A PROOF OF PROPOSITION 2

Proof A stationary distribution μ of (8) means $\partial_t \mu = 0$. Assuming $\mu = p(\boldsymbol{\theta}|\, \mathbf{X}) \triangleq p$, then we need to prove that

$$\nabla \cdot ((\mathbf{W} * p)p) = 0 \, .$$

By the definition of $\mathbf{W}$ in (9), and applying Stein's identity Liu and Wang (2016a), we have $\mathbf{W} * p = 0$. Consequently, we have $\nabla \cdot ((\mathbf{W} * p)p) = 0$.

The above argument indicates $p(\boldsymbol{\theta}|\, \mathbf{X})$ is a stationary distribution of (8). This completes the proof. ■

B MORE DETAILS ON LEMMA 3

We first specify the conditions the energy functional E needs to satisfy in Assumption 1.

Assumption 1 *The energy functional is assumed to*

- proper*:* $D(E) \triangleq \{\boldsymbol{\theta} \in \Omega : E(\boldsymbol{\theta}) < +\infty\} \neq \emptyset$*.*
- coercive*: There exists* $\tau_0 > 0$*,* $\boldsymbol{\theta}_0 \in \Omega$ *such that* $\inf \left\{ \frac{1}{2\tau_0} W_2^2(\boldsymbol{\theta}_0, \mathbf{v}) + E(\mathbf{v}) : \mathbf{v} \in \Omega \right\} > -\infty$*.*
- lower semicontinuous*: For all* $\boldsymbol{\theta}_n, \boldsymbol{\theta} \in \Omega$ *such that* $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$*,* $\liminf_{n \to \infty} E(\boldsymbol{\theta}_n) \geq E(\boldsymbol{\theta})$*.*
- convex*: E is convex in the sense that given* $\lambda \in \mathbb{R}$ *and a curve* $\boldsymbol{\theta}_\alpha \in \Omega$*,*

$$E(\boldsymbol{\theta}_\alpha) \leq (1 - \alpha) E(\boldsymbol{\theta}_0) + \alpha E(\boldsymbol{\theta}_1) \, .$$

Proof [Sketch proof of Lemma 3] Our case is just a simplification of Theorem 3.5.1 in Craig (2014), where we restrict the energy functional to be convex instead the more general case of λ-convex. For λ-convex energy functional, Craig (2014) proves that $W_2^2(\tilde{\mu}_{T/h}, \mu_T) \leq \sqrt{3} |\partial E|(\mu) e^{3\lambda^- T} \sqrt{Th}$ with $\lambda^- \triangleq \max\{0, -\lambda\}$. Our result follows by simply letting $\lambda^- = 0$, which is for the case of convex E. ■

C DERIVATION OF (22)

The derivation of (22) relies on the following Lemma from Carrillo et al. (2017).

Lemma 6 (Proposition 3.12 in Carrillo et al. (2017)) *Let* $F : (0, \infty) \to \mathbb{R}$ *belongs to* $C^2(0, +\infty)$ *and satisfy* $\lim_{s \to +\infty} F(s) = +\infty$ *and* $\liminf_{s \to 0} F(s)/s^\beta > -\infty$ *for some* $\beta > -2/(d+2)$*. Define*

$$\mathcal{F}(\mu) \triangleq \int F \circ (K * \mu) \mathrm{d}\mu \, ,$$

where $\circ$ *denotes function composition, i.e.,* F *is evaluated on the output of* $K * \mu$*. Then we have*

$$\nabla \frac{\delta \mathcal{F}}{\delta \mu} = \nabla \varphi_\epsilon * (F' \circ (\varphi_\epsilon * \mu) \mu) + (F' \circ (\varphi_\epsilon * \mu)) \nabla \varphi_\epsilon * \mu \, . \tag{25}$$

Now it is ready to derive (22). In this case, $F = \log(\cdot)$. Let $F_1 = \nabla \varphi_\epsilon * (F' \circ (\varphi_\epsilon * \mu) \mu)$, $F_2 = (F' \circ (\varphi_\epsilon * \mu)) \nabla \varphi_\epsilon * \mu$. We use particles to approximate μ, *e.g.*, $\mu \approx \frac{1}{M} \sum_{i=1}^M \delta_{\boldsymbol{\theta}^{(i)}i}$. We have

$$\begin{aligned} F_1(\boldsymbol{\theta}) &= \left(\nabla K * \frac{\mu}{K * \mu} \right) (\boldsymbol{\theta}) \\ &\approx \left(\nabla K * \left(\frac{1}{M} \sum_{i=1}^M \frac{\delta_{\boldsymbol{\theta}^{(i)}}}{(K * (\frac{1}{M} \sum_j \delta_{\boldsymbol{\theta}^{(j)}}))(\boldsymbol{\theta}^{(i)})} \right) \right) (\boldsymbol{\theta}) \\ &= \left(\nabla K * \left(\sum_{i=1}^M \frac{\delta_{\boldsymbol{\theta}^{(i)}}}{\sum_j K(\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(j)})} \right) \right) (\boldsymbol{\theta}) \\ &= \sum_{i=1}^M \frac{\nabla K(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})}{\sum_j K(\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(j)})} \, . \end{aligned} \tag{26}$$

$$\begin{aligned} F_2(\boldsymbol{\theta}) &= \left(\frac{\nabla K * \mu}{K * \mu} \right) (\boldsymbol{\theta}) \\ &\approx \left(\frac{\nabla K * \frac{1}{M} \sum_{i=1}^M \delta_{\boldsymbol{\theta}^{(i)}}}{K * \frac{1}{M} \sum_{i=1}^M \delta_{\boldsymbol{\theta}^{(i)}}} \right) (\boldsymbol{\theta}) \\ &= \frac{\sum_{i=1}^M \nabla K(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})}{\sum_{i=1}^M K(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})} \, . \end{aligned} \tag{27}$$

Combing (26) and (27) gives the formula for $\mathbf{v}$ in (22).

