L2 norm burst during BNN training (3)

-Summary-

23/09/19

Summary of heuristics (Review)

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

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

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

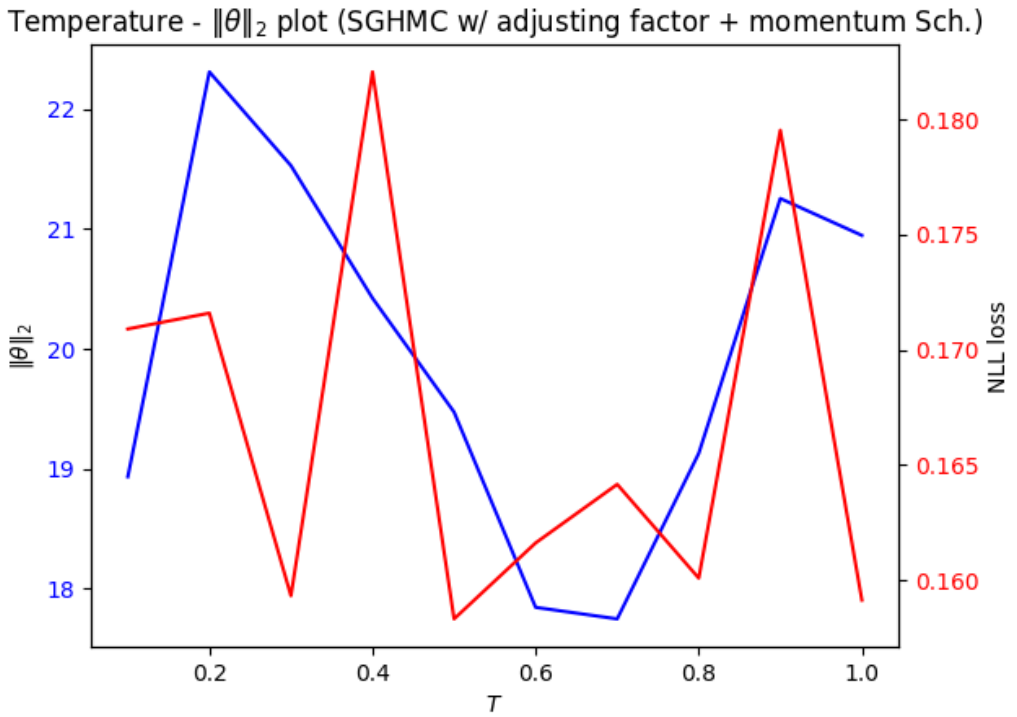
$$\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + \alpha^{-1}\gamma C M^{-1} r)] + 2T\gamma \cdot tr(C)$$ (= $2\gamma \cdot tr(C)$ if w/o cold posterior)

- 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. **-> Is it really true? (Our current question)**

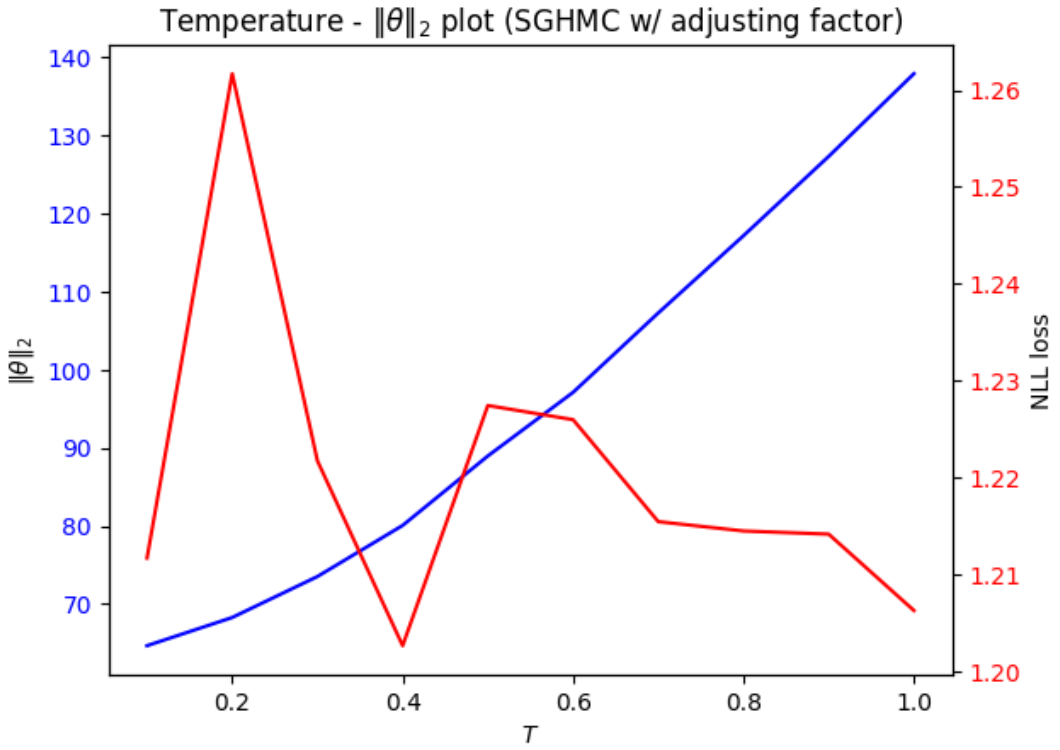
    (This could be the reason why some paper claims “SGMCMC is good enough w/o cold posterior”)

Summary of heuristics (Review)

- Results of revised method



MNIST

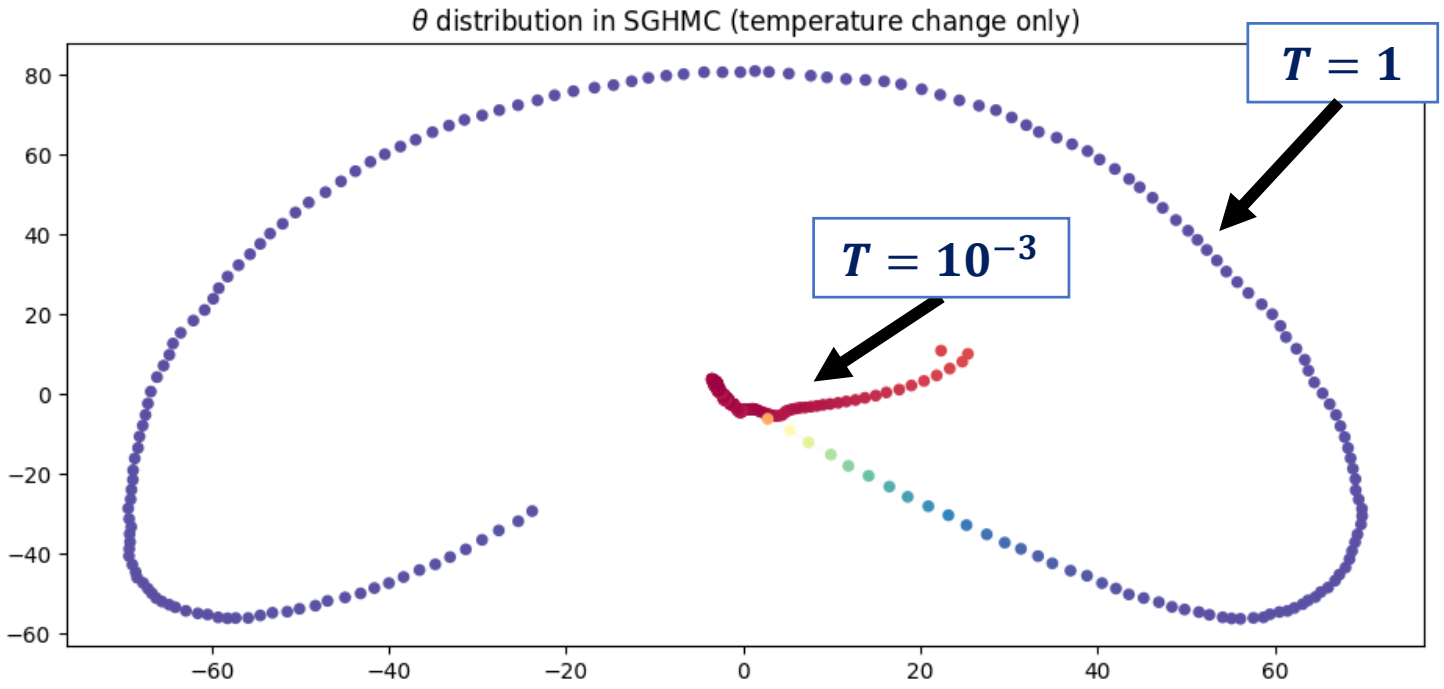


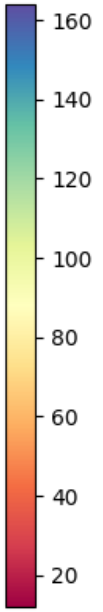
CIFAR-10 (w/o data augmentation)

