## A. More Details about the Notation

- One may notice the different use of $\boldsymbol{\theta}$ and $\theta$. $\boldsymbol{\theta}$ is mostly used for the interpretation of the theory; and $\theta$ is only used for the interpretation of algorithms, which means $\theta$ often appears with $k$ (which stands for the $k$th interation ) like $\theta_k$. The rules also apply for the results in Appendix.
- The symbol $\mathbf{1}(H_1 \leq H_2)$ in Theorem 3 means

$$\mathbf{1}(H_1 \leq H_2) = \begin{cases} 1 & H_1 \leq H_2 \\ 0 & H_1 > H_2 \end{cases} \tag{13}$$

  and the symbol $H_3 \wedge H_4$ means $\min\{H_3, H_4\}$

- The relationship between RBF kernel $\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|^2}{2\eta^2})$ and the function $K(\boldsymbol{\theta}) = \exp(-\frac{\|\boldsymbol{\theta}\|^2}{2\eta^2})$ can be interpreted as $\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}') = K(\boldsymbol{\theta} - \boldsymbol{\theta}')$ in detail.

## B. The Positive Constants in Theorem 1, Theorem 2 and Theorem 3

For the sake of clarity, we present the following constants which are used in our theorems.

$$\begin{aligned}
C_1 &= \frac{2(H_{\nabla K} + H_\theta)}{\sqrt{M}(\beta^{-1} - \frac{5}{2}H_\theta L_K - L_F - 2L_{\nabla K})} \\
C_2 &= \sqrt{2(\beta^{-1}L_F + 2L_K H_1 + H_K L_F + L_{\nabla K})^2 + 2} \\
C_3 &= \beta^{-1}m_F - 2L_F - 3H_1 L_K \\
C_4 &= \beta^{-1}D_F + 4D_{\nabla^2 K} + 4H_2 L_{\nabla K} + 2L_F H_{\nabla K} + 2H_1 L_K + L_F H_K \\
C_5 &= 2\beta^{-1}\sigma^2 + 2H_K^2\sigma^2
\end{aligned}$$

## C. Convergence Guarantees for SAGA-LD, SVRG-LD and SVRG-LD$^+$

In this section, we present the convergence guarantees for SAGA-LD, SVRG-LD and SVRG-LD+ from (Chatterji et al., 2018; Zou et al., 2018).

**Assumption 5**

- *(Sum-decomposable) The $F(\boldsymbol{\theta})$ is decomposable, i.e., $F(\boldsymbol{\theta}) = \sum_{j=1}^N F_j(\boldsymbol{\theta})$.*
- *(Smoothness) $F(\boldsymbol{\theta})$ is Lipschitz continuous with some positive constant, i.e., for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, $\|F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_2)\| \leq L_F \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.*
- *(Strong convexity) $F(\boldsymbol{\theta})$ is a $m_F$-strongly convex function, i.e., $(F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_2))(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq m_F \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.*
- *(Hessian Lischitz) There exits such a positive constant such that $\|\nabla F(\boldsymbol{\theta}_1) - \nabla F(\boldsymbol{\theta}_2)\| \leq D_F \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.*

**Assumption 6** *(Bound Variance)*[2] *There exits a constant $\sigma \geq 0$, such that for all j*

$$\mathbb{E}[\|F_j(\boldsymbol{\theta}) - \frac{1}{N}\sum_{j=1}^N F_j(\boldsymbol{\theta})\|^2] \leq d\sigma^2/N^2$$

**Theorem 4** *Under Assumption 5, let the step size $h < \frac{B}{8NL_F}$ and the batch size $B \geq 9$, then we can have the following bound for $\mathcal{W}_2(\mu_T, \mu^*)$ for the SAGA-LD algorithm:*

$$\begin{aligned}
\mathcal{W}_2(\mu_T, \mu^*) \leq & 5\exp(-\frac{m_F h}{4}T)\mathcal{W}_2(\mu_0, \mu^*) + \\
& \frac{2hD_F d}{m_F} + \frac{2hL_F{}^{\frac{3}{2}}\sqrt{d}}{m_F} + \frac{24hL_F\sqrt{dN}}{\sqrt{m_F}B}
\end{aligned}$$

[2] This assumption is a little different from that in (Zou et al., 2018) since we adopt different definition of $F_j$