Table 4: Hyper-parameter settings for MNIST on FNN.

Datasets	MNIST	
Batch Size	100	100
Step Size	5×10^{-4}	5×10^{-4}
#Epoch	150	150
RMSProp	0.99	0.99
Network (hidden layers)	[400, 400]	[800, 800]
Variance in prior	1	1

D EXPERIMENTAL SETTING

We list some experimental settings in Table 4 and Table 5. For evaluation on BNNs, following a standard Bayesian treatment, we use ensemble of particle predictions to compute the test accuracy. We will need to store all the M particles in order to do particle optimization, thus the time and memory complexity would be proportional to the number of particles. In practice, however, we can reduce the complexity by only treating a small part of the parameters as particles (*e.g.*, the parameters of the last layer of a BNN), and leaving others as single values.

Table 5: Hyper-parameter settings for CIFAR-10 on CNN.

Datasets	`CIFAR10`
Batch size	128
Step size	0.01 ($< 5e^3$), 0.001 ($< 1e^4$) 0.0001
#Epoch	200
Filter size	3×3
Channels	C64-C128-C256
Network (hidden layers)	[1024]
Variance in prior	1

E EXTRA EXPERIMENTS

We further optimize 50 particles to approximate different distributions. The optimized particles are plotted in Figure 4, which shows that w-SGLD seems to be able to learn better particles due to the concentration property of the Wasserstein regularization term.

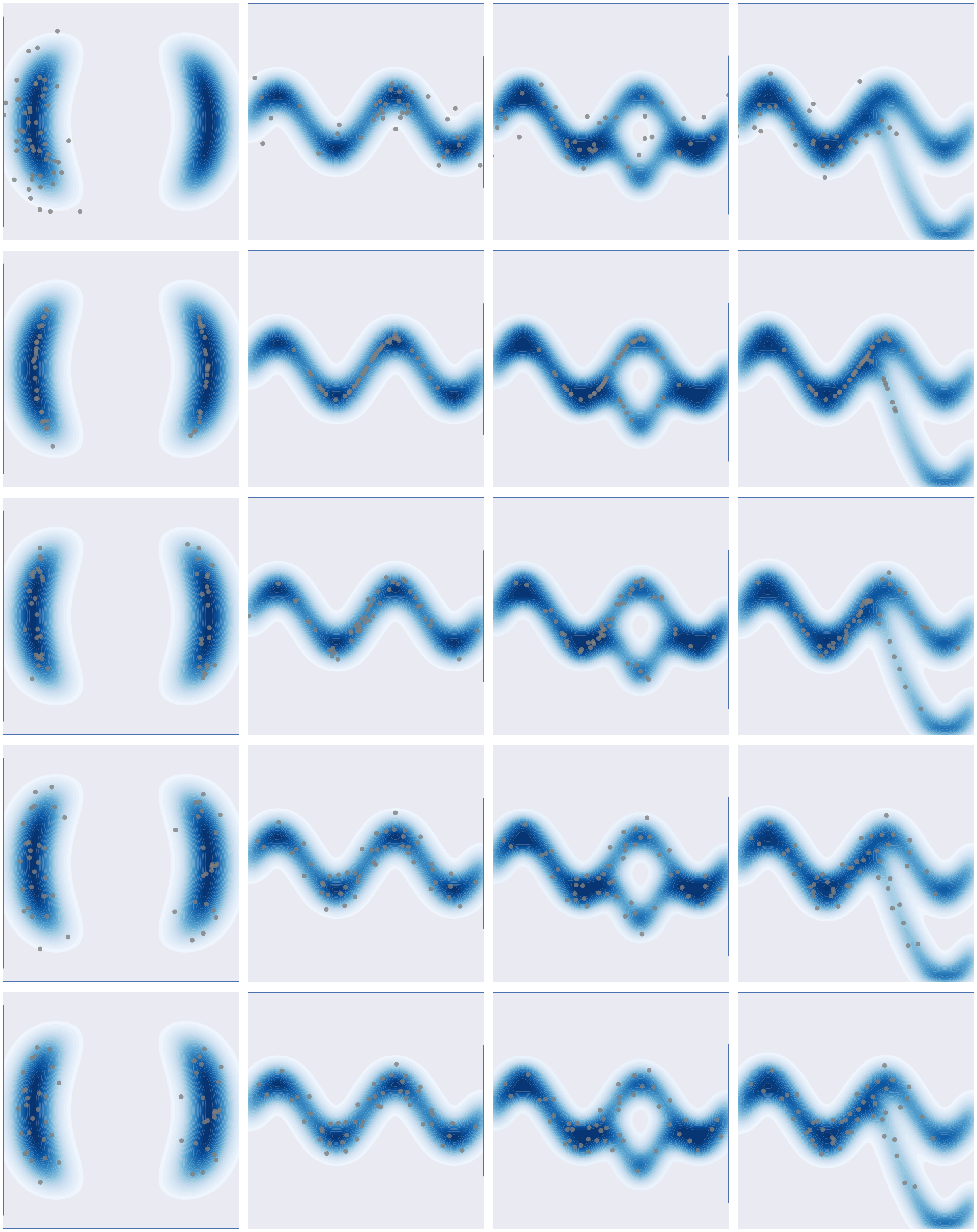

Figure 4: Illustration of different algorithms on toy distributions with 50 particles. Each column is a distribution case. 1st row: standard SGLD; 2nd row: w-SGLD; 3rd row: w-SGLD-B; 4th row: SVGD; 5th row: π-SGLD. The blue shapes are ground true density contours.