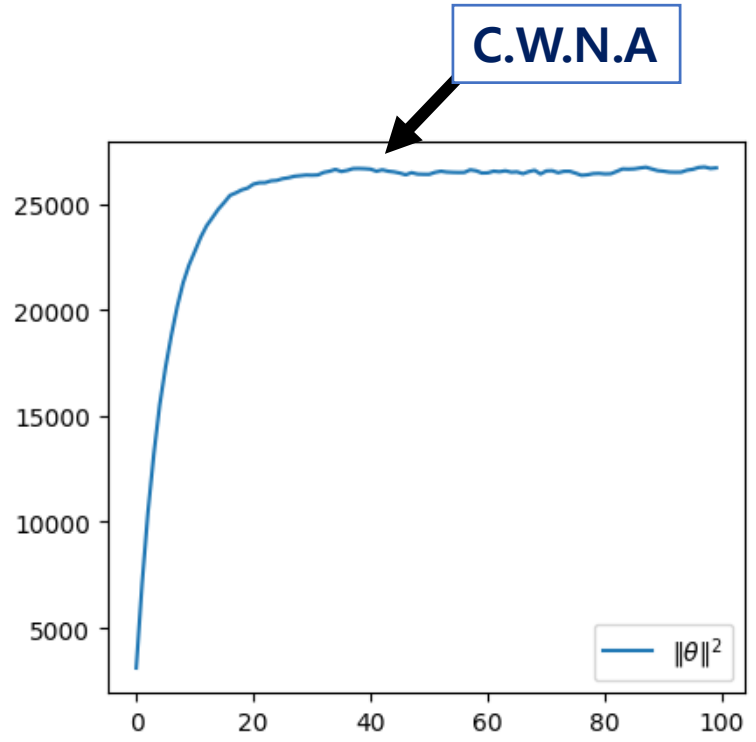
- **Critical problem : for optimal choice of C, M, NLL loss : 0.1326 w/ $T = 0.0001$ (MNIST)**
- **Can we perform better than cold posterior by using $T = 1$??**

Observed problem

- Suspicious behavior observed in large dataset: (we use SGHMC only, here)
  - The weight norm keeps (almost) constant on the end-tail of the training (also high value).





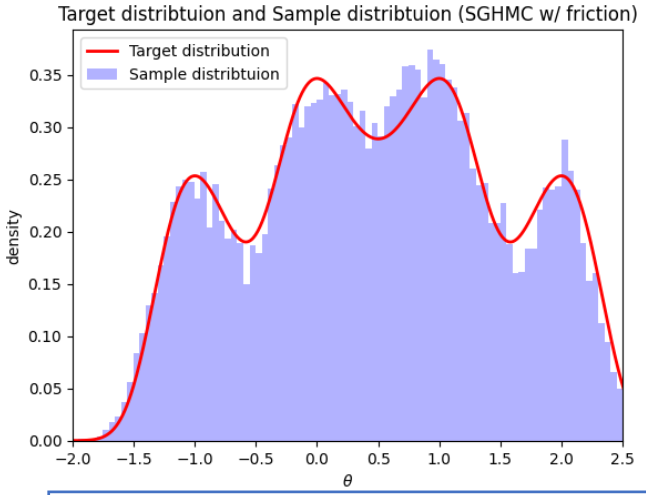


Projected θ by PCA in MNIST (color bar : $\|\theta\|_2$)

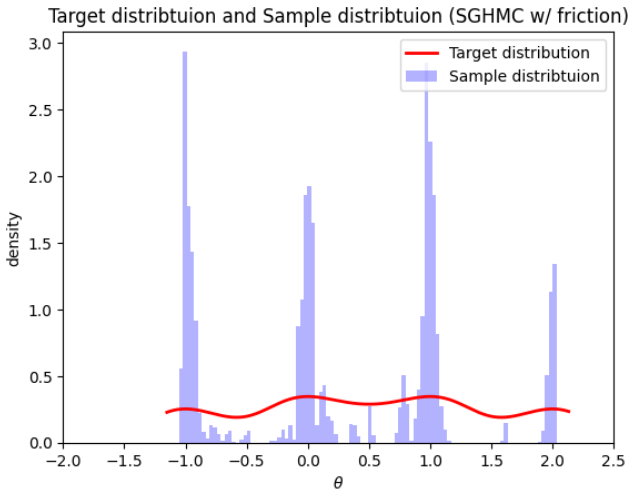
Constant weight norm area (C.W.N.A)

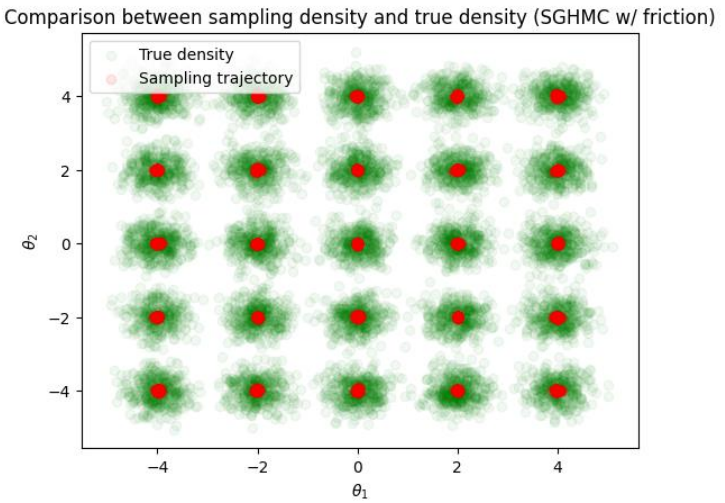
Observed problem

- Question : why does the suspicious behavior is observed ?
  - One hypothesis is that very large numbers of local modes hamper the efficient sampling, or there is a intrinsic problem of SGHMC that make C.W.N.A appears.
  - In practice, we do not require the samples nearby local modes (due to the limited # of samples), but the samples which locates on the exact local modes.



Not favorable for deep BNN





Favorable for deep BNN

Observed problem

- Question : why does the suspicious behavior is observed ?
  - Furthermore, those samples does not perform well (may be) due to the high weight norm. (acc avg. ≅ 90.0, while 93~94 for cold posterior individual samples)
  - We previously suggested a heuristic to adopt momentum resampling to reduce weight norm. To understand the problem clearly, let's consider basic toy example:

$$p(\theta|\mathcal{D}) = N(0, I_d), \quad (d = 5)$$

$$\begin{bmatrix} d\theta \\ dr \end{bmatrix} = \underbrace{\begin{bmatrix} M^{-1}r \\ -\nabla U(\theta) - CM^{-1}r \end{bmatrix} dt}_{\text{Drift term}} + \underbrace{\begin{bmatrix} 0 \\ N(0, 2CTdt) \end{bmatrix}}_{\text{Diffusion term}}$$

  - In this setting, we set $T = 0$ to observe behavior determined by "drift term"

Observed problem

- To understand the problem, let's consider basic toy example: $p(\theta|\mathcal{D}) = N(0, I_d)$, $(d = 5)$
  - Then, $U(\theta) = \nabla(-\log p(\theta|\mathcal{D})) = \theta$ (when $\theta|\mathcal{D} \sim N(0, I_d)$), and it follows that

$$\dot{z} = \begin{bmatrix} 0 & M^{-1} \\ -I_d & CM^{-1} \end{bmatrix} z = Az$$

  where $z = [\theta, r]^T$

  - This is just a linear system, and the solution is given by below: (Now, assume $C, M \in \mathbb{R}$)

$$z = z(0)e^{At}$$

  - And, the characteristic polynomial of A is given by : $\left(\lambda^2 + \lambda CM^{-1} + M^{-1}\right)^{2d}$
    - The two roots of the equation : $\lambda_{sol} = \frac{-CM^{-1} \pm \sqrt{(CM^{-1})^2 - 4M^{-1}}}{2}$

Observed problem

- The two roots of the equation : $\lambda_{sol} = \frac{-CM^{-1} \pm \sqrt{(CM^{-1})^2 - 4M^{-1}}}{2}$

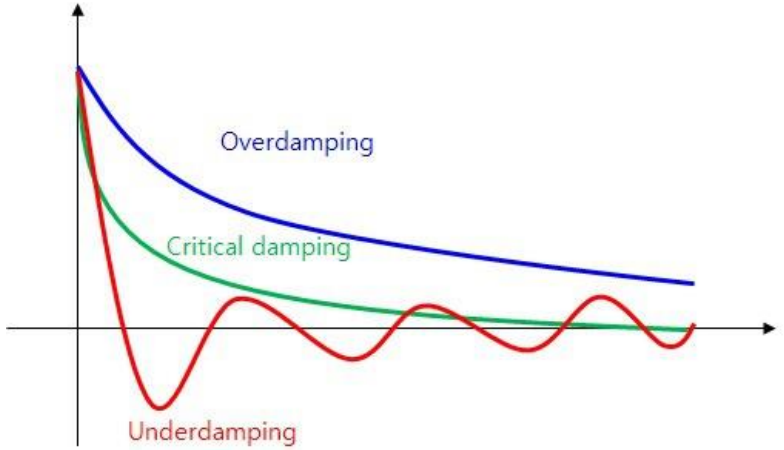
  - Note that

    $(CM^{-1})^2 - 4M^{-1} > 0 \rightarrow$ Overdamping

    $(CM^{-1})^2 - 4M^{-1} = 0 \rightarrow$ Critical damping

    $(CM^{-1})^2 - 4M^{-1} < 0 \rightarrow$ Underdamping

  - The condition for Critical damping is $C = 2\sqrt{M}$, also as the $CM^{-1} \uparrow$, we get faster decay.

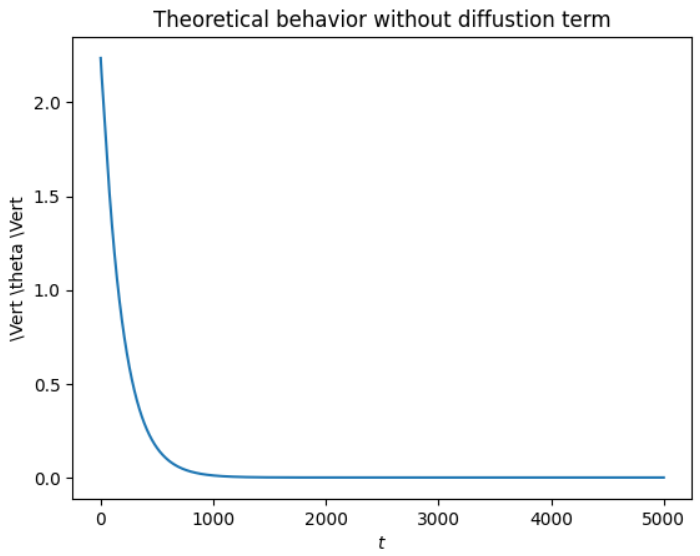


- Question : what happen if we adopt momentum resampling, now? (we do experiments, now.)

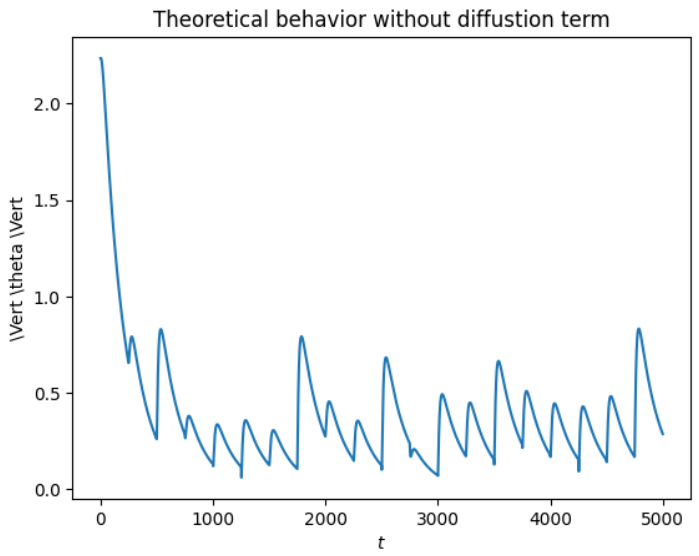
Observed problem

- Question : what happen if we adopt momentum resampling, now?
  - If we use momentum resampling, the exact analytic solution can be obtained by time-shifting with new initial condition $(\theta_{prev}, r_{sampled})$ [Here, $C = 4,\ M = 1 \Rightarrow C > 2\sqrt{M}$] (over-damping case)

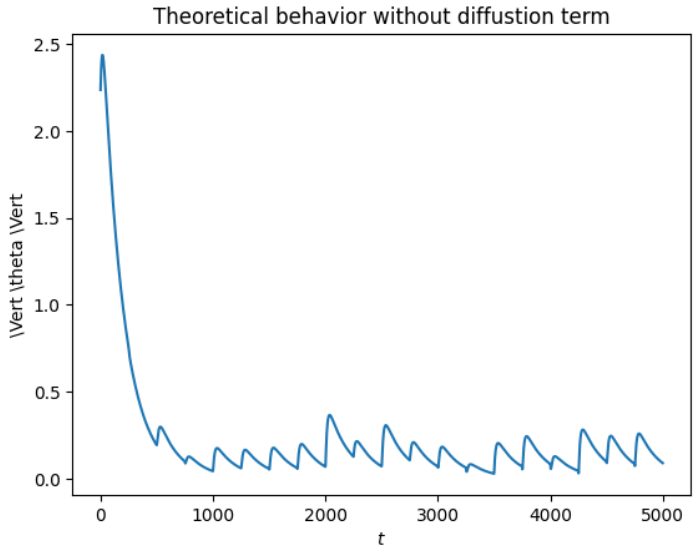
$$\boldsymbol{r \sim N(0, \beta M)}$$



Without momentum resampling



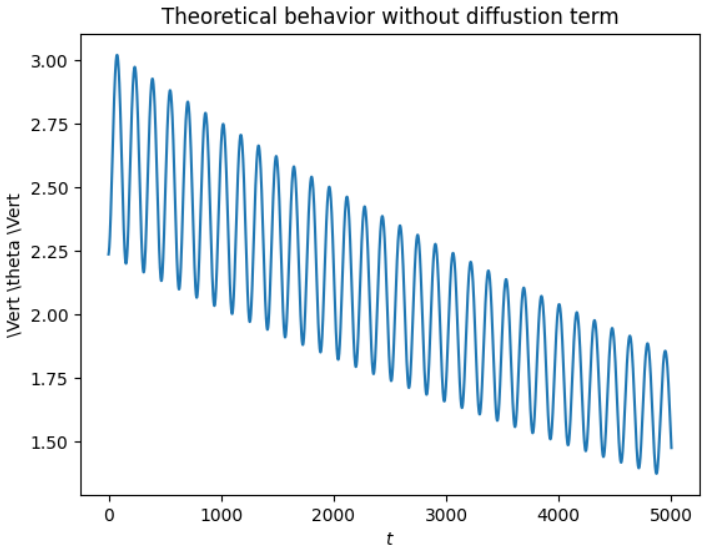
With momentum resampling



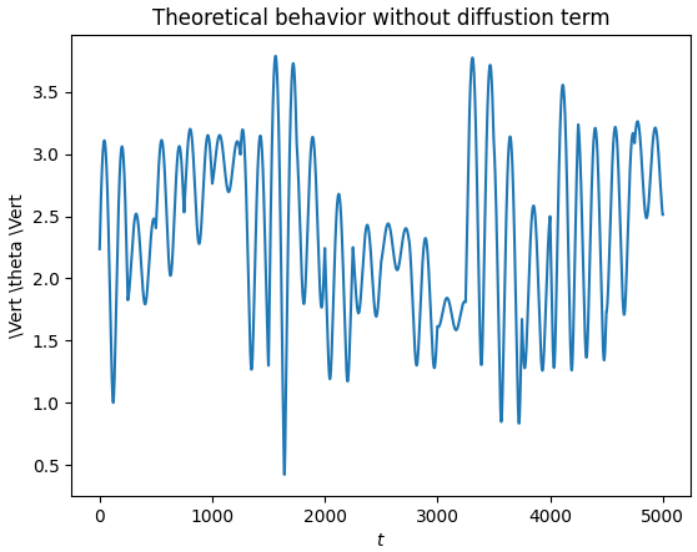
With momentum resampling ($\beta = 0.5$)