**Theorem 5** *Under Assumption 5, if we choose Option $I$ and set the step size $h < \frac{1}{8L_F}$, the batch size $B \geq 2$ and the epoch length $\tau \geq \frac{8}{m_F h}$, then we can have the following bound for all ($T$ mod $\tau$ =0) for the SVRG-LD algorithm:*

$$\begin{aligned}\mathcal{W}_2(\mu_T, \mu^*) \leq & \exp(-\frac{m_F h}{56}T)\frac{\sqrt{L_F}}{\sqrt{m_F}}\mathcal{W}_2(\mu_0, \mu^*) + \\ & \frac{2hD_F d}{m_F} + \frac{2hL_F{}^{\frac{3}{2}}\sqrt{d}}{m_F} + \frac{64L_F^{\frac{3}{2}}\sqrt{hd}}{m_F\sqrt{B}}\end{aligned}$$

*If we choose Option $II$ and set the step size $h < \frac{\sqrt{B}}{4\tau C_2}$, then we can have the following bound for all T for the SVRG-LD algorithm:*

$$\begin{aligned}\mathcal{W}_2(\mu_T, \mu^*) \leq & \exp(-\frac{m_F h}{4}T)\mathcal{W}_2(\mu_0, \mu^*) + \\ & \frac{\sqrt{2}hD_F d}{m_F} + \frac{5hL_F{}^{\frac{3}{2}}\sqrt{d}}{m_F} + \frac{9hL_F\tau\sqrt{d}}{\sqrt{Bm_F}}\end{aligned}$$

**Theorem 6** *Under Assumption 5 and Assumption 6, if we set the step size $h \leq min\{(\frac{BC_3}{24C_2{}^4\tau^2})^{\frac{1}{3}}, \frac{1}{6\tau(C_5{}^2/b+C_2)}\}$, we can have the following bound for all T for the SVRG-LD$^+$ algorithm:*

$$\begin{aligned}&\mathcal{W}_2(\mu_T, \mu^*) \leq (1 - hm_F/4)^T\mathcal{W}_2(\mu_0, \mu^*) + \\ &\frac{3\sigma d^{1/2}}{m_F b^{1/2}}\mathbf{1}(b \leq N) + \frac{2hD_4 d}{m_F} + \frac{2hL_F{}^{3/2}d^{1/2}}{m_F} \\ &+ \frac{4hL_F(\tau d)^{1/2} \wedge 3h^{1/2}d^{1/2}\sigma}{\sqrt{Bm_F}}\end{aligned}$$

## D. Proof of the theorems in Section 4

In this section, we prove the theorems in Section 4. We have simplified our proofs because we want to make it easier to understand. Our proof is based on the idea of (Zhang et al., 2020) and borrow some results from (Chatterji et al., 2018; Zou et al., 2018). We first have the following update equation for SPOS:

$$\begin{aligned}\mathrm{d}\boldsymbol{\theta}_t^{(i)} = & -\beta^{-1}F(\boldsymbol{\theta}_t^{(i)})\mathrm{d}t - \frac{1}{M}\sum_{q=1}^{M}K(\boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^{(q)})F(\boldsymbol{\theta}_t^{(q)})\mathrm{d}t \\ & + \frac{1}{M}\sum_{q=1}^{M}\nabla K(\boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^{(q)})\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_t^{(i)} \quad \forall i\end{aligned} \tag{14}$$

As mention in Section 2.3, we denote the distribution of $\boldsymbol{\theta}_t^{(i)}$ in Eq.(14) as $\nu_t$. From the proof of Theorem 5 in (Zhang et al., 2020), we can derive that

$$\mathcal{W}_2(\nu_\infty, \mu^*) \leq \frac{2(H_{\nabla K} + H_\theta)}{\sqrt{M}(\beta^{-1} - \frac{5}{2}H_\theta L_K - L_F - 2L_{\nabla K})} \tag{15}$$

