Cold posterior effect

-Summary-

23/10/25

Observed problem (Review)

- Constant weight norm area phenomenon via this toy example:
  - By our relations on weight norm: (Assume standard gaussian posterior)

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

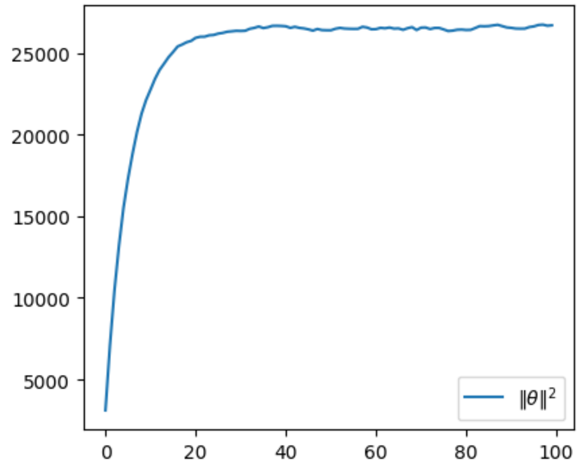
$\frac{d}{dt}\mathbb{E}[\|r\|^2] = -2\mathbb{E}[r^T(\nabla U(\theta) + CM^{-1}r)] + 2T \cdot tr(C)$ (= $2tr(C)$ if w/o cold posterior)

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

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

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

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



Observed problem

- Q: does the constant weight norm behavior is wrong circumstance?
  - A: No, it is a good signal to imply sampling around a 'typical set'
- Observation :
  - **[Typical set perspective]:** Area where the volume integral ($= p(\theta|x)dw$) is maximized

    [Recall: Posterior predictive => $p(y|x) = \int p(y|\theta, x)p(\theta|x)dw$]

    If we assume isotropic gaussian posterior (i.e: $\theta|x \sim MN(0, I_d)$) and use symmetry of sphere,

$$p(\theta|x)\frac{d\theta}{d\|\theta\|} \propto \exp\left(-\frac{1}{2}\|\theta\|^2\right)\|\theta\|^{(d-1)}, \qquad \sqrt{(\boldsymbol{d-1})} = \operatorname{argmax}_{\|\theta\| \in \mathbb{R}^+} p(\theta|x)\frac{d\theta}{d\|\theta\|}$$

c.f : $\mathbb{E}\left[\|\boldsymbol{\theta}\|^2\right] = \boldsymbol{T} \cdot \boldsymbol{tr}(\boldsymbol{I_d}) = \boldsymbol{Td}$ (Very similar when $T = 1$)

Observed problem

- Furthermore, we can prove the following fact.

  - Under the stochastic system suggested in [YA Ma, 2015] : $dz = f(z)dt + \sqrt{2D(z)}dW$

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

Theorem 1

Under above stochastic system, if we assume $D(z), Q(z)$ are constant matrix satisfying the condition : $D(z)$ is P.S.D and $Q(z)$ is skew symmetric (that is, $D(z) = \begin{bmatrix} A & H \\ G & F \end{bmatrix} \succcurlyeq 0, Q(z) = \begin{bmatrix} 0 & B \\ -B & 0 \end{bmatrix}$), the following asymptotic relation holds when $U(\theta) \propto -\log\left(\exp\left(-\frac{1}{2}\|\theta\|^2\right)\right)$ [Standard gaussian posterior]:

$$\mathbb{E}\left[\|\boldsymbol{r}\|^2\right] = \boldsymbol{Md}, \qquad \mathbb{E}\left[\|\boldsymbol{\theta}\|^2\right] = \boldsymbol{tr}(\boldsymbol{I}_d) = \boldsymbol{d}$$