Observed problem

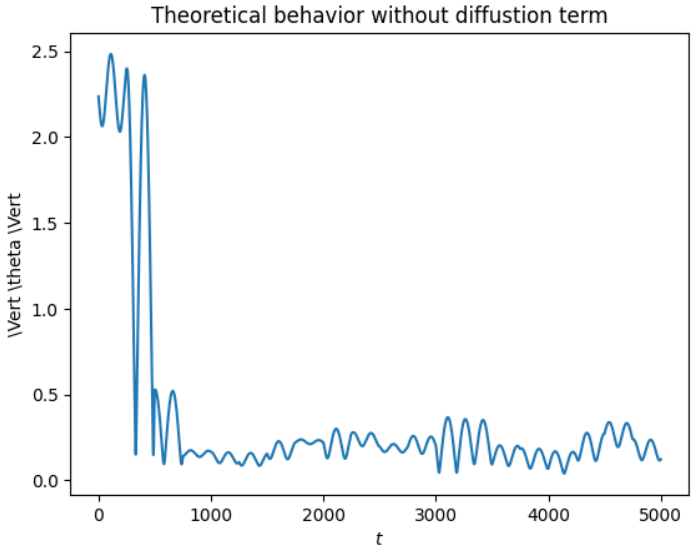
- In over-damping case, there is no interesting phenomenon even we add momentum resampling. But, what about under-damping case? [$C = 0.01, M = 1 \Rightarrow C < 2\sqrt{M}$]
- It significantly helps to approach toward the target solution $\theta^* = \theta_{MAP}$ $\quad\quad$ $\boldsymbol{r \sim N(0, \beta M)}$



Without momentum resampling



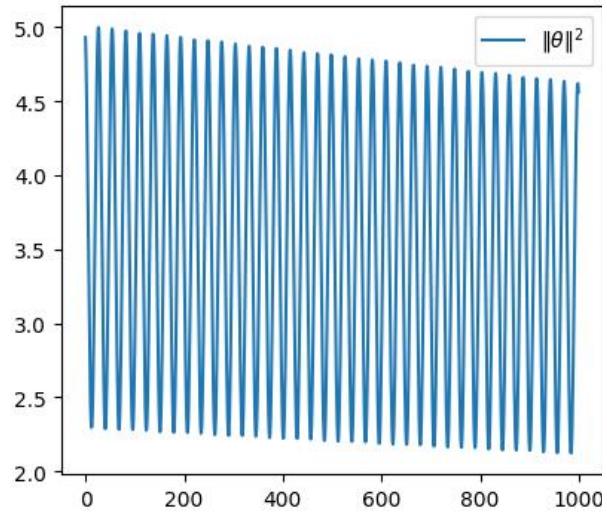
With momentum resampling



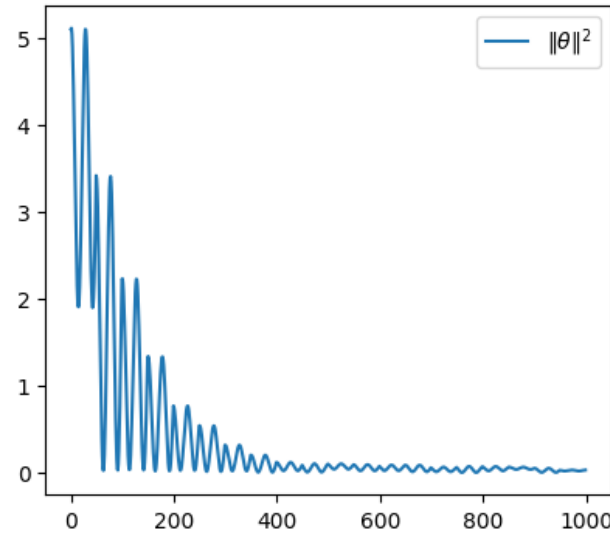
With momentum resampling ($\beta = 0.1$)

Observed problem

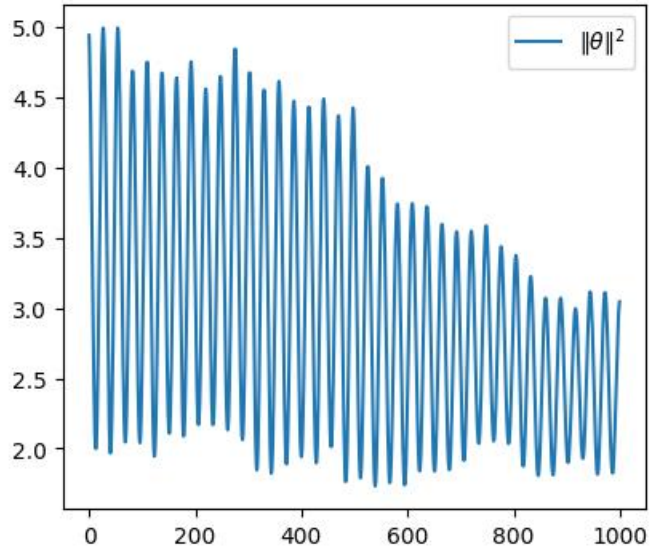
- Obviously, if we use our SGHMC framework, the result is the similar (when $T \cong 0$)



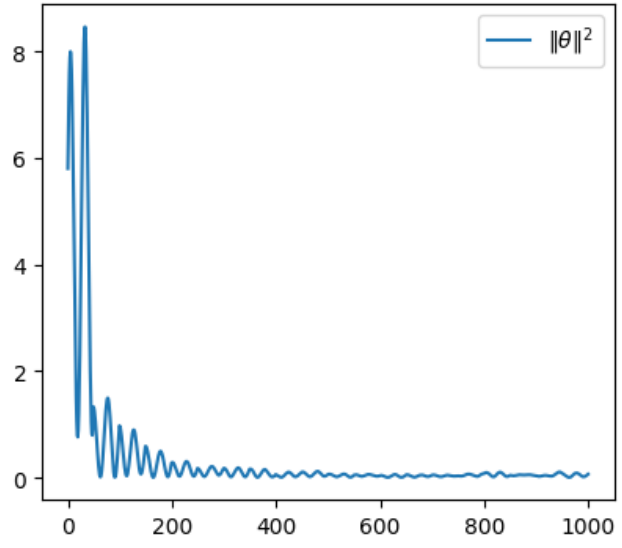
Without M.R



With M.R ($\beta = 0.1$)



Without M.R ($T = 1$)



With M.R ($\beta = 0.1$, $T = 1$)