In order to bound $\mathcal{W}_2(\mu_T, \mu^*)$, we need to bound $\mathcal{W}_2(\mu_T, \nu_\infty)$. We borrow the idea in (Zhang et al., 2020), by concatenating the particles at each time into a single vector representation, We define a new parameter at time $t$ as $\boldsymbol{\Theta}_t \triangleq [\boldsymbol{\theta}_t^{(1)}, \cdots, \boldsymbol{\theta}_t^{(M)}] \in \mathbb{R}^{Md}$. Consequently, $\boldsymbol{\Theta}_t$ is driven by the following linear SDE:

$$\mathrm{d}\boldsymbol{\theta}_t = -F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathcal{W}_t^{(Md)}\ , \tag{16}$$

where $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_t) \triangleq [\beta^{-1}F(\boldsymbol{\theta}_t^{(1)}) - \frac{1}{M}\sum_{q=1}^{M}\nabla K(\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(q)}) + \frac{1}{M}\sum_{q=1}^{M}K(\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(q)})F(\boldsymbol{\theta}_t^{(q)}), \cdots, \beta^{-1}F(\boldsymbol{\theta}_t^{(M)}) - \frac{1}{M}\sum_{q=1}^{M}\nabla K(\boldsymbol{\theta}_t^{(M)} - \boldsymbol{\theta}_t^{(q)}) + \frac{1}{M}\sum_{q=1}^{M}K(\boldsymbol{\theta}_t^{(M)} - \boldsymbol{\theta}_t^{(q)})F(\boldsymbol{\theta}_t^{(q)})]$ is a vector function $\mathbb{R}^{Md} \to \mathbb{R}^{Md}$, and $\mathcal{W}_t^{(Md)}$ is

Brownian motion of dimension $Md$.

Define the $F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_t) \triangleq [\beta^{-1}F_j(\boldsymbol{\theta}_t^{(1)}) - \frac{1}{MN}\sum_{q=1}^{M}\nabla K(\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(q)}) + \frac{1}{M}\sum_{q=1}^{M}K(\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(q)})F_j(\boldsymbol{\theta}_t^{(q)}), \cdots, \beta^{-1}F_j(\boldsymbol{\theta}_t^{(M)}) - \frac{1}{MN}\sum_{q=1}^{M}\nabla K(\boldsymbol{\theta}_t^{(M)} - \boldsymbol{\theta}_t^{(q)}) + \frac{1}{M}\sum_{q=1}^{M}K(\boldsymbol{\theta}_t^{(M)} - \boldsymbol{\theta}_t^{(q)})F_j(\boldsymbol{\theta}_t^{(q)})]$. We can find the $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_t)$ and $F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_t)$ defined above satisfy the following theorem.

**Theorem 7**

- *(Sum-decomposable) The $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$ is decomposable, i.e., $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \sum_{j=1}^{N}F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$.*
- *(Strong convexity) $F^{\boldsymbol{\Theta}}$ is a $(\beta^{-1}m_F - -2L_F - 3H_1L_K)$-strongly convex function, it i.e., $\left(F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\right)(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2) \leq (\beta^{-1}m_F - 2L_F - 3H_1L_K)\left\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\right\|$.*
- *(Hessian Lischitz) The function $F^{\boldsymbol{\Theta}}$ is Hessian Lipschitz, i.e., $\left\|\nabla F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - \nabla F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\right\| \leq (\beta^{-1}D_F + 4D_{\nabla^2 K} + 4H_2L_{\nabla K} + 2L_FH_{\nabla K} + 2H_1L_K + L_FH_K)\left\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\right\|$.*
- *(Smoothness) $F^{\boldsymbol{\Theta}}$ is Lipschitz continuous with some positive constant, i.e., for all $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathbb{R}^{Md}$,$\|F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\| \leq \sqrt{2(\beta^{-1}L_F + 2L_KH_1 + H_KL_F + L_{\nabla K})^2 + 2}\left\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\right\|$.*
- *(Bound Variance) There exits a constant, $\sigma \geq 0$, such that for all $j$,*

$$\mathbb{E}[\|F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) - \frac{1}{N}\sum_{j=1}^{N}F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})\|^2] \leq Md(2\beta^{-1} + 2H_K^2)\sigma^2/N^2$$

**Proof**

- The sum-decomposable property of $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$ is easy to verify.
- (Strong convexity)

$$(F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1) - F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2))(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2) = \frac{1}{M}\sum_{i,q}^{M}(\xi_{iq}^1 + \xi_{iq}^2 + \xi_{iq}^3 + \xi_{iq}^4) \tag{17}$$

where

$$\xi_{iq}^1 = \beta^{-1}\left(F(\boldsymbol{\Theta}_1^{(i)}) - F(\boldsymbol{\Theta}_2^{(i)})\right)\cdot\left(\boldsymbol{\Theta}_1^{(i)} - \boldsymbol{\Theta}_2^{(i)}\right)$$

$$\xi_{iq}^2 = -\left(\nabla K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(q)}) - \nabla K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(q)})\right)\cdot\left(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\right)$$

$$\xi_{iq}^3 = \left(F(\boldsymbol{\theta}_1^{(q)})K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(q)}) - F(\boldsymbol{\theta}_2^{(q)})K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(q)})\right)\cdot\left(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\right)$$

$$\xi_{iq}^4 = \left(F(\boldsymbol{\theta}_2^{(q)})K(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_1^{(q)}) - F(\boldsymbol{\theta}_2^{(q)})K(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\theta}_2^{(q)})\right)\cdot\left(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\right)$$

For the $\xi_{iq}^1$ terms, applying the convex condition for $F$, we have

$$\begin{aligned}\sum_{iq}\xi_{iq}^1 &= \sum_{iq}\beta^{-1}\left(F(\boldsymbol{\theta}_1^{(i)}) - F(\boldsymbol{\theta}_2^{(i)}\right)\cdot\left(\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\right)\\ &\geq \beta^{-1}m_FM\sum_{i}\left\|\boldsymbol{\theta}_1^{(i)} - \boldsymbol{\theta}_2^{(i)}\right\|^2\end{aligned} \tag{18}$$

For the $\xi_{iq}^2$ term, applying the concave condition for $K$ and $\nabla K$ is odd, we have

$$\sum_{iq}\xi_{iq}^2 = -\sum_{iq}^{M}\left(\nabla K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})-\nabla K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})\right)\cdot\left(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right)$$

$$\begin{aligned}
&=-\frac{1}{2}\sum_{iq}^{M}\sum_{iq}^{M}\left(\nabla K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})-\nabla K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})\right)\cdot\left(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)}-(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})\right)\\
&=\frac{1}{2}L_{\nabla K}\sum_{ij}\left\|\theta_\tau^{(i)}-\bar{\theta}_\tau^{(i)}-(\theta_\tau^{(j)}-\bar{\theta}_\tau^{(j)})\right\|^2\geq -2L_{\nabla K}M\mathbb{E}\sum_{i}\left\|\theta_\tau^{(i)}-\bar{\theta}_\tau^{(i)}\right\|^2
\end{aligned}\tag{19}$$

For the $\xi_{iq}^3$ terms, after applying the $L_F$-Lipschitz property of $F$, we have

$$\begin{aligned}
\sum_{iq}\xi_{iq}^3&=\sum_{iq}(F(\boldsymbol{\theta}_1^{(q)})K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})-F(\boldsymbol{\theta}_2^{(q)})K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)}))\cdot\left(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right)\\
&\geq-\sum_{iq}L_F\left\|\boldsymbol{\theta}_1^{(q)}-\boldsymbol{\theta}_2^{(q)}\right\|\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|\\
&\geq-2L_FM\sum_{i}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|^2
\end{aligned}\tag{20}$$

For the $\xi_{iq}^4$ terms, we have

$$\begin{aligned}
\sum_{iq}\xi_{iq}^4&=\sum_{iq}(F(\boldsymbol{\theta}_2^{(q)})K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})-F(\boldsymbol{\theta}_2^{(q)})K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)}))\cdot\left(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right)\\
&\geq-H_1L_K\sum_{iq}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)}-(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})\right\|\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|\\
&\geq-3H_1L_KM\sum_{i}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|^2
\end{aligned}\tag{21}$$

Combining these bounds, we arrive at:

$$\begin{aligned}
&\left(F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1)-F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\right)(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2)\\
&\geq(\beta^{-1}m_F-2L_{\nabla K}-2L_F-3H_1L_K)\sum_{i}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|\\
&\geq(\beta^{-1}m_F-2L_{\nabla K}-2L_F-3H_FL_K)\left\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\right\|
\end{aligned}\tag{22}$$

- Next, we will prove the third result:

$$\begin{aligned}
&\left\|\nabla F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1)-\nabla F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\right\|\\
&\leq\beta^{-1}\sum_{i=1}^{M}\left\|\nabla F(\boldsymbol{\theta}_1^{(i)})-\nabla F(\boldsymbol{\theta}_2^{(i)})\right\|+\sum_{i=1}^{M}\frac{2}{M}\sum_{q=1}^{M}\left\|\nabla^2K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})-\nabla^2K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})\right\|+\\
&\frac{2}{M}\sum_{i=1}^{M}\sum_{q=1}^{M}\|\nabla K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})F(\boldsymbol{\theta}_1^{(q)})-\nabla K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_2^{(q)})\|\\
&+\sum_{i=1}^{M}\sum_{q=1}^{M}\frac{1}{M}\|K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})\nabla F(\boldsymbol{\theta}_1^{(q)})-K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})\nabla F(\boldsymbol{\theta}_2^{(q)})\|\\
&\leq\sum_{i=1}^{M}\beta^{-1}D_F\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\|+4D_{\nabla^2K}\sum_{i=1}^{M}\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\|+
\end{aligned}$$

$$
\begin{aligned}
&\frac{2}{M}\sum_{i=1}^{M}\sum_{q=1}^{M}\|\nabla K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})F(\boldsymbol{\theta}_1^{(q)})-\nabla K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_1^{(q)})\|+\\
&\frac{2}{M}\sum_{i=1}^{M}\sum_{q=1}^{M}\|\nabla K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_1^{(q)})-\nabla K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_2^{(q)})\|\\
&+\frac{1}{M}\sum_{i=1}^{M}\sum_{q=1}^{M}\|K(\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_1^{(q)})F(\boldsymbol{\theta}_1^{(q)})-K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_1^{(q)})\|+\\
&\frac{1}{M}\sum_{i=1}^{M}\sum_{q=1}^{M}\|K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_1^{(q)})-K(\boldsymbol{\theta}_2^{(i)}-\boldsymbol{\theta}_2^{(q)})F(\boldsymbol{\theta}_2^{(q)})\|\\
&\le\sum_{i=1}^{M}\beta^{-1}D_F\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|+4D_{\nabla^2K}\sum_{i=1}^{M}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|+\\
&4\sum_{i=1}^{M}H_2L_{\nabla K}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|+2\sum_{i=1}^{M}L_FH_{\nabla K}\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|+\\
&2\sum_{i=1}^{M}H_1L_K\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|+\sum_{i=1}^{M}L_FH_K\left\|\boldsymbol{\theta}_1^{(i)}-\boldsymbol{\theta}_2^{(i)}\right\|\\
&\le(\beta^{-1}D_F+4D_{\nabla^2K}+4H_2L_{\nabla K}+2L_FH_{\nabla K}+2H_1L_K+L_FH_K)\,\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\| \qquad (23)
\end{aligned}
$$

Similarly, we can easily verify that

$$\|F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_1)-F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_2)\|\le\sqrt{2(\beta^{-1}L_F+2L_KH_1+H_KL_F+L_{\nabla K})^2+2}\,\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2\|$$

- Finally, we prove the last result.

$$
\begin{aligned}
&\mathbb{E}[\|F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})-\frac{1}{N}\sum_{j=1}^{N}F_j^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})\|^2]=\\
&\sum_{i=1}^{M}\mathbb{E}[\|\beta^{-1}F_j(\boldsymbol{\Theta}^{(i)})-\beta^{-1}\frac{1}{N}\sum_{j=1}^{N}F_j(\boldsymbol{\theta}^{(i)})+\frac{1}{M}\sum_{q=1}^{M}K(\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(q)})F_j(\boldsymbol{\theta}_t^{(q)})-\frac{1}{MN}\sum_{j=1}^{N}\sum_{q=1}^{M}K(\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(q)})F_j(\boldsymbol{\theta}_t^{(q)})\|^2]\\
&\le\sum_{i=1}^{M}[2\mathbb{E}\|\beta^{-1}F_j(\boldsymbol{\theta}^{(i)})-\beta^{-1}\frac{1}{N}\sum_{j=1}^{N}F_j(\boldsymbol{\theta}^{(i)})\|^2+2\frac{H_K^2}{M^2}\mathbb{E}\|\sum_{q=1}^{M}\left(F_j(\boldsymbol{\theta}^{(q)}-\frac{1}{N}\sum_{j=1}^{N}F_j(\boldsymbol{\theta}^{(q)})\right)\|^2]\\
&\le\sum_{i=1}^{M}(2d\sigma^2+2H_K^2d\sigma^2)/N^2\\
&\le Md(2\sigma^2+2H_K^2\sigma^2)/N^2 \qquad (24)
\end{aligned}
$$

We apply Euler-Maruyama discretization to Eq.(16) and substitute $G_k^{\Theta}$ for $F^{\Theta}(\Theta_k)$ to derive the following equation:

$$\Theta_{k+1}=\Theta_k-G_k^{\Theta}h+\sqrt{2\beta^{-1}h}\Xi_k,\ \ \Xi_k\sim\mathcal{N}(\mathbf{0},_{Md\times Md})$$

Hence, different $G_k^{\Theta}$ correspond to different algorithms for $\Theta_k$, *e.g.*, the SAGA-LD, SVRG-LD and SVRG-$LD^+$ algorithms. It is worth noting that the SAGA-LD, SVRG-LD and SVRG-LD$^+$ algorithms of $\Theta_k$ is actually the corresponding SAGA-POS, SVRG-POS and SVRG-POS$^+$ algorithm of $\{\theta_k^{(i)}\}$.

This result is very important for our proof, which bridges the gap between the variance reduction in stochastic gradient Langevin dynamics (SGLD) and variance reduction in stochastic particle-optimization sampling (SPOS). Thanks to the Theorem 7, we can can verify $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$ satisfies the Assumption 5 and Assumption 6 (please notice the $F^{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$ corresponds to the $\nabla F$ in (Chatterji et al., 2018)). Hence, we can borrow the results in (Chatterji et al., 2018; Zou et al., 2018) and derive some new results for the variance reduction techniques in stochastic particle-optimization sampling (SPOS).

We denotes the distribution of $\boldsymbol{\Theta}$ in Eq.(16) and the distribution of $\Theta_k$ in Eq.(25) as $\Gamma_t$ and and $\Lambda_k$. Now we can derive the following theorems, with $C_1$,$C_2$,$C_3$,$C_4$ and $C_5$ defined in Section 4.

**Theorem 8** *Let the step size $h < \frac{B}{8NC_1}$ and the batch size $B \geq 9$. We have the bound for $\mathcal{W}_2(\Lambda_T, \Gamma_\infty)$ in the SAGA-LD algorithm of $\Theta_k$ as:*

$$\begin{aligned}\mathcal{W}_2(\Lambda_T, \Gamma_\infty) \leq & 5\exp(-\frac{C_3 h}{4}T)\mathcal{W}_2(\Lambda_0, \Gamma_\infty) + \\ & \frac{2hC_4 Md}{C_3} + \frac{2hC_2{}^{\frac{3}{2}}\sqrt{Md}}{C_3} + \frac{24hC_2\sqrt{MdN}}{\sqrt{C_3}B}\end{aligned}$$

**Theorem 9** *If we choose Option $I$ and set the step size $h < \frac{1}{8C_2}$, the batch size $B \geq 2$ and the epoch length $\tau \geq \frac{8}{C_3 h}$, we can have the bound for all (T mod $\tau$ =0) in the SVRG-LD algorithm of $\Theta_k$ as:*

$$\begin{aligned}\mathcal{W}_2(\Lambda_T, \Gamma_\infty) \leq & \exp(-\frac{C_3 h}{56}T)\frac{\sqrt{C_2}}{\sqrt{C_3}}\mathcal{W}_2(\Lambda_0, \Gamma_\infty) + \\ & \frac{2hC_4 Md}{C_3} + \frac{2hC_2{}^{\frac{3}{2}}\sqrt{Md}}{C_3} + \frac{64C_2^{\frac{3}{2}}\sqrt{hMd}}{\sqrt{B}C_3}\end{aligned}$$

*If we choose Option $II$ and set the step size $h < \frac{\sqrt{B}}{4\tau C_2}$, we can have the bound for all T in the SVRG-LD algorithm of $\Theta_k$ as:*

$$\begin{aligned}\mathcal{W}_2(\Lambda_T, \Gamma_\infty) \leq & \exp(-\frac{C_3 h}{4}T)\mathcal{W}_2(\Lambda_0, \Gamma_\infty) + \\ & \frac{\sqrt{2}hC_4 Md}{C_3} + \frac{5hC_2{}^{\frac{3}{2}}\sqrt{Md}}{C_3} + \frac{9hC_2\tau\sqrt{Md}}{\sqrt{BC_3}}\end{aligned}$$

**Theorem 10** *If we set the step size $h \leq min\{(\frac{BC_3}{24C_2{}^4\tau^2})^{\frac{1}{3}}, \frac{1}{6\tau(C_5{}^2/b + C_2)}\}$, we can have the bound for all T in the algorithm SVRG-LD$^+$ of $\Theta_k$ as:*

$$\begin{aligned}&\mathcal{W}_2(\Lambda_T, \Gamma_\infty) \leq (1 - hC_2/4)^T\mathcal{W}_2(\mu_0, \mu^*) + \\ &\frac{3C_5(Md)^{1/2}}{C_3 b^{1/2}}\mathbf{1}(b \leq N) + \frac{2h(C_4 Md)}{C_3} + \frac{2hC_2{}^{3/2}(Md)^{1/2}}{C_3} \\ &+ \frac{4hC_2(\tau Md)^{1/2} \wedge 3h^{1/2}(Md)^{1/2}C_5}{\sqrt{BC_3}}\end{aligned}$$

Next, we derive a proposition, which will be useful to connecting the $\mathcal{W}_2(\Lambda_T, \Gamma_\infty)$ and $\mathcal{W}_2(\mu_T, \nu_\infty)$ mentioned above. For simplicity of notations, we directly use $\theta$ and $\Theta$ themselves to denote their own distributions.

**Proposition 11** *Define $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ as $\boldsymbol{\Theta}_1 \triangleq [\boldsymbol{\theta}_1^{(1)}, \cdots, \boldsymbol{\theta}_1^{(M)}] \in \mathbb{R}^{Md}$ and $\boldsymbol{\theta}_2 \triangleq [\boldsymbol{\theta}_2^{(1)}, \cdots, \boldsymbol{\theta}_2^{(M)}] \in \mathbb{R}^{Md}$. We have*

$$\sum_{i=1}^{M}\mathcal{W}_2^2(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}) \leq \mathcal{W}_2^2(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{25}$$

**Proof** According to the Eq.(4.2) in (Soheil Feizi & Tse), we can write the $\mathcal{W}_2(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)})$ in the following optimizaition:

$$\begin{aligned}\mathcal{W}_2^2(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}) &= \mathbb{E}\|\boldsymbol{\theta}_1^{(i)}\|^2 + \mathbb{E}\|\boldsymbol{\theta}_2^{(i)}\|^2 \\ +2 &\sup_{\phi:convex}\{-\mathbb{E}[\phi(\boldsymbol{\theta}_1^{(i)})] - \mathbb{E}[\phi^*(\boldsymbol{\theta}_2^{(i)})]\}\end{aligned} \tag{26}$$

where $\phi^*(\theta) \triangleq \sup_v (v^T\theta - \phi(\theta))$ is the convex-conjugate of the function $\phi$. Assume $\phi_i$ is the optimal function of Eq.26. It is trivial to verify that $\Psi(\mathbf{\Theta}) \triangleq \sum_{i=1}^M \phi_i(\boldsymbol{\theta}^{(i)})$ is a convex function. Due to the property of conjugate functions, we have $\Psi(\mathbf{\Theta})^* = \sum_{i=1}^M \phi_i^*(\boldsymbol{\theta}^{(i)})$. Now we can derive the following result:

$$\begin{aligned}\sum_{i=1}^M \mathcal{W}_2^2(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}) &= \sum_{i=1}^M \{\mathbb{E}\|\boldsymbol{\theta}_1^{(i)}\|^2 + \mathbb{E}\|\boldsymbol{\theta}_2^{(i)}\|^2 + 2(-\mathbb{E}[\phi_i(\boldsymbol{\theta}_1^{(i)})] - \mathbb{E}[\phi_i^*(\boldsymbol{\theta}_2^{(i)})])\} \\ &= \mathbb{E}\|\mathbf{\Theta}_1\|^2 + \mathbb{E}\|\mathbf{\Theta}_2\|^2 + 2(-\mathbb{E}[\Psi(\mathbf{\Theta}_1)] - \mathbb{E}[\Psi^*(\mathbf{\Theta}_2)]) \\ &\leq \mathcal{W}_2^2(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \ ,\end{aligned}$$

which finishes our proof.

We should notice that due to the exchangeability of the $M$-particles system $\{\theta_k^{(i)}\}$ in our SPOS-type sampling, the distribution of each particle $\theta_T^{(i)}$ at the same time is identical. Hence, using Proposition 11, we can derive

$$\mathcal{W}_2(\mu_T, \nu_\infