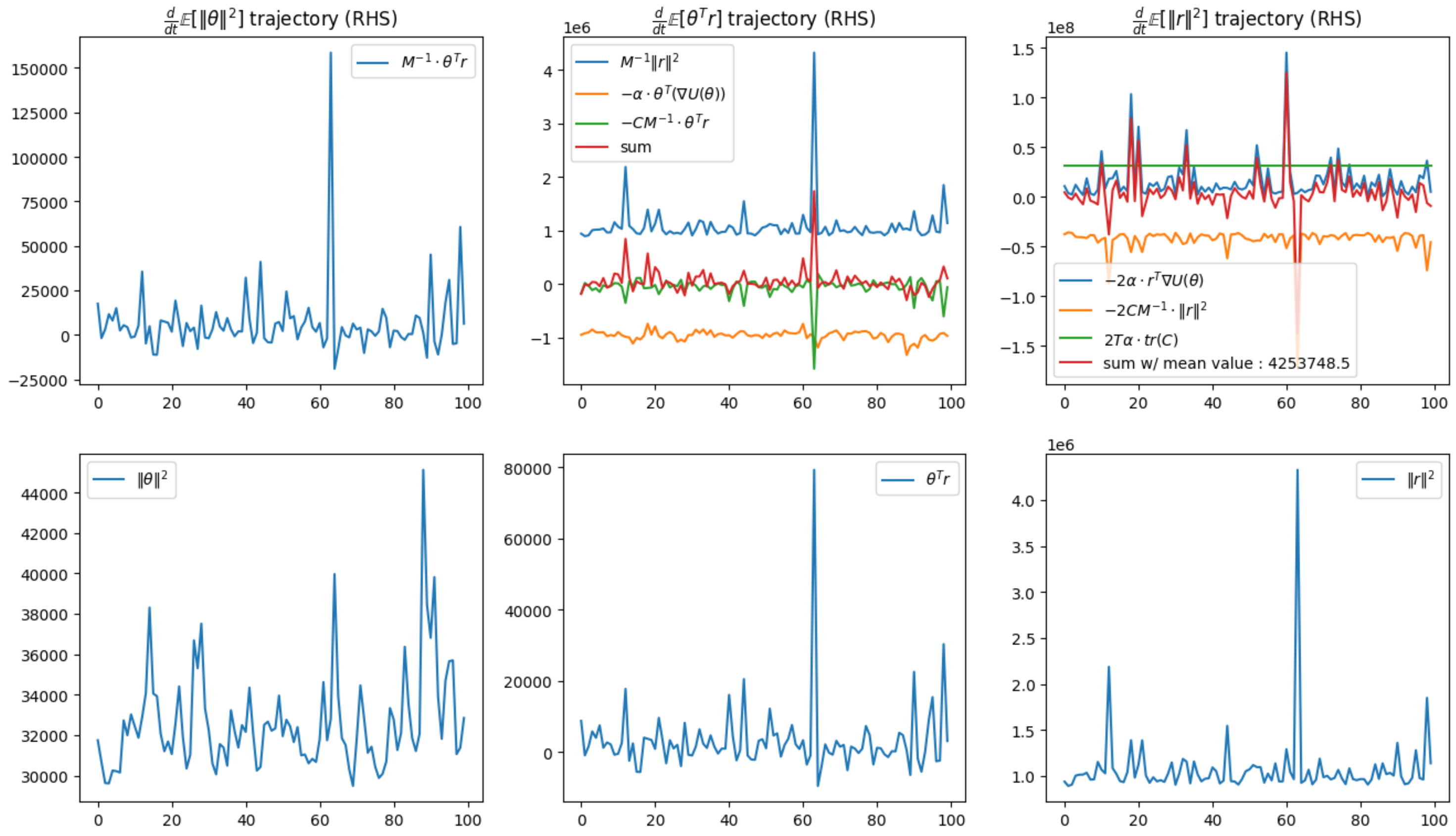
Phenomenon analysis (Experiments)

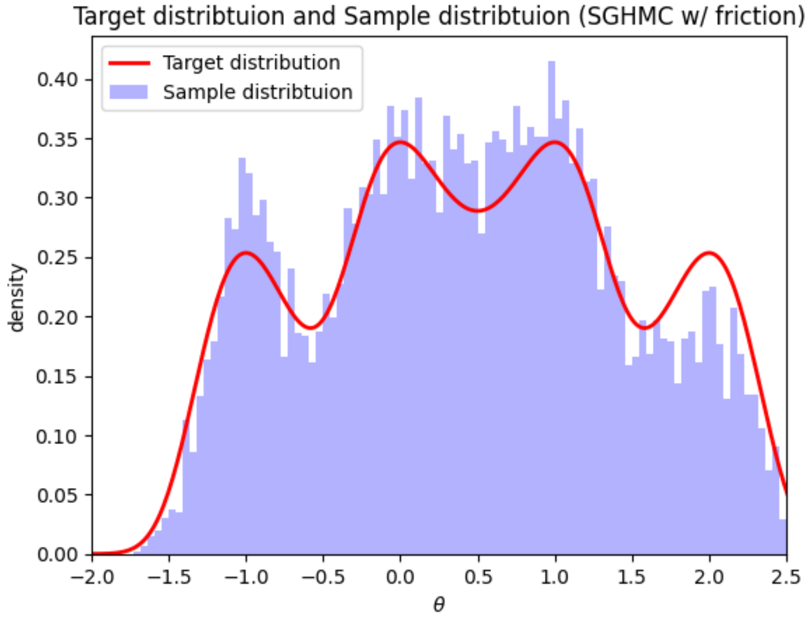




$\boxed{\alpha = 2, \gamma = 1}$

Phenomenon analysis (Experiments)