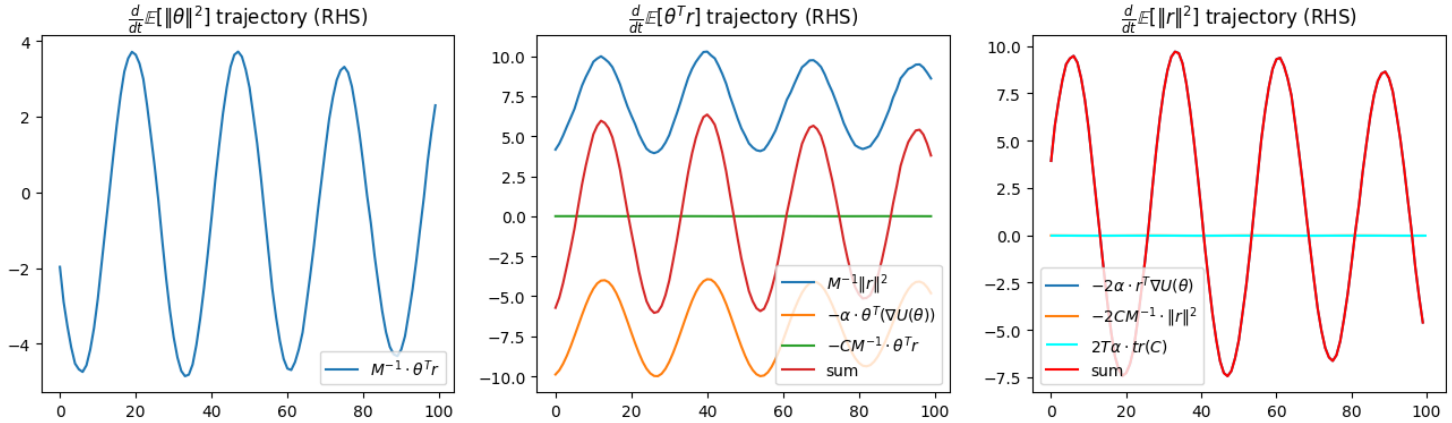
- One important observation is that **the attenuation of $\|\theta\|_2$ *toward the θ_{MAP}* is observed similarly in SDE, too,** which illustrates the $\|\theta\|_2$ attenuation effect during our experiments.

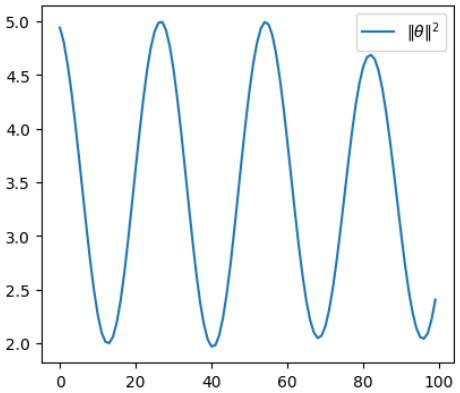
Observed problem (Ablation)

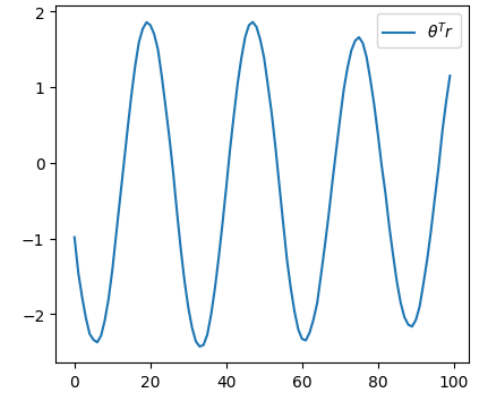
- During this simple toy problem, we can also verify the weight norm behavior explained by

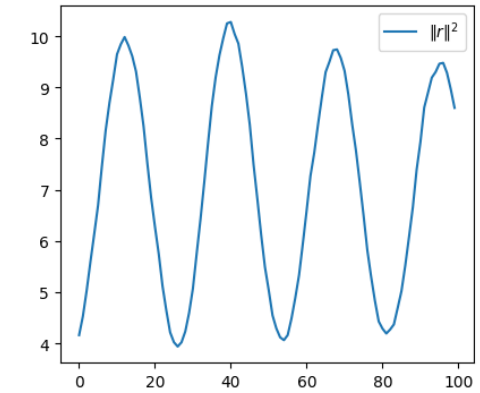
$$\frac{d}{dt}\mathbb{E}[\|\theta\|^2] = 2M^{-1}\mathbb{E}[\theta^T r], \qquad \frac{d}{dt}\mathbb{E}[\theta^T r] = \mathbb{E}[M^{-1}\|r\|^2 - \theta^T(\nabla U(\theta) + CM^{-1}r)]$$

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





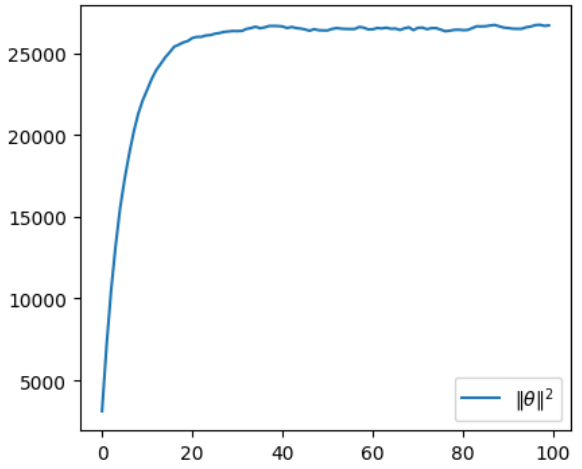




Note:
The derivative of 2nd row is approximately the same as the 1st row value.

Observed problem (Ablation)

- Now, we highlight the phenomenon of C.W.N.A via this toy example:



- If we assume the SGHMC achieves C.W.N.A ***(Q: Is it guaranteed??)***

  (i.e: $\mathbb{E}[\|\theta\|^2] \to L$ for some constant L as $t \to \infty$)

  - By our relations on weight norm:

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

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

  (Take $\mathbb{E}[\theta^T r] = 0, \nabla U(\theta) = \theta$)

  - We have the followings: (assuming $C, M \in \mathbb{R}$)

$$M^{-1}\mathbb{E}[\|r\|^2] = \mathbb{E}[\|\theta\|^2], \qquad \mathbb{E}[\|r\|^2] = \frac{2TM}{C} \cdot tr(C \cdot I_d) = TMd$$

$$\therefore \mathbb{E}[\|\theta\|^2] = T \cdot tr(I_d) = Td \quad \textbf{(dependent on } \boldsymbol{T}, \boldsymbol{d} \textbf{ only)}$$

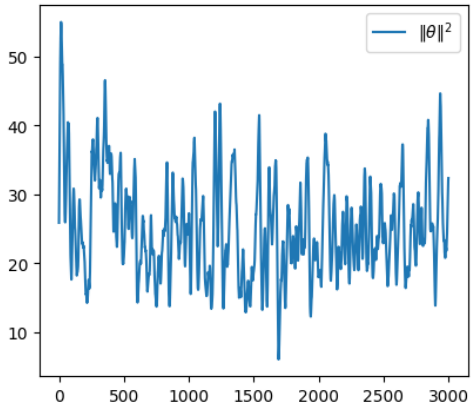
Observed problem (Ablation)

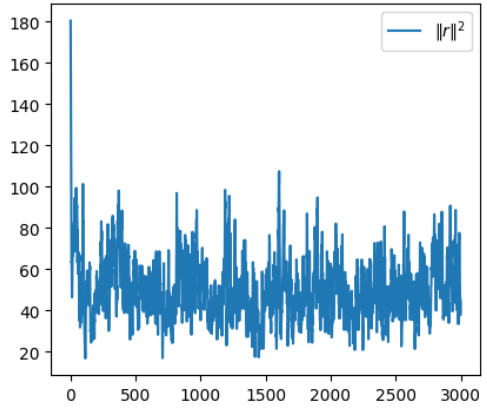
- We have the followings: (assuming $C, M \in \mathbb{R}$)

$$M^{-1}\mathbb{E}[\|r\|^2] = \mathbb{E}[\|\theta\|^2], \qquad \mathbb{E}[\|r\|^2] = \frac{2TM}{C} \cdot tr(C \cdot I_d) = 2TMd$$

$$\therefore \mathbb{E}[\|\theta\|^2] = 2T \cdot tr(I_d) = 2Td \quad \textbf{(dependent on } \boldsymbol{T}, \boldsymbol{d} \textbf{ only)}$$

- Is our analysis accurate under experiments? → Yes (SGHMC, $C = 1, M = 2, T = 0.1, d = 25$)