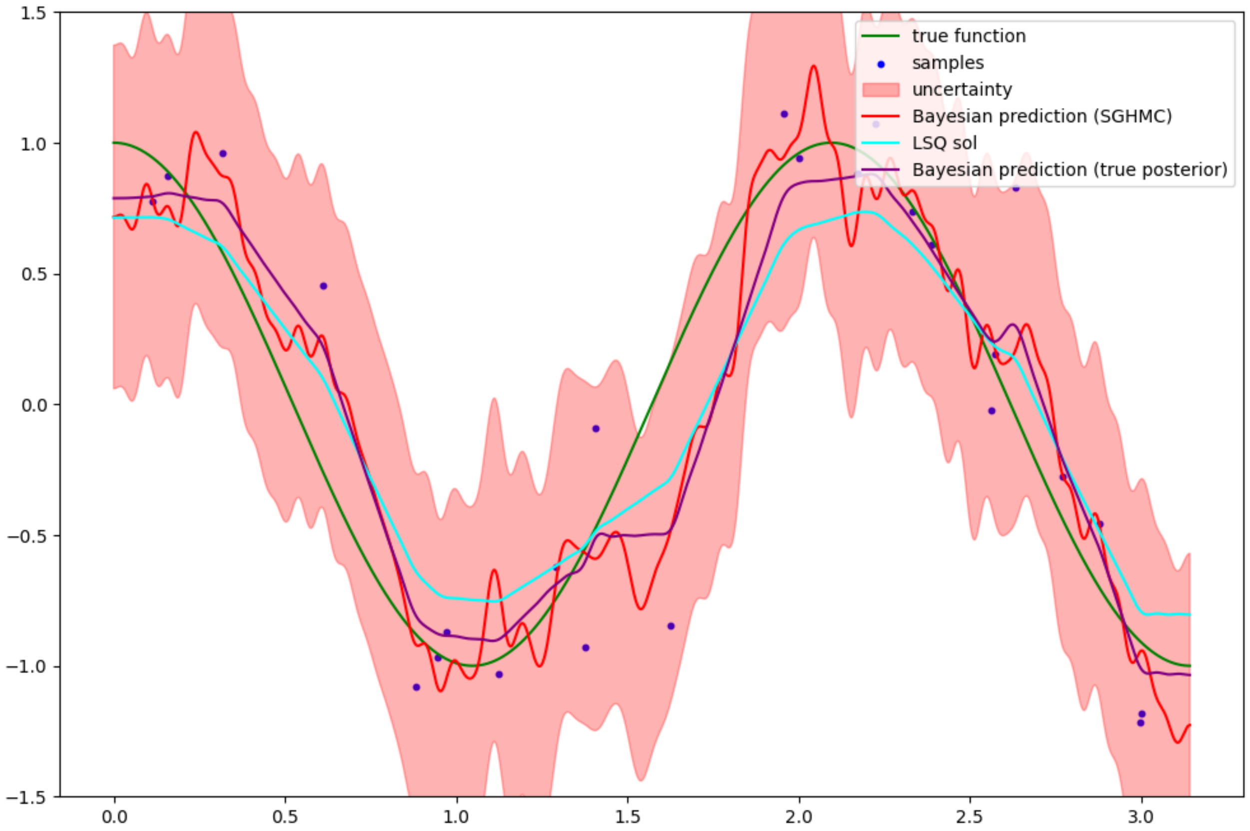
Observed problem

- Some fact about typical set:
  - As the dimension of θ gets bigger, ***the radius of typical set get far away from the origin***, and the ***thickness of typical set become thinner.*** (Which is the fact we have observed so far)

- ***Back to the origin of problem…***
  - This context implies there is no difficulty to reach the samples from typical set in practical BNN.
  - Then, the empirical results should show the SGHMC is almost exact compared to HMC in the sense of approximation error (thanks to SGHMC strong ability to capture typical sets)

Observed problem

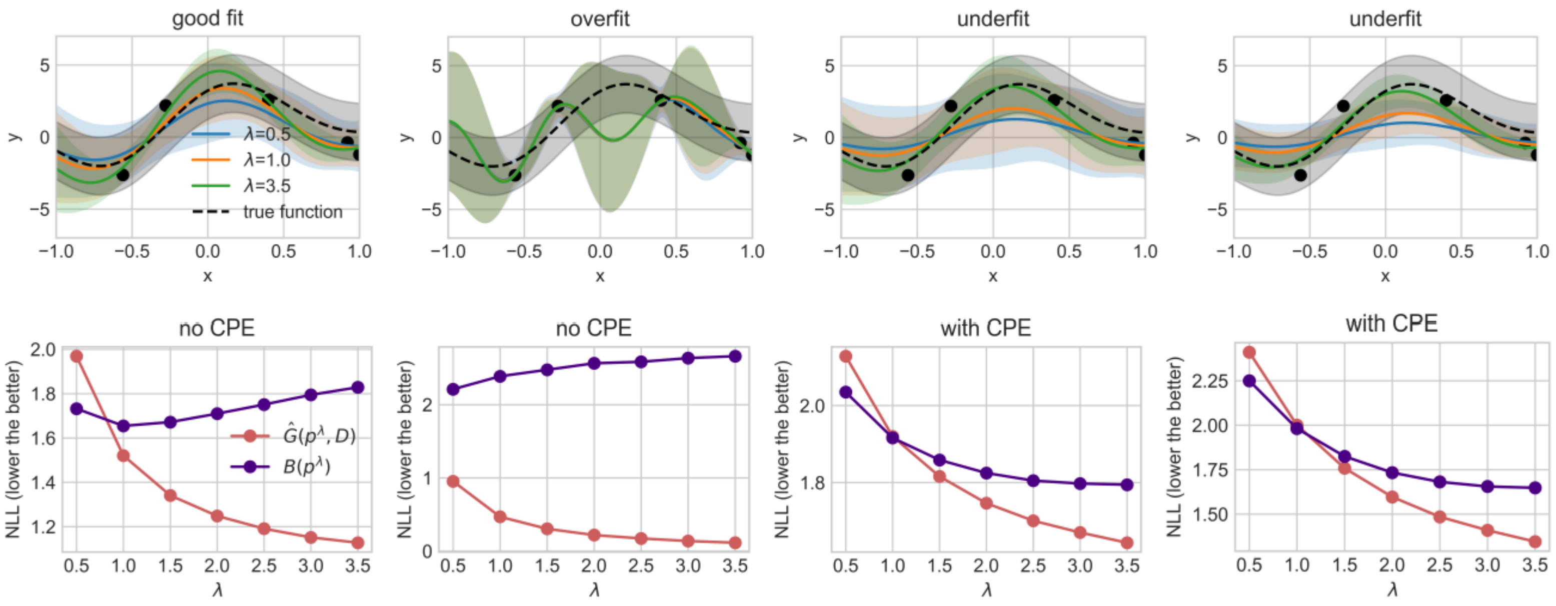
- Understanding the approximation error in Bayesian linear regression:
  - Experiment setting : Linear regression of true function $\cos(3x)$ with n =25 samples
  - Set feature dim : $p = 100$ / use attenuated cosine kernel : $\{\cos(mx)\ /m\}_{m=1}^{p}$ *(for overparametrized circumstance)*



- *Result:*
  - The approximation error does not seem to big when $T = 1$

Observed problem

- Q: Then, what is the fundamental reason of cold posterior??
  - Recently [Y. Zhang, 2023] suggested underfitting of BNN as the critical factor for cold posterior.



(a) No misspecification (b) Misspecified likelihood I (c) Misspecified likelihood II (d) Misspecified prior

Notations:

- $B(p^\lambda)$ = **Test NLL loss by B.M.A with** $T = 1/\lambda$
- $\hat{G}(p^\lambda, \mathcal{D})$ = **Train NLL loss of BNN with** $T = 1/\lambda$ **(may be the last sample train NLL loss)**
- **Misspecified likelihood :** **set** $p(y|x,\theta) = N(\theta^T x, \tilde{\sigma}^2)$, **where** $\tilde{\sigma} \neq \sigma$ **(true std)**
- **Misspecified prior :** **set** $p(\theta) = N(0, \sigma_p^2)$ **with** $\sigma_p \cong 0$ **(very informative prior)**

Observed problem

- Furthermore, [Y. Zhang, 2023] suggested one necessary condition for cold posterior effect

Proposition 3. *A necessary condition for the presence of the CPE, as defined in Definition 1, is that*

$$\hat{G}(p^{\lambda=1}, D) > \min_{\boldsymbol{\theta}} - \ln p(D|\boldsymbol{\theta}) \, .$$

- That is, if the cold posterior effect appears, the model cannot achieve the minimum train loss (underfitting).