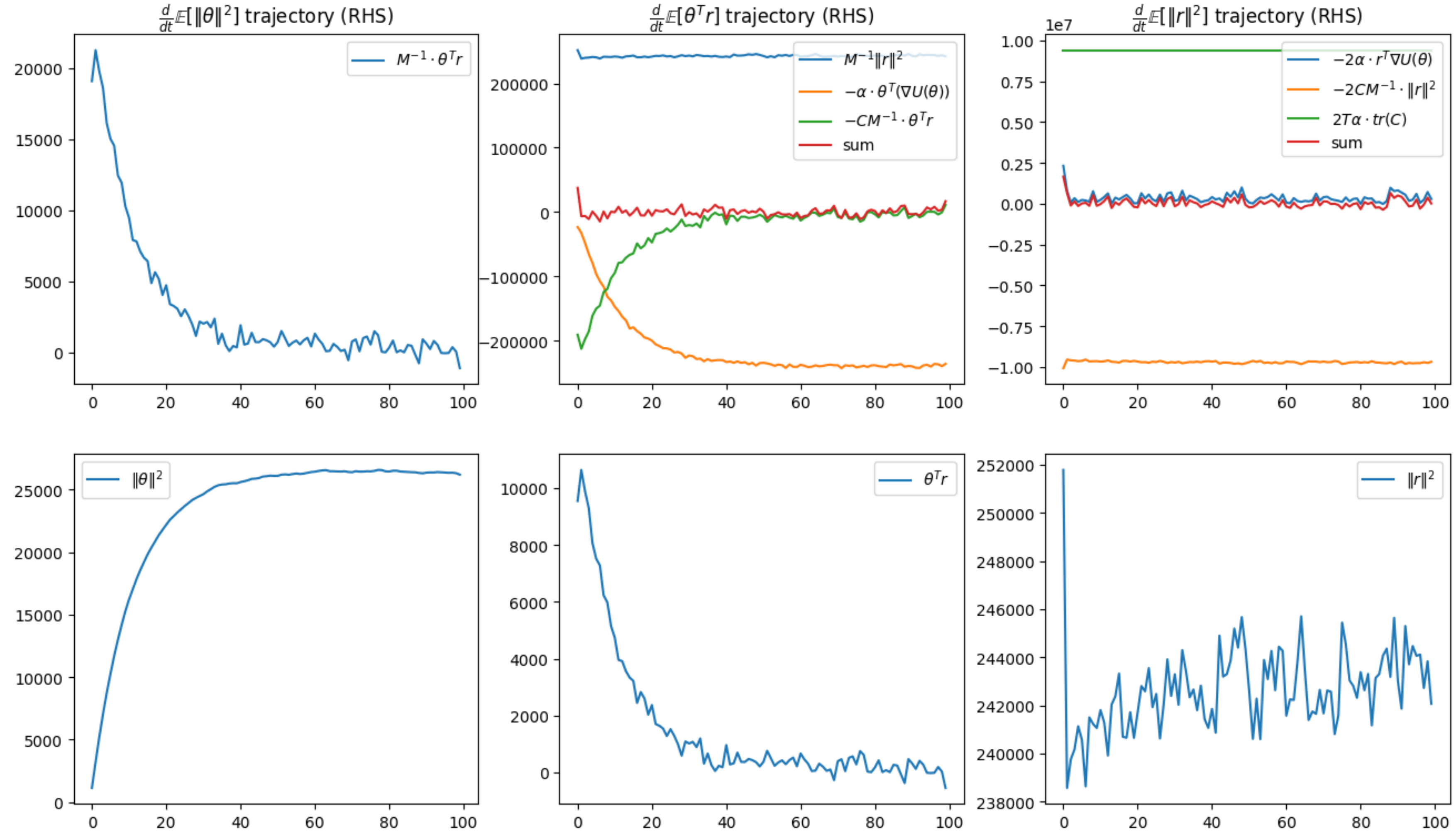


$\alpha = 5, \gamma = 1$

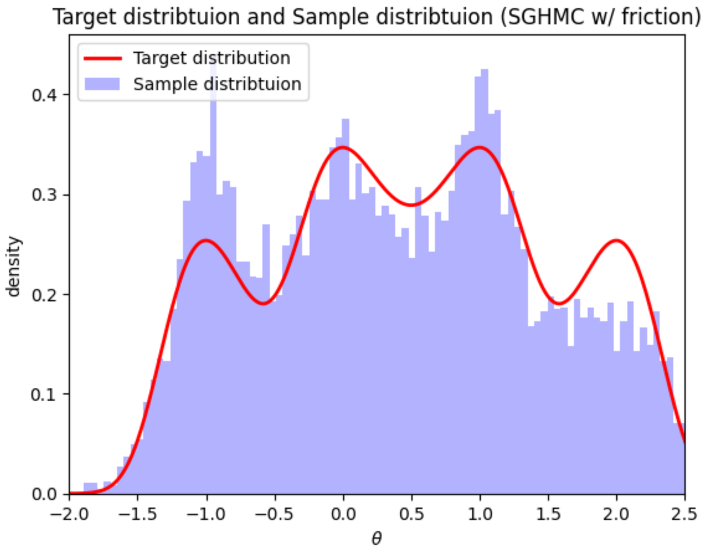
Higher α → high focus on ∇U
∴ mixing focused on local modes

※ Problem :
it seems that too high α leads to a sampling with high weight norm.

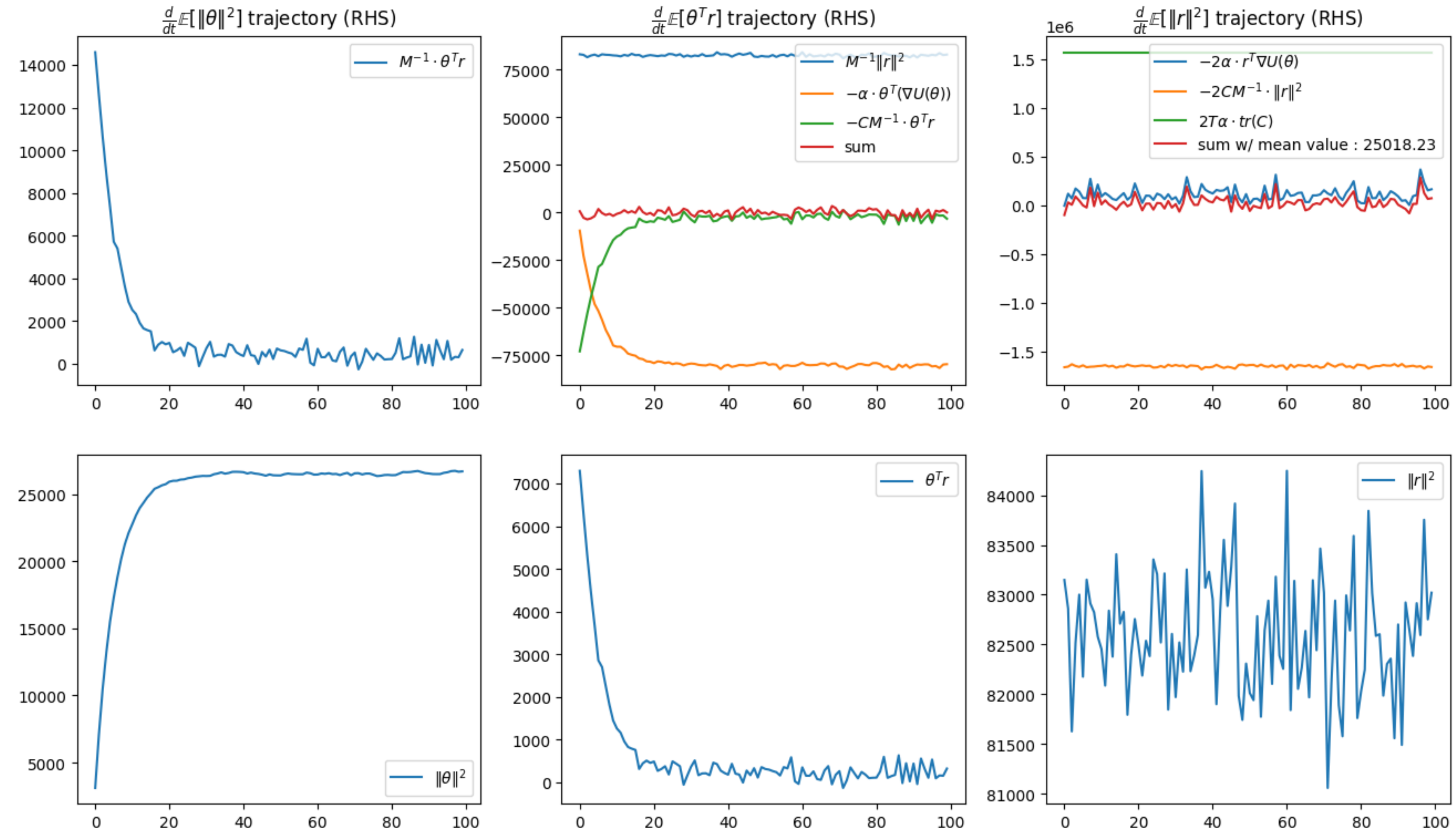
Phenomenon analysis (Experiments)