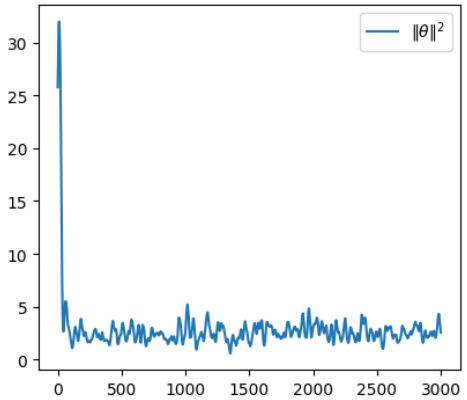


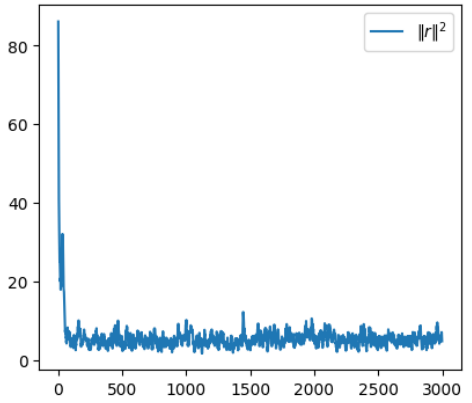


$\boldsymbol{T = 1}$

$\mathbb{E}\left[\|\boldsymbol{\theta}\|^2\right] = \mathbf{24.0733}$
$\mathbb{E}\left[\|\boldsymbol{r}\|^2\right] = \mathbf{48.8929}$

(from iter: 1k~3k)





$\boldsymbol{T = 0.1}$

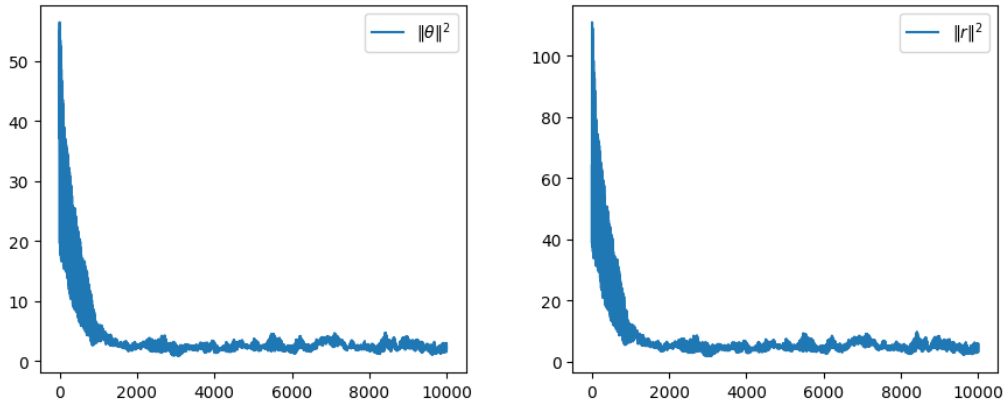
$\mathbb{E}\left[\|\boldsymbol{\theta}\|^2\right] = \mathbf{2.5538}$
$\mathbb{E}\left[\|\boldsymbol{r}\|^2\right] = \mathbf{5.2167}$

(from iter: 1k~3k)

Observed problem (Ablation)

- Then, what is the role of C, M? → changing the convergence speed toward C.W.N.A

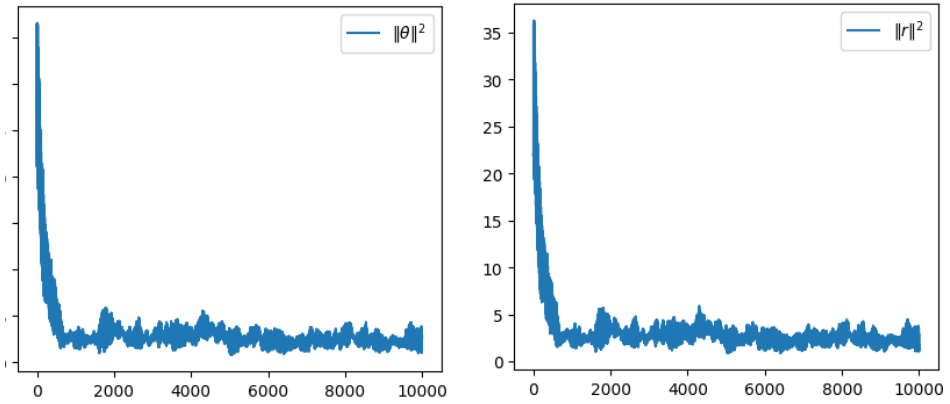
$$\lambda_{sol} = \frac{-CM^{-1} \pm \sqrt{(CM^{-1})^2 - 4M^{-1}}}{2}$$



$M = 2, C = 0.01, T = 0.1$

$\mathbb{E}[\|\theta\|^2] = 2.5507$
$\mathbb{E}[\|r\|^2] = 5.1100$

(from iter: 7k~10k)



$M = 1, C = 0.01, T = 0.1$

$\mathbb{E}[\|\theta\|^2] = 2.4345$
$\mathbb{E}[\|r\|^2] = 2.4421$

(from iter: 7k~10k)

- Note : this phenomenon (C.W.N.A and convergence speed) appears similarly in the MNIST

Observed problem (Ablation)

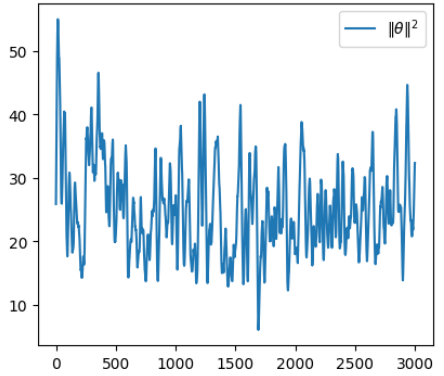
- Remained important questions for cold posterior effect :

1. ***Why is the convergence of*** $\mathbb{E}\left[\|\theta\|^2\right]$ ***achieved as*** $t \to \infty$ ***in such SDE?? (What is the asymptotic behavior?)***
2. ***The behavior of NN is similar as if we assumed*** $\theta|\mathcal{D} \sim MN\left(\mu, \sigma^2 I_d\right)$***, Is there any possibility that the posterior of NN approximately follows the MN as*** $d \to \infty$?
3. ***Can we guarantee that the samples attained on the high C.W.N.A are better than the samples attained on the low C.W.N.A ?*** (if true, it directly implies the advantage of cold posterior effect)
   - Empirically true by checking the samples' test accuracy by adjusting temperature T: $1.0 \to 0.1$
   - Can be related to the Rademacher complexity, but seems to be a little weak evidence.
     (Desired claim : "<u>local minima</u> at low $\|\theta\|^2$ **have better generalization performance than <u>local minima</u> at high** $\|\theta\|^2$ **in average.) (Because we need to consider weight norm regularization via high** λ **in prior)**
4. Can we devise a new strategy enjoying the Bayesian concept while outperforms the standard SGD?
   1. Use temperature scheduling (seems to be the best, and simple method)
   2. Use scaled momentum scheduling (but, it may not achieve target distribution $p^s(\theta) \propto exp\left(-U(\theta)\right)$)

Curse of dimensionality in SGHMC (Ablation)

- Recall that the $\|\theta\|^2$ cannot touch the MAP line ($\|\theta\|^2 = 0$) on C.W.N.A even when $d = 25$.
  - This alludes a ***curse of dimensionality problem*** in SGHMC.
  - Note that SGHMC consider the volume of integration : $\mathbb{E}[h(X)] = \int_X h(x) f(x) dx$

    (Not only the density $f(x) \rightarrow$ Can it be problematic if # of samples are very low while the dimension is very high?)



- Note that common NN use parameters ≅ 10k~100k per each layer

  → never touch the MAP point via SGHMC (= showing very bad mixing)

- How to make the SGHMC can touch the MAP line frequently during the sampling??
  1. Use narrow, but tall NN (Impossible ∵ Curse of dimensionality appears too fast (ex: $d = 25$))
  2. Remove the C.W.N.A (can we achieve this?)

     → Abandon $T = 1$ and use temperature scheduling to enforce mixing. (or make it not converge)

Observed problem (Ablation)

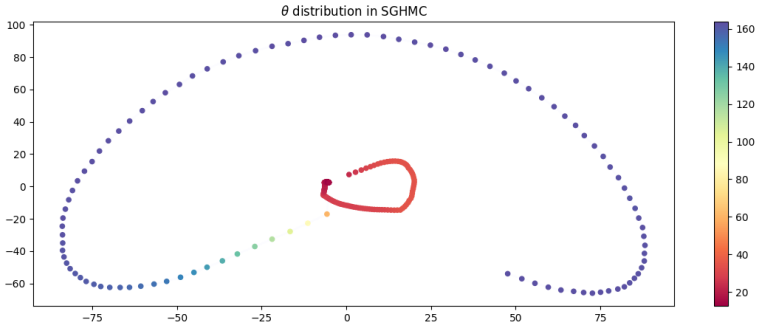
- Back to the original problem, we could explain why the momentum resampling is effective to attenuate $\|\theta\|_2$ toward θ_{MAP} (observed from the theoretical analysis of drift term in simple MN posterior)

- Now, can we perform better than $T \cong 0$ (cold posterior) with aid of momentum resampling? : Here are some crucial factors to consider:

    1. Since we have limited # of samples, **the samples must be chosen from (local) modes, not <u>near</u> the (local) modes.**
    2. To enjoy the advantages of Bayesian concept, **we should get rid of the phenomenon : " the constant norm area" at the end of the training** → Make the sample trajectory attenuated by using the β scheduler.

Strategies to be better (Ablation)

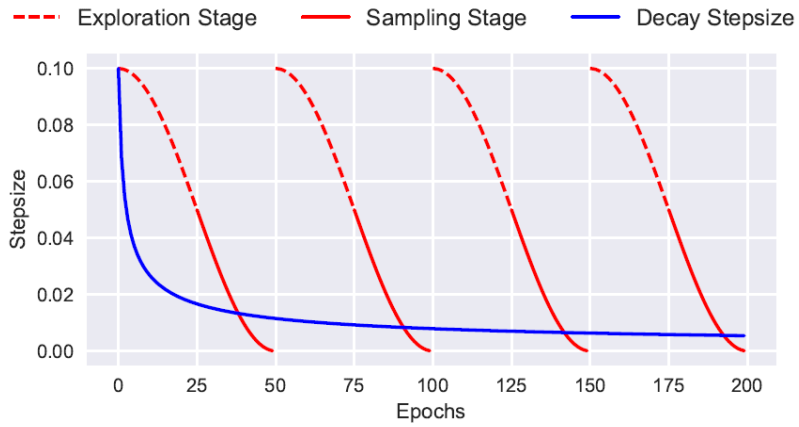
- Criteria 1 : **The samples must be chosen from (local) modes, not near the (local) modes.**
  - It turned out that the random direction of momentum by frequent M.R is harmful in practice, which leads to significant performance drop in MNIST. [NLL : 0.1601 > 0.1326 (cold posterior)]



  - A new strategy to resolve this is to use momentum decaying: ⇒ $\|\theta\|_2$ **is attenuated toward** θ_{MAP}
    - For each iteration, $r = r \times \zeta$ (where $\zeta = 0.9$, in practice)
    - After taking a sample,
      - Sample $r \sim N(0, \beta M)$ (where smaller β decrease the constant of weight norm zone)

Strategies to be better (Ablation)

- Criteria 2 : **“ the constant norm area” at the end of the training should be resolved.**
- This is possible via the “momentum scaler β scheduler”



- A new strategy to resolve this is to use : β scheduler
  - For each iteration, $r = r \times \zeta$ (where $\zeta = 0.9$, in practice)
  - After taking a sample,
    - Sample $r \sim N(0, \beta M)$ (where smaller β decrease the constant of weight norm zone)

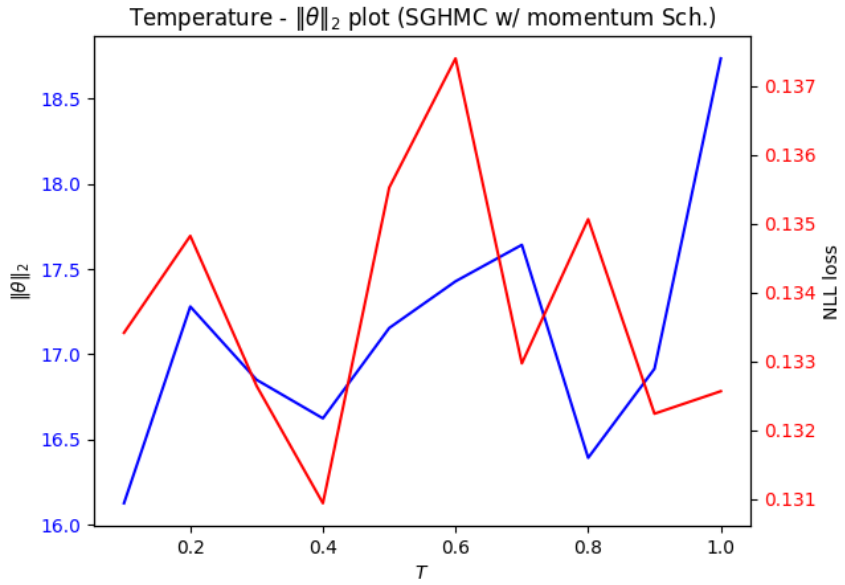
      where $\boldsymbol{\beta = [0.5, 0.4, 0.3, \ldots 0.01]}$ decreases as the step t increases.

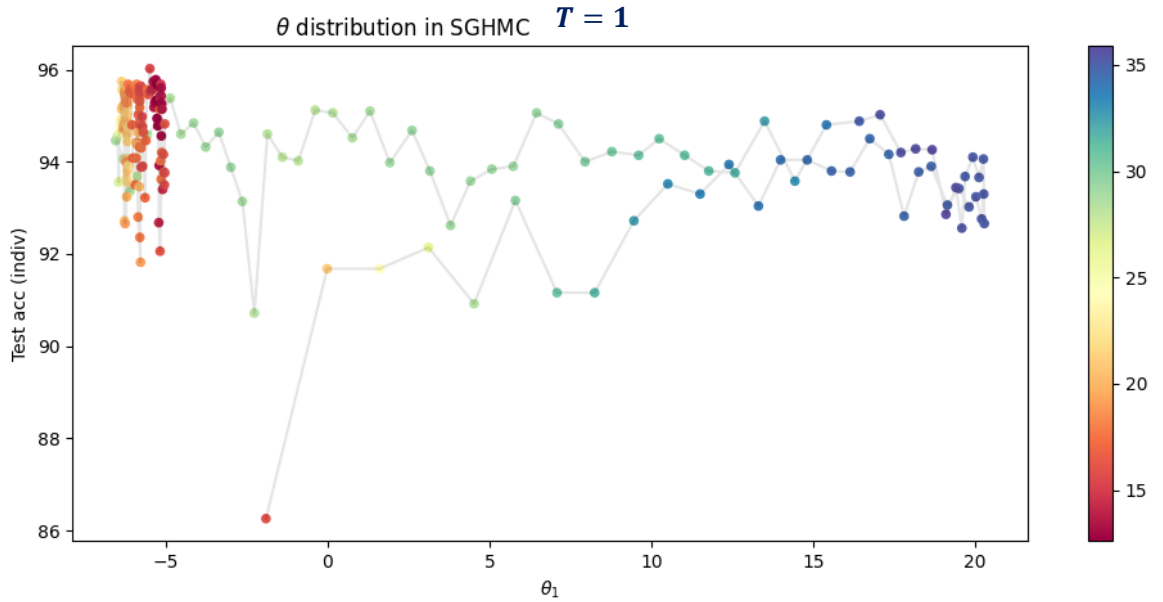
- ***One critical issue : This strategy may lose the target distribution $p^s(\theta) \propto exp\left(-U(\theta)\right)$ due to the momentum decaying step: $r = r \times \zeta$ for each iteration***

Strategies to be better (Ablation)

Problem : originally 95.5% is target in C.P
→ we require to find another parameter for C.P

- Results in MNIST / CIFAR-10 (data aug): (Note: all data is pruned → 10% train set + 50% test set)





MNIST (pruned)

- Note: CIFAR-10 (full data) – R18 – 200 epoch (w/ batch size = 100):

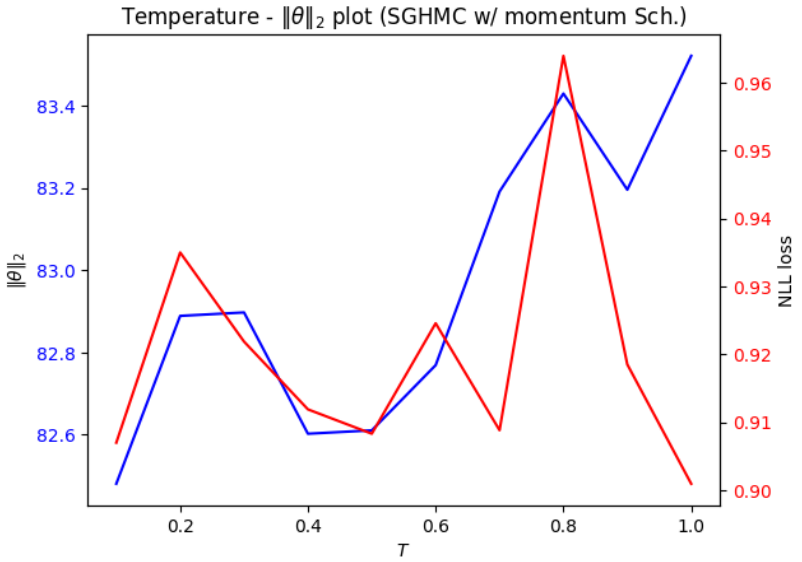
For full CIFAR-10, we used the same hyper parameter as in MNIST(pruned)

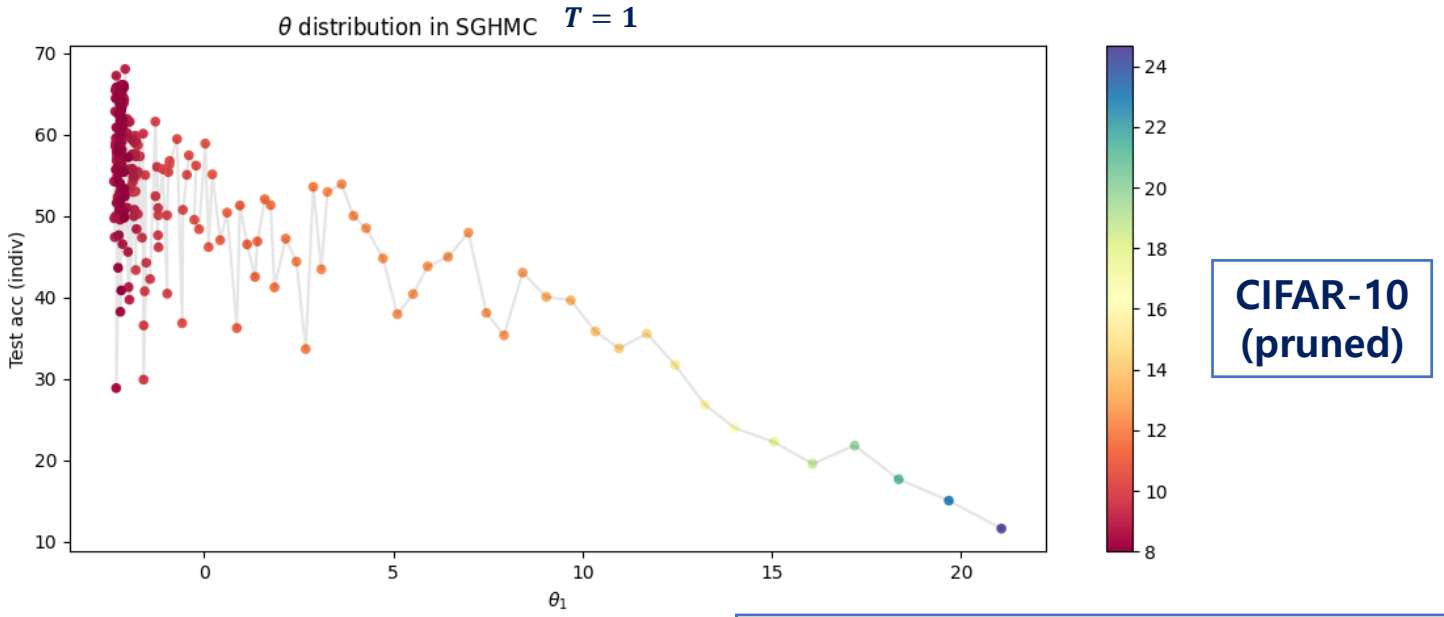
  - Momentum scheduler ($T = 1$): 94.87% w/ NLL : 0.2129
  - Cold posterior : 91.91% w/ NLL : 0.3096 ($T = 0.1$) , 94.39% w/ NLL : 0.2353 ($T = 0.01$)
  - Baseline (SGD : lr = 0.1, wd = 5e-4, momentum = 0.9 w/ cosine-annealing) : 95.45%

Strategies to be better (Ablation)

Problem : originally 95.5% is target in C.P
→ we require to find another parameter for C.P

- Results in MNIST / CIFAR-10 (data aug): (Note: all data is pruned → 10% train set + 50% test set)





CIFAR-10 (pruned)

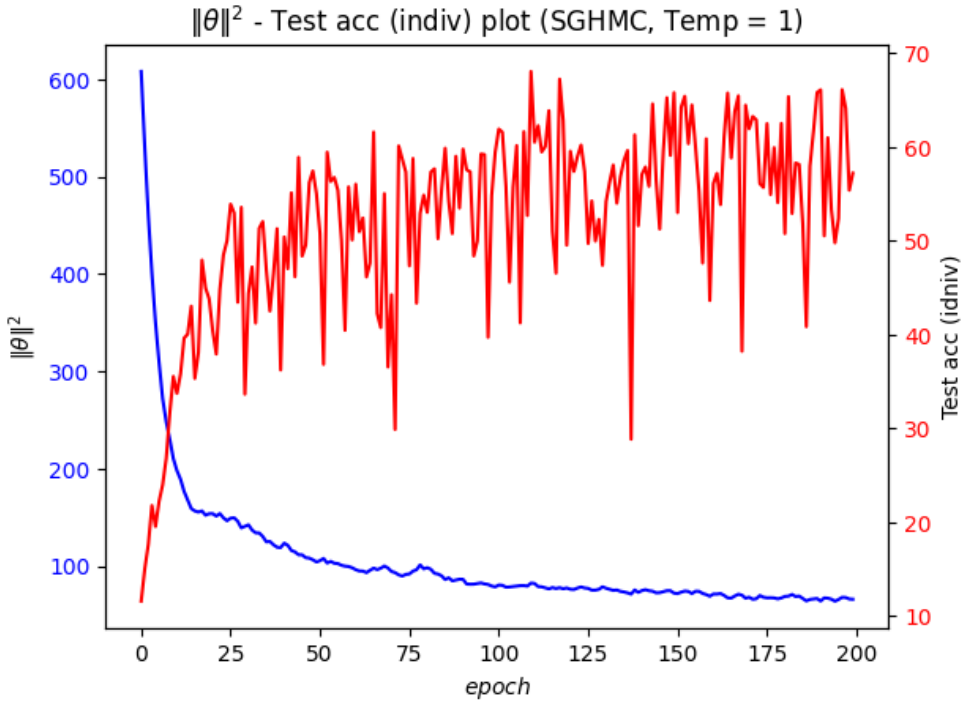
- Note: CIFAR-10 (full data) – R18 – 200 epoch (w/ batch size = 100):

For full CIFAR-10, we used the same hyper parameter as in MNIST(pruned)

  - Momentum scheduler ($T = 1$): 94.87% w/ NLL : 0.2129
  - Cold posterior : 91.91% w/ NLL : 0.3096 ($T = 0.1$) , 94.39% w/ NLL : 0.2353 ($T = 0.01$)
  - Baseline (SGD : lr = 0.1, wd = 5e-4, momentum = 0.9 w/ cosine-annealing) : 95.45%

Strategies to be better (Ablation)

- Question : Does the individual test accuracy increases as the $\theta^{(s)} \to \theta_{MAP}$ (which has low norm)?



CIFAR-10 (pruned)

- Not only in MNIST, but also the same thing occurs in CIFAR-10 (pruned)

- Note: By this reason, we may benefits NLL loss from taking sample $\theta^{(s)} \cong \theta_{MAP}$ as the $\beta \to 0$.

  (The reason why the sample with low-norm is beneficial can be addressed by Rademacher complexity)