- *Similarly, our experiments in MNIST, Fashion-MNIST, CIFAR-10 shows significant underfitting problem under BNN training.*

Observed problem

- Another attractive research is [S. Kapoor, 2022], which claims the softmax function misrepresent the belief about aleatoric uncertainty by lowering the model's confidence significantly.
- [Idea of the paper]

$$f(x) = (f_1(x, w), \ldots, f_C(x, w))$$

1. Recall that $p(w|\mathcal{D}) \propto p(w)\Pi_{x,y\in\mathcal{D}}\, f_y(x,w)$, where $f_y(x,w) = f_y(x)$ = y-th index predicted probability of example x by weight w.
2. Observe : $p(y|f(x)) = f_y(x) \propto Dir(1, \ldots 1)\big(f(x)\big) \cdot f_y(x)$, where Dir is pdf of Dirichlet distribution (Just introduce auxiliary uniform prior $Dir(1, \ldots 1)f(x)$)

Observed problem

$\boxed{\boldsymbol{Dir(\alpha_1, \ldots \alpha_K)(x)}}$ $f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

3. Observe : $p(y|f(x)) = f_y(x) \propto Dir(1, \ldots 1)\big(f(x)\big) \cdot f_y(x)$, where Dir is pdf of Dirichlet distribution (Just introduce uniform prior $Dir(1, \ldots 1)$ for $f(x)$)

4. Note that $y|f(x) \sim \text{Categorical}\big(f_1(x), \ldots f_C(x)\big)$ and $f(x) \sim Dir(1, \ldots 1)$. Hence, by the conjugate prior property of Dirichlet distribution, $f(x)|y \sim Dir(1, \ldots 2 \text{ (y th)}, \ldots 1)$

   $\Rightarrow \mathbb{E}\big[f_y(x)|y\big] = \frac{2}{C+1} \cong 2\%$ (C = # of classes) $\boxed{\textbf{Assume } C = 100}$

   (That is, the confidence change per an observation (x, y) is very low)

5. Also $p(y|f(x)) \propto Dir(1, \ldots, 2 \text{ } (y \text{ th}), \ldots 1)\big(f(x)\big)$ by red formula, which again shows the relation: $p(w|D) \propto p(w)\Pi_{x,y \in \mathcal{D}} Dir(1, \ldots, 2, \ldots 1)(f(x))$

Observed problem

Q: What happen on confidence change if we adopt cold posterior?

1. Now, we use $p_{cold}(w|D) \propto p(w)\Pi_{(x,y)\in\mathcal{D}}\, f_y(x,w)^{1/T}$ as a posterior of $w|D$

2. Similarly, $p_{cold}(w|D) \propto p(w)\Pi_{(x,y)\in\mathcal{D}}\, \mathrm{Dir}\left(1,\dots,1+\frac{1}{T},\dots 1\right)\left(f(x)\right)$

3. Then, $f(x)|y \sim Dir\left(1,\dots,1+\frac{1}{T},\dots 1\right) \Rightarrow \mathbb{E}\left[f_y(x)|y\right] = \frac{T+1}{CT+1} \cong 50.5\%$ *(if* $T = 10^{-2}$*)*

In other words, the confidence gain per an observation (x, y) can be amplified by T.

Q: Can we achieve the same confidence gain without using cold posterior?

Observed problem

Q: Can we achieve the same confidence gain without using cold posterior?

⇒ Use Dirichlet model.