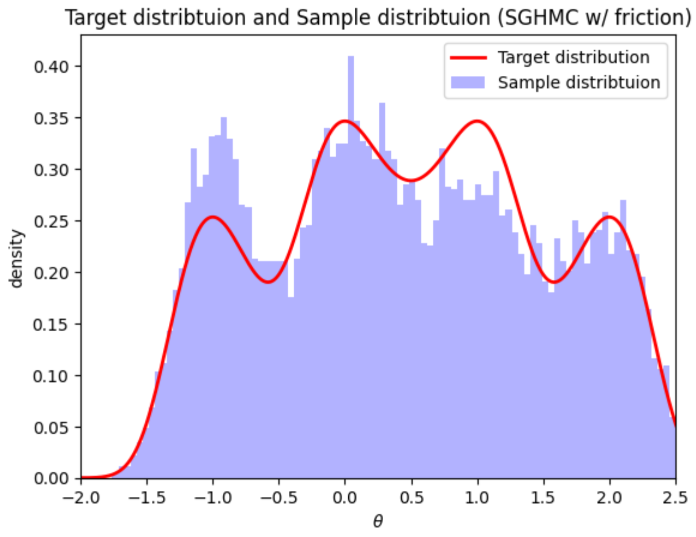


$\boldsymbol{\alpha = 1, \gamma = 1}$



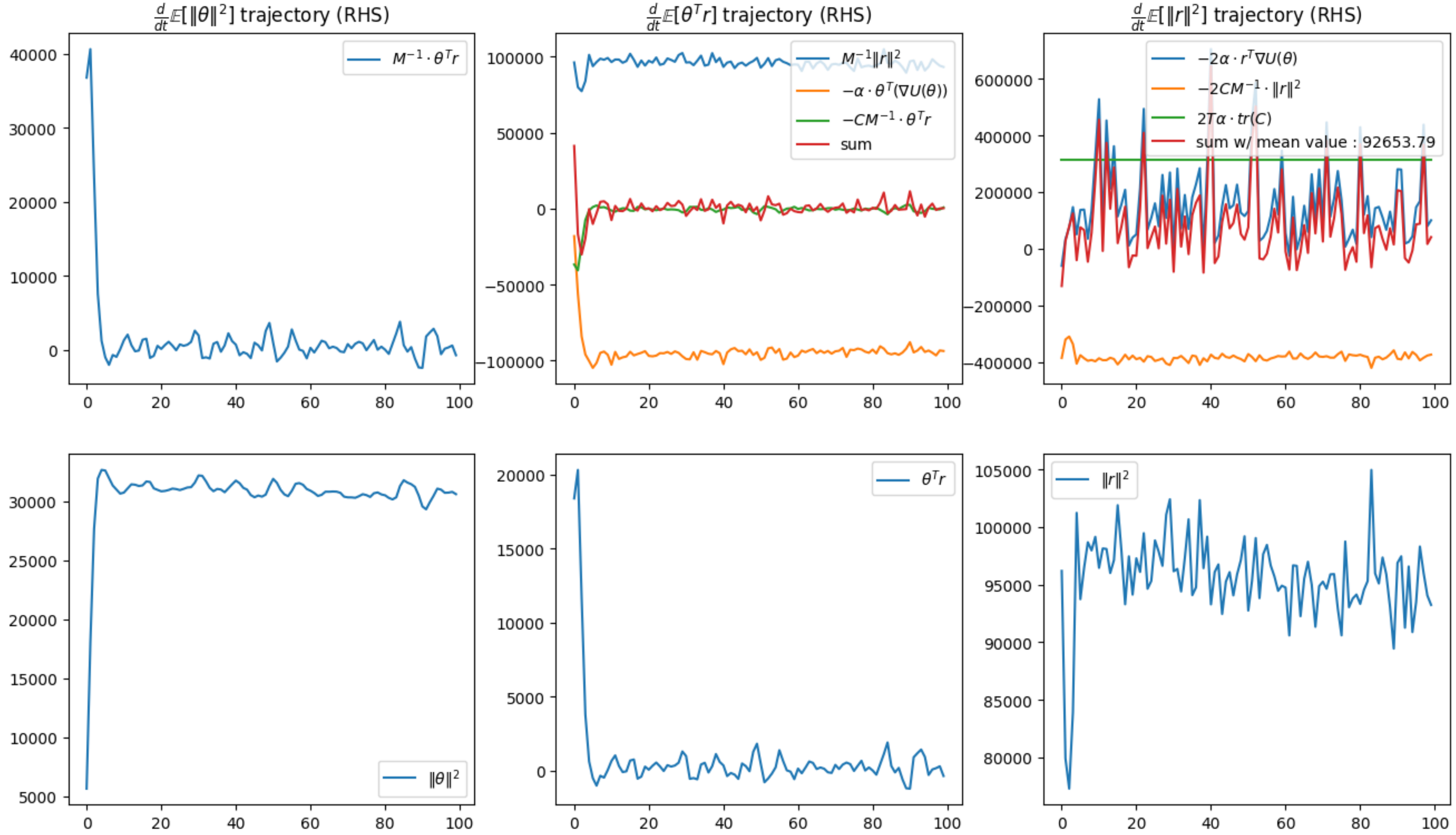
Phenomenon analysis (Experiments)

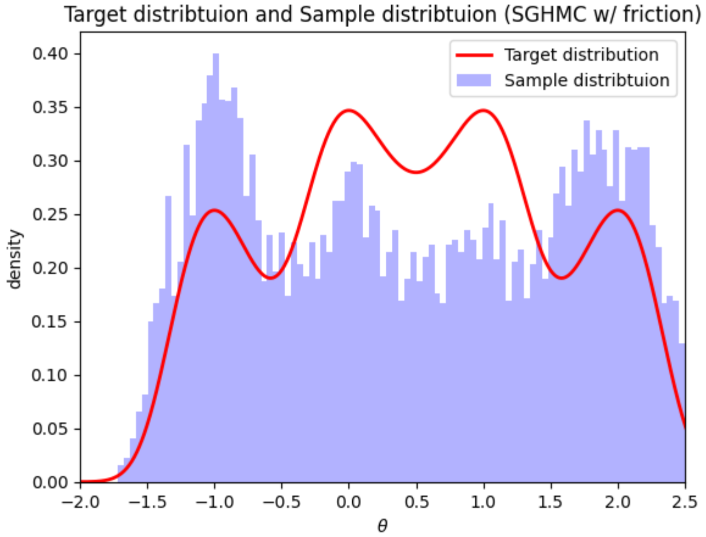




$\boldsymbol{\alpha = 1, \gamma = 0.5}$

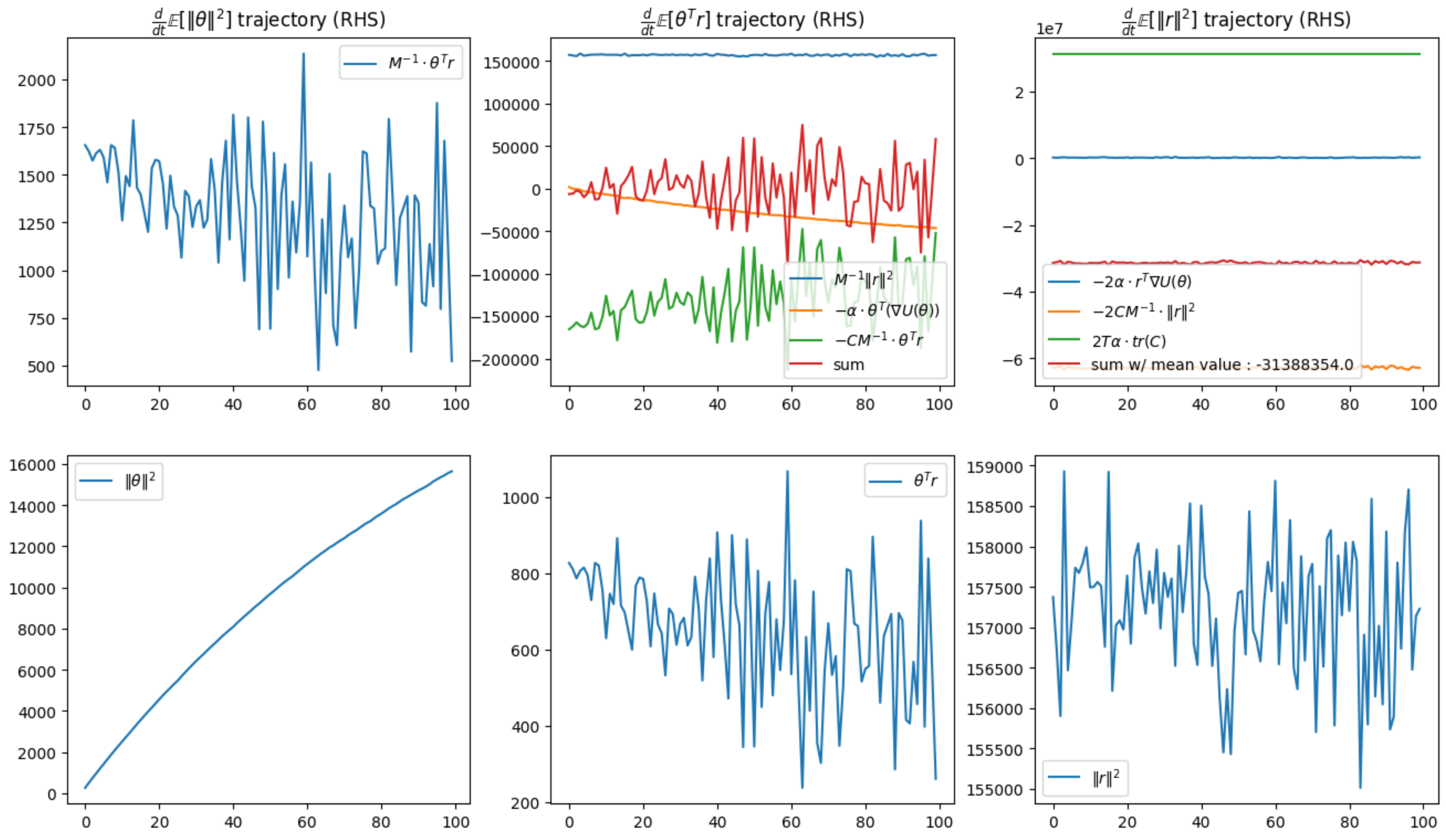
Phenomenon analysis (Experiments)

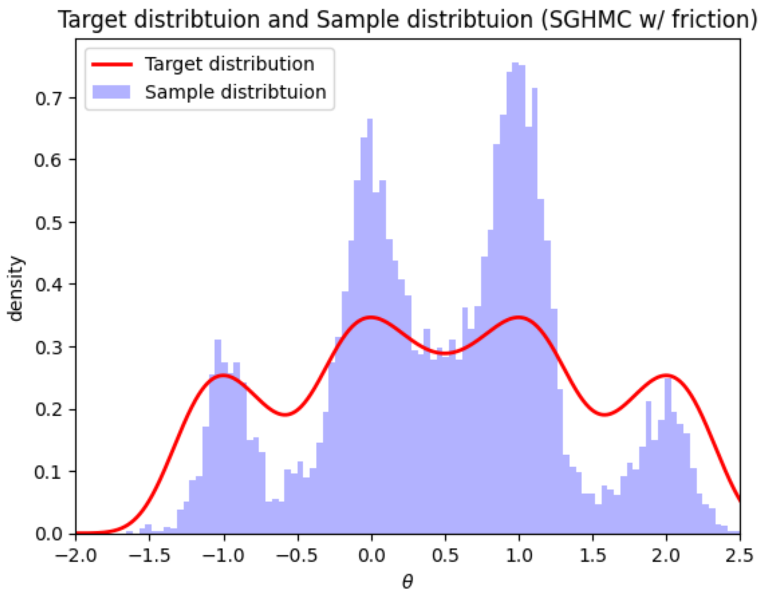




$\boldsymbol{\alpha = 1, \gamma = 0.1}$

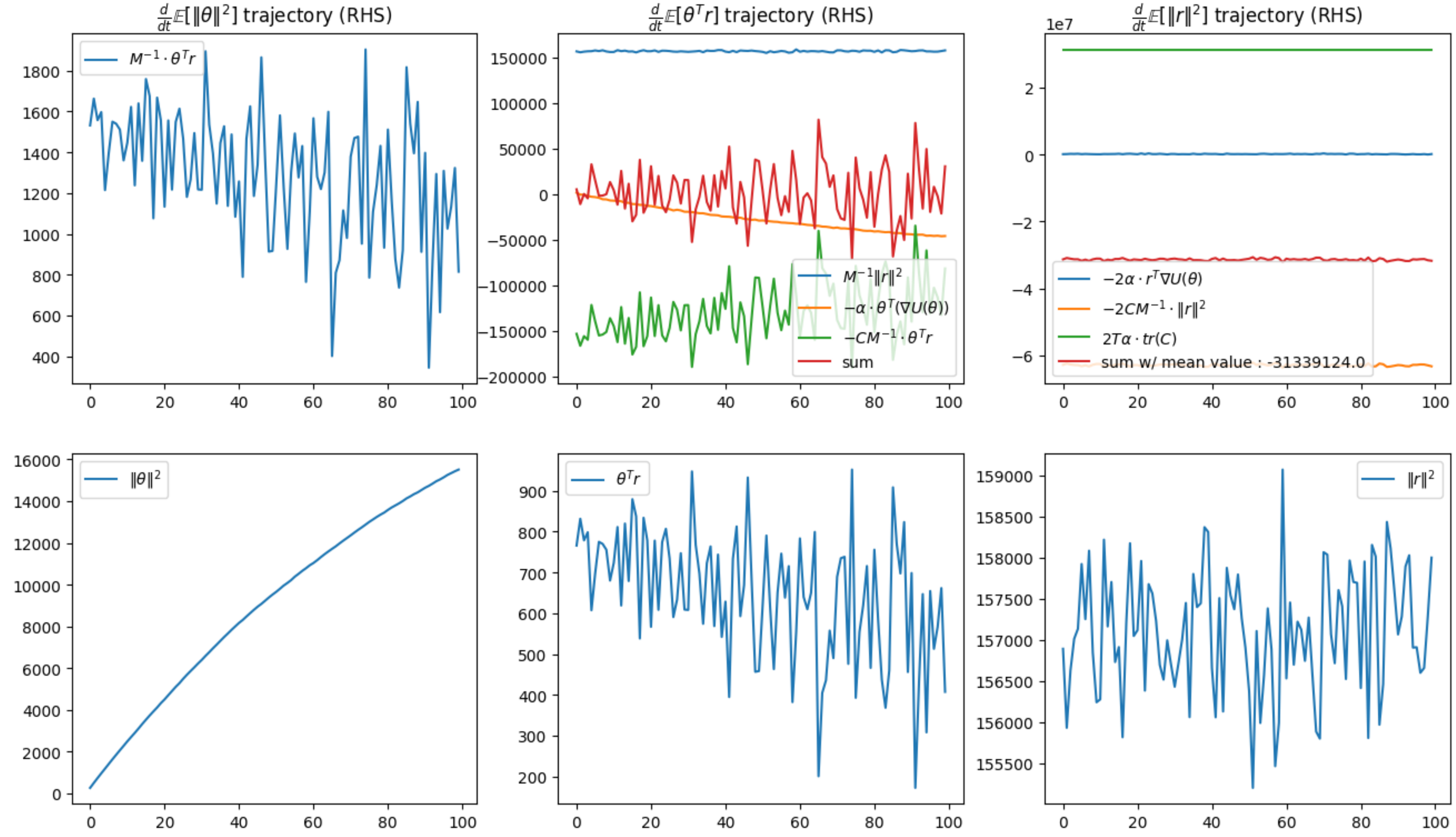
Phenomenon analysis (Experiments) – w/ momentum resampling

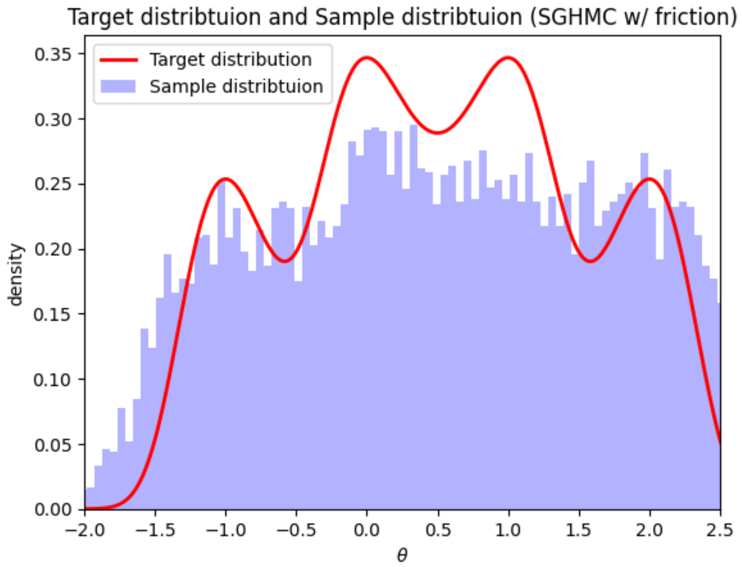




$\boldsymbol{\alpha = 1, \gamma = 1\ ,\ \beta\ =\ 0.001}$

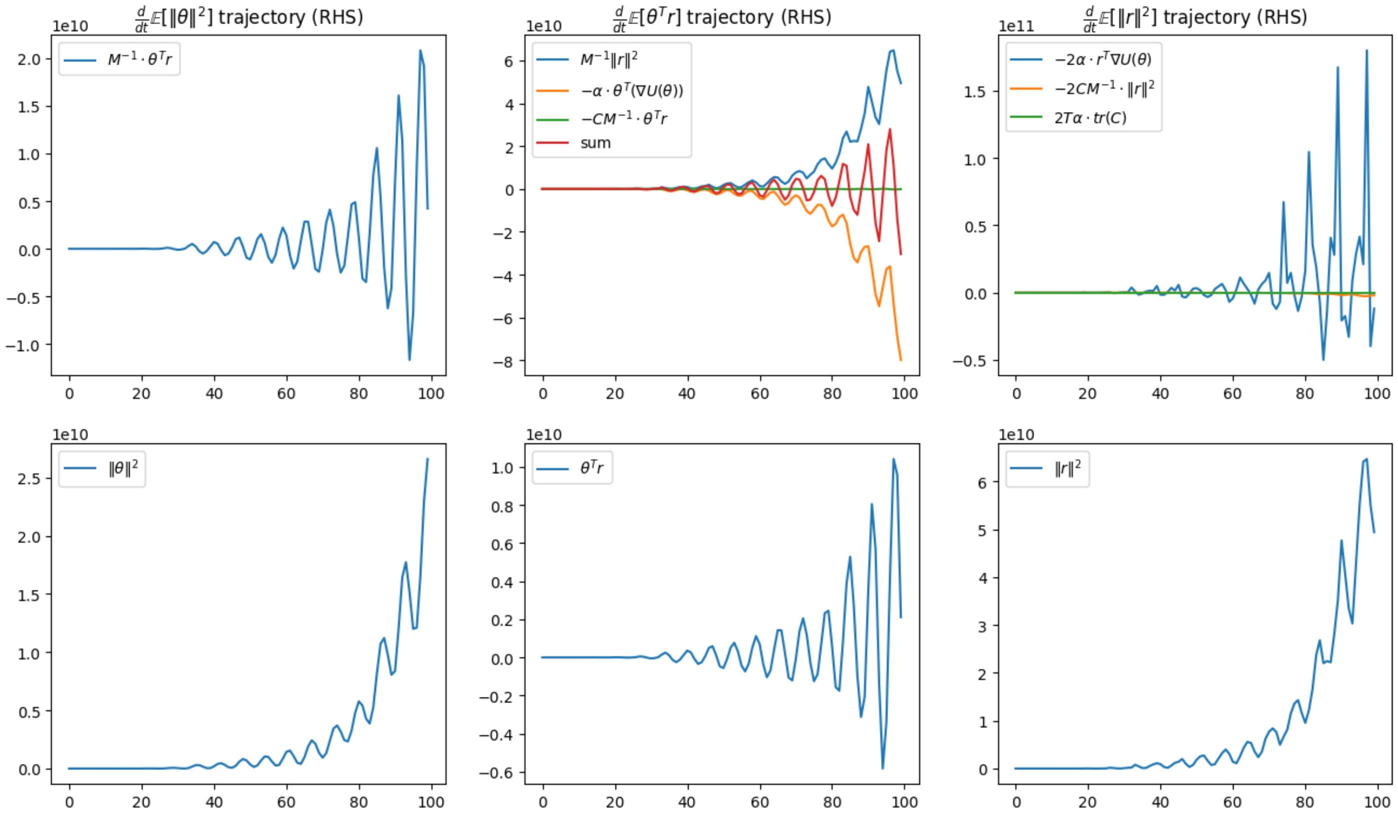
Phenomenon analysis (Experiments) – w/ momentum resampling

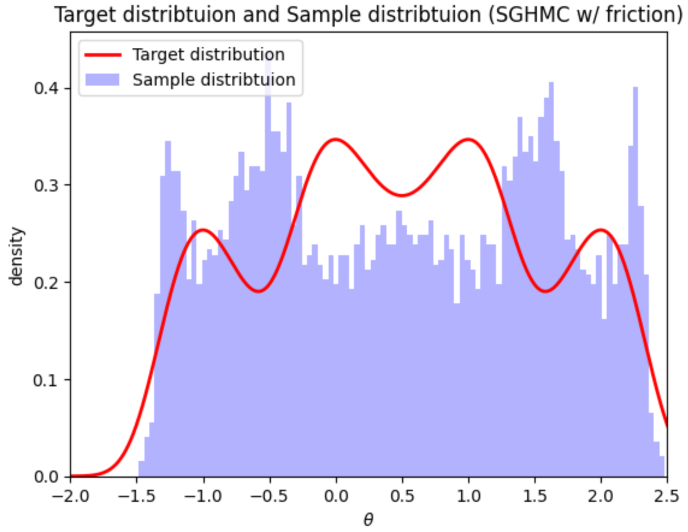




$\alpha = 1, \gamma = 1\ , \beta = 2$

Phenomenon analysis (Experiments)





$\boldsymbol{\alpha = 1, \gamma = 0.001}$

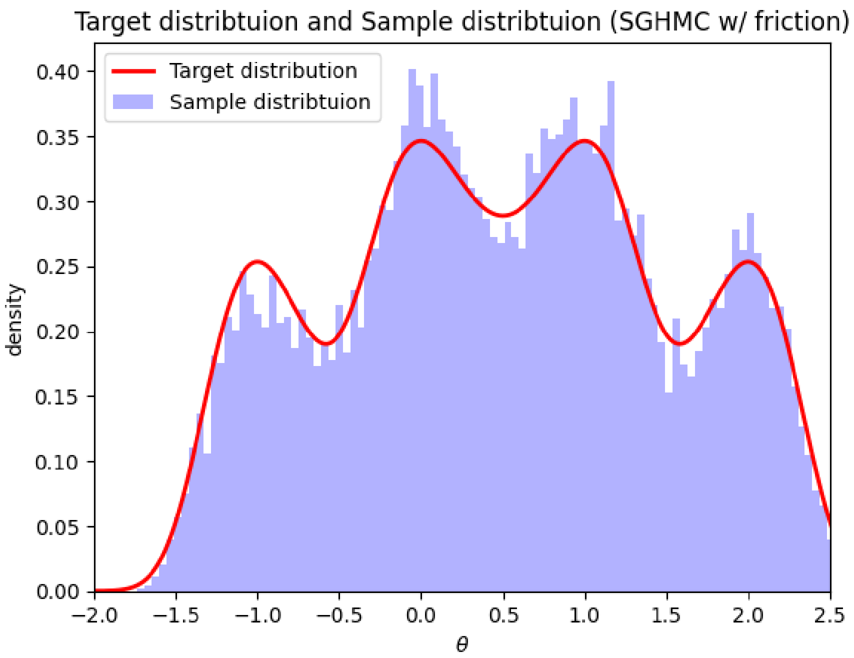
Note:
The oscillating effect originates from small friction coefficient.
$\rightarrow$ helps to appear $\|\theta\|^2$ decreasing zone if combined with M. resampling

Phenomenon analysis (Experiments)

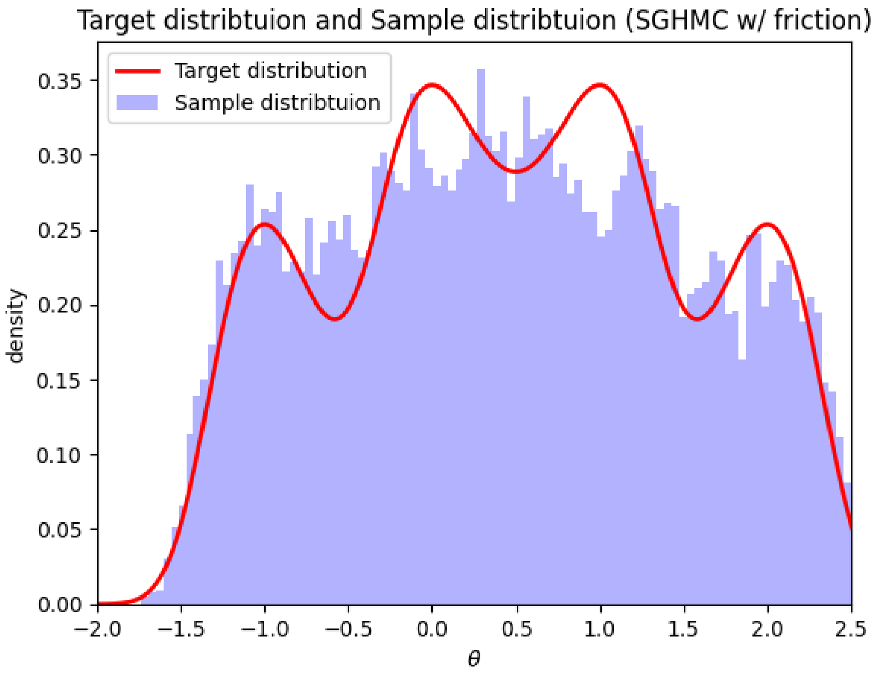
- Can we mimic the cold posterior by using these parameters??

<Effect of parameters>

1. High α : **mixing focused on local modes**



$\boldsymbol{\alpha = 1, \gamma = 1}$



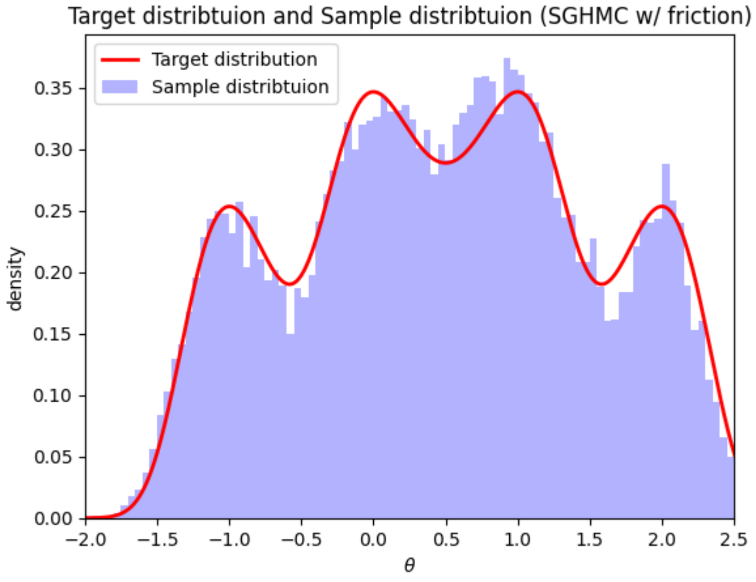
$\boldsymbol{\alpha = 5, \gamma = 1}$

Phenomenon analysis (Experiments)

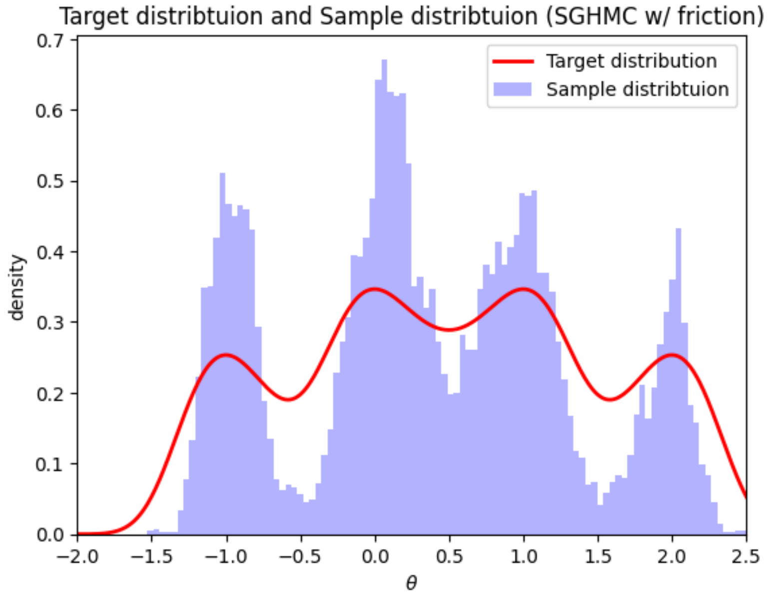
- By exploiting the momentum resampling as a tool to escape local modes for high α …

<Effect of parameters>

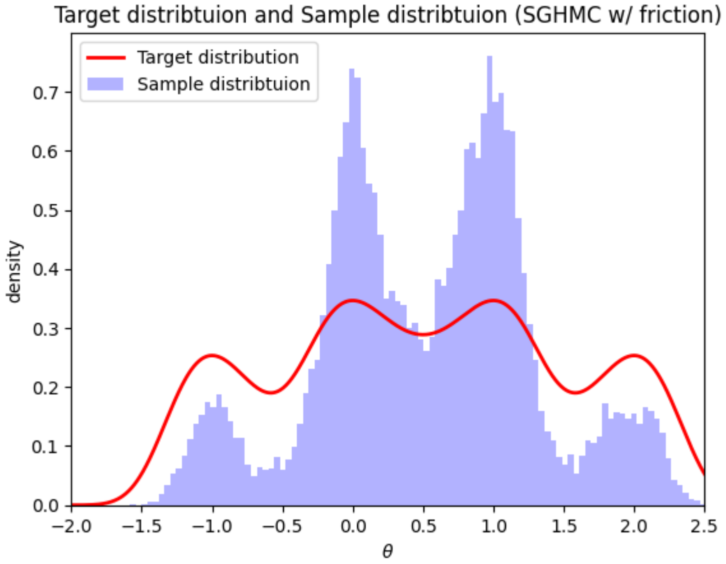
1. High α : **mixing focused on local modes** → **effectively explore modes when momentum resampling is adopted.**



$\boldsymbol{\alpha = 1, \gamma = 1, \beta = 1}$



$\boldsymbol{\alpha = 5, \gamma = 1, \beta = 1}$



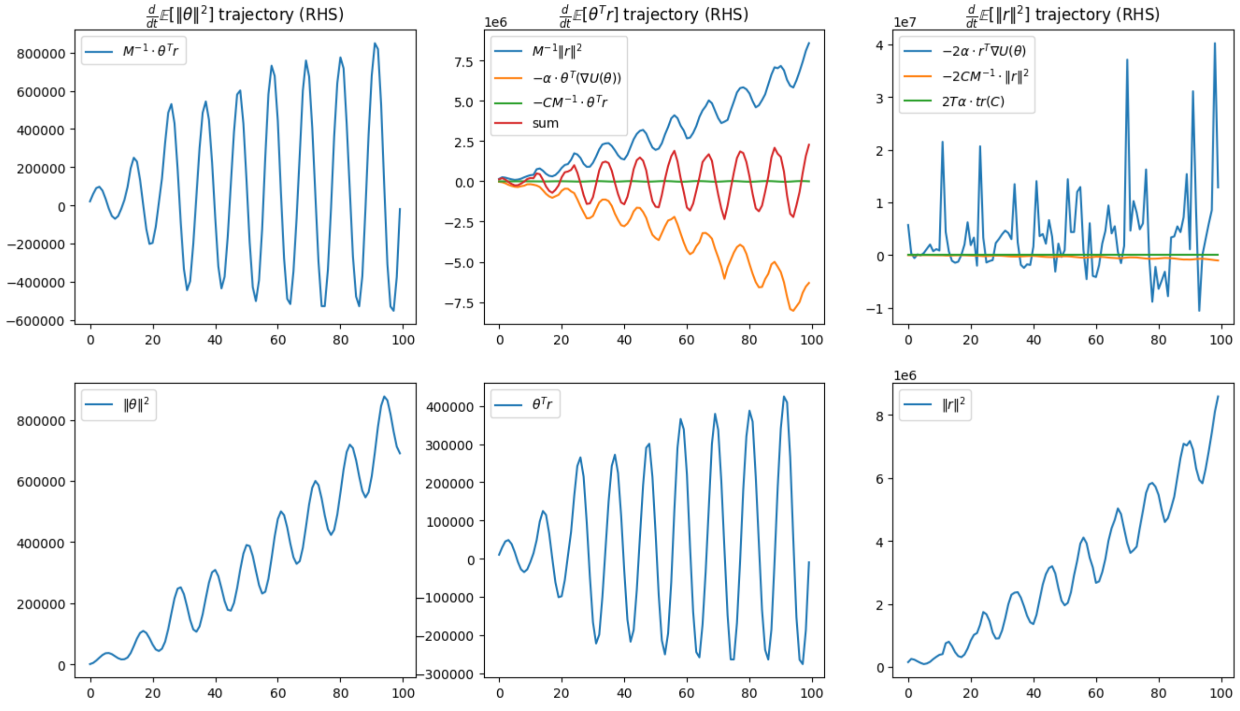
$\boldsymbol{\alpha = 1, \gamma = 1, T = 0.5}$
(cold posterior)

Phenomenon analysis (Experiments)

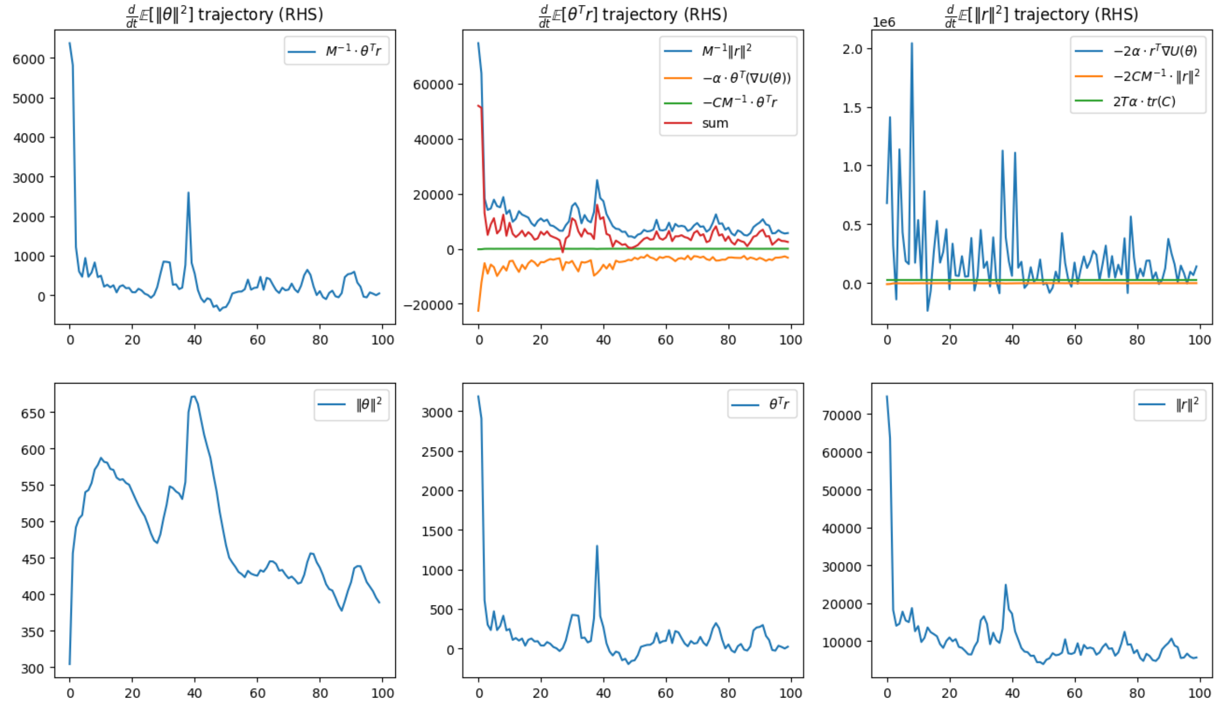
- By exploiting the momentum resampling as a tool to escape local modes for high α …

<Effect of parameters>

2. Low γ (friction coeff.): **helps to make oscillation behavior of $\|\theta\|_2$ & decreasing zone**



Norm-adjusting SGHMC (w/o momentum resampling)



Norm-adjusting SGHMC (w/ momentum resampling)

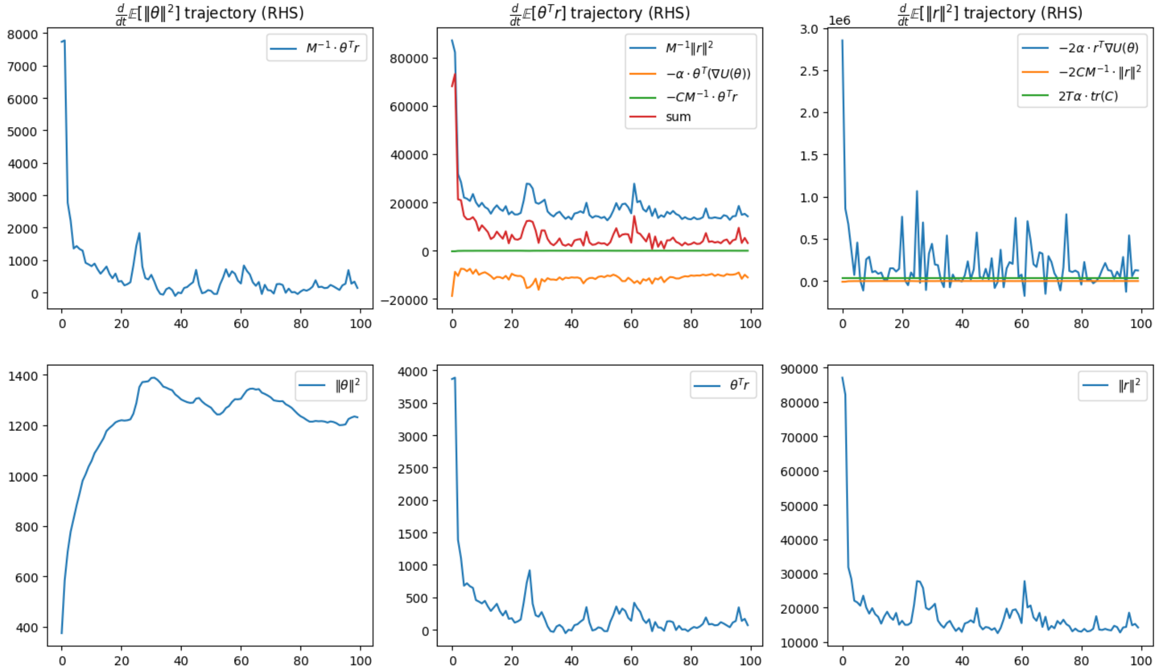
Phenomenon analysis (Experiments)

- By exploiting the momentum resampling as a tool to escape local modes for high α …

<Effect of parameters>

3. Low β (momentum resampling scaler) : **regulate the weight norm $\|\theta\|_2$**

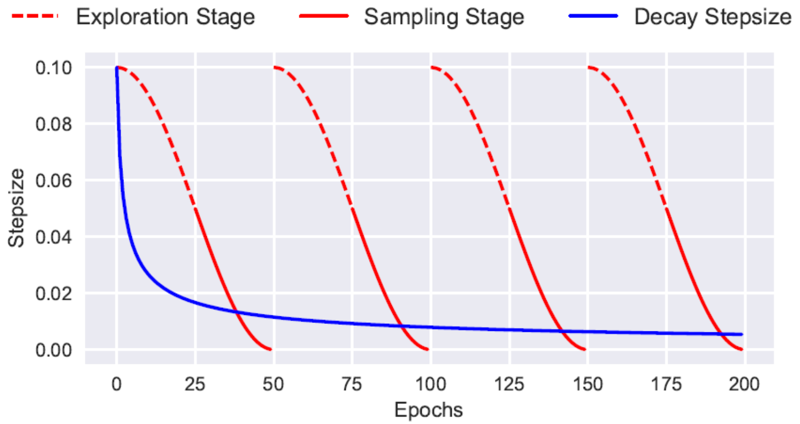
Norm-adjusting SGHMC (w/ $\beta = 0.001$)

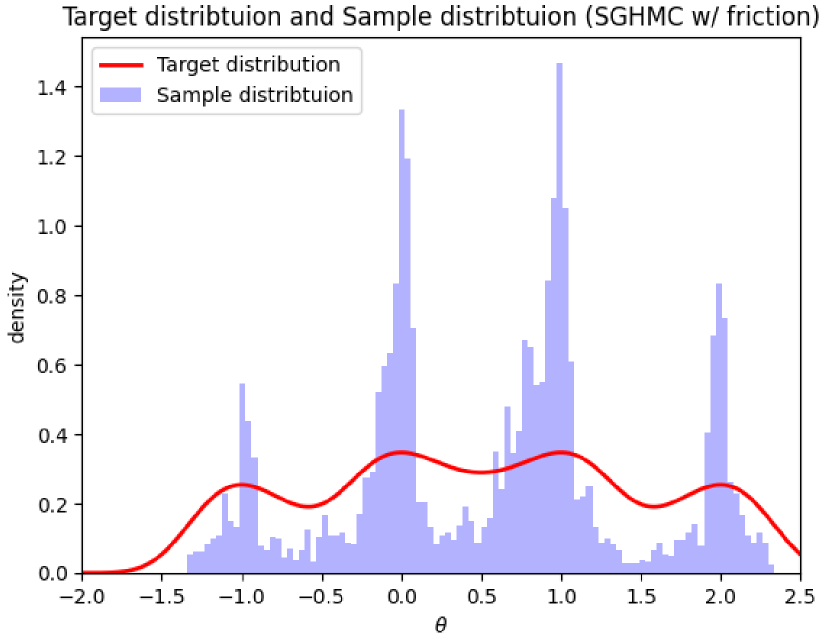


Norm-adjusting SGHMC (w/ $\beta = 0.1$)

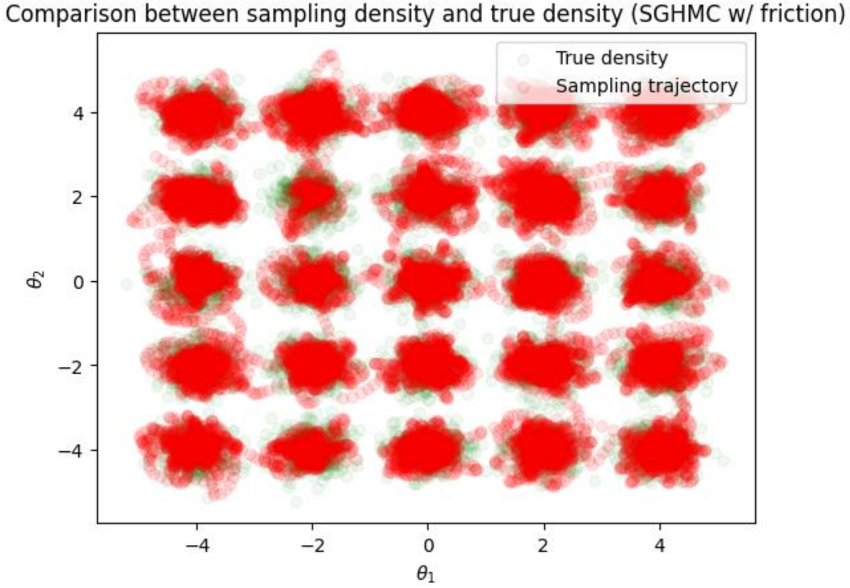
One interesting toy experiment (Appendix)

- What happen if we schedule the momentum scaler β as we did in CSG-MCMC??
  - Increased β (exploration) & decreased β (sampling)
  - Then, we can explore local modes very effectively without aid of cyclic step size scheduler.

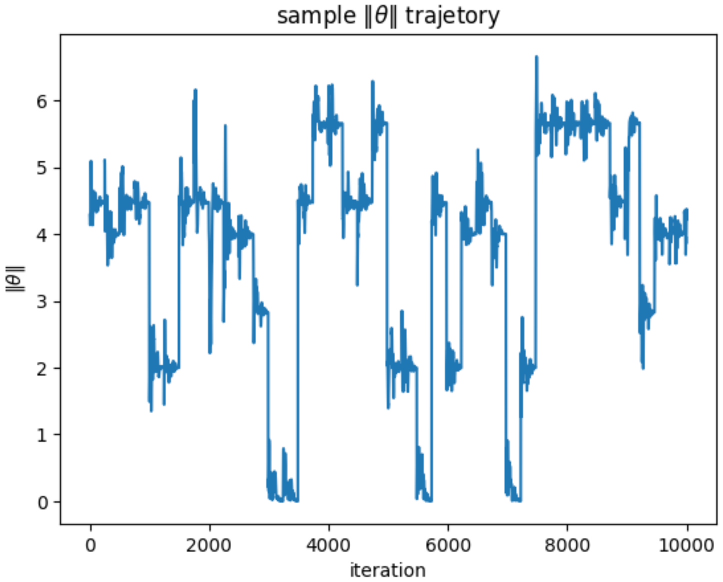




New method + momentum cyclic scheduler (1D)



New method + momentum cyclic scheduler (2D)



Sample norm trajectory (2D)

Summary of heuristics

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

- For parameters, we take $\alpha > 1, \ \ \gamma, \beta \ll 1$.
- By Fokker-Planck equation:

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

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

- Note that this method is just nothing but original SG-HMC with different parameters M, C.
  - It reveals that importance of mass M and friction coefficient C to regulate $\|\theta\|^2$ when it combined with momentum resampling.

    (This could be the reason why some paper claims “SGMCMC is good enough w/o cold posterior”)