- Idea : Drop uniform prior for $f(x) \sim Dir(1 \ldots .1)$, and use attenuated concentration parameters : $f(x) \sim Dir(\alpha_\epsilon, \ldots \alpha_\epsilon)$ where $\alpha_\epsilon \ll 1$
- Then, $p_{ND}(w|D) \propto p(w)\Pi_{(x,y)\in\mathcal{D}}\mathrm{Dir}(\alpha_\epsilon, \ldots \alpha_\epsilon + 1, \ldots \alpha_\epsilon)(f(x))$ [Noisy Dirichlet model]
- In this case, $\mathbb{E}\left[f_y(x)\right] = \frac{\alpha_\epsilon+1}{C\alpha_\epsilon+1} \cong 50.5\%$ if $\alpha_\epsilon = 10^{-2}$ (Same effect with $T = 10^{-2}$)

Observed problem

- Q: Can we deploy this scheme into SGMCMC methods? (Yes)
  - Observe $p_{ND}(w|D) \propto p(w)\Pi_{(x,y)\in\mathcal{D}}\mathrm{Dir}(\alpha_\epsilon, \ldots \alpha_\epsilon, \ldots \alpha_\epsilon)\big(f(x)\big) \cdot f_y(x) = q_{ND}(w) f_y(w)$

    where $q_{ND}(w) = p(w)\Pi_{(x,y)\in\mathcal{D}}\mathrm{Dir}(\alpha_\epsilon, \ldots \alpha_\epsilon, \ldots \alpha_\epsilon)\big(f(x)\big)$ **(data-dependent prior)**
  - Since we exactly calculate $p(w)\Pi_{(x,y)\in\mathcal{D}}\mathrm{Dir}(\alpha_\epsilon, \ldots \alpha_\epsilon, \ldots \alpha_\epsilon)\big(f(x)\big)$, it is possible to use SGMCMC.
- But, it turns out that it is unstable numerically ⇒ Adopt Noisy Dirichlet Gaussian approx.

$$p_{NDG}(w \mid D) \propto p(w) \prod_{x,y\in\mathcal{D}} \prod_{c=1}^{C} \mathcal{N}(z_c(x) \mid \mu_c, \sigma_c^2), \quad \text{with}$$

$$\alpha_c = 1 + \alpha_\epsilon \cdot I[c = y], \quad \sigma_c^2 = \log(1/\alpha_c + 1), \quad \mu_c = \log(\alpha_c) - \frac{\sigma_c^2}{2},$$

Where $z_c(x)$ is c th logit value by model

Note:

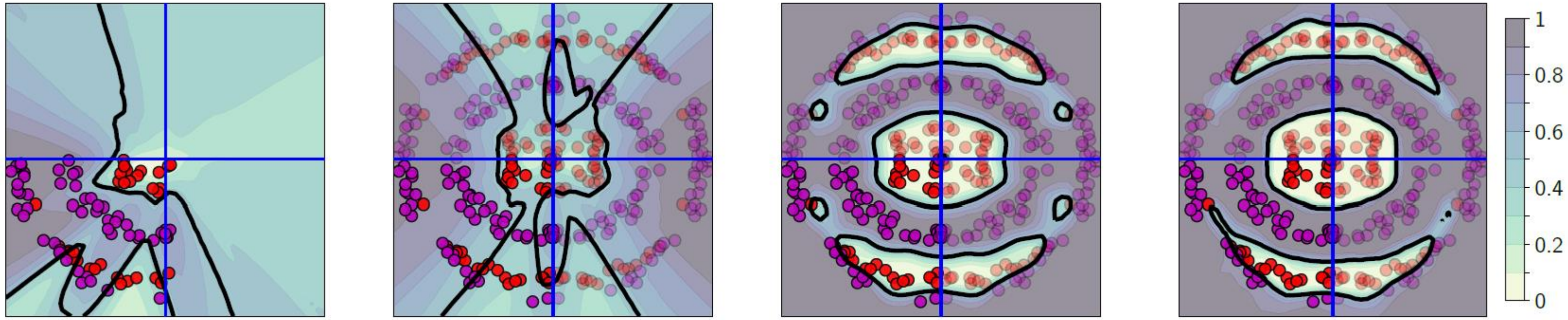
We get rid of softmax and use batch-wise approximation as in SGMCMC.

Observed problem

- Intuition of ND ? : $p_{ND}(w|D) \propto p(w)\Pi_{(x,y)\in\mathcal{D}}\mathrm{Dir}(\alpha_\epsilon, \ldots \alpha_\epsilon, \ldots \alpha_\epsilon)\big(f(x)\big) \cdot f_y(x)$
- Standard softmax method : we believe the aleatoric uncertainty are high for all training data
- Tempered softmax method : we believe the aleatoric uncertainty are low for all training data, ***even with some of unseen data (problematic)***
- Noisy Dirichlet method : we believe the aleatoric uncertainty are low for all training data, but high for unseen data **(better)**
- ***[Important Recall] : This paper coincide the observation of [Y. Zhang, 2023] by pinpointing the misspecified likelihood, and they fixed this problem by adopting data-dependent prior.***

Observed problem

- [Empirical results] (w/ cSGLD and $p(w) \sim N(0,1)$)

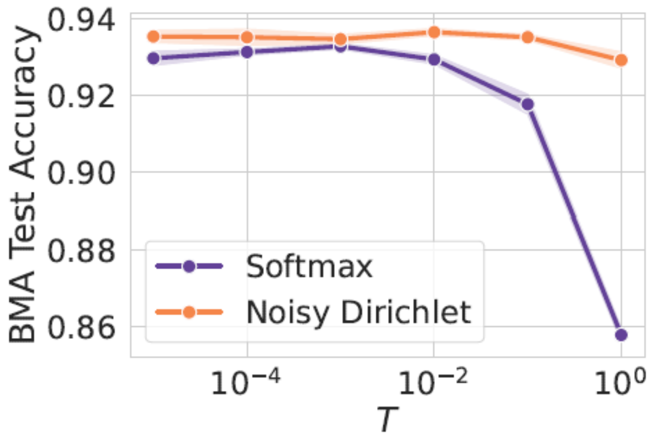


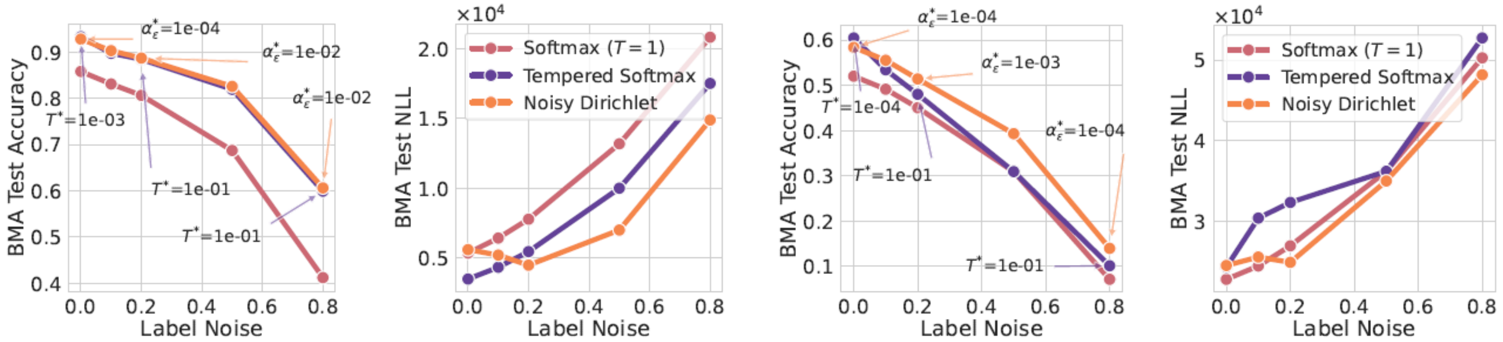
(a) Softmax Lik. (SL) (b) SL+Data Aug.(DA) (c) SL+DA+Tempering (d) Noisy Dirichlet+DA

- While $T = 1$ [(a),(b)] failed to fit even training data, Cold temperature or ND [(c),(d)] succeeded to classify correctly even with data augmentation

Observed problem

- [Empirical results]





(a) CIFAR-10 (b) Tiny Imagenet

- Importantly, the cold posterior effect disappeared and show better results compared to cold posterior when there are intense label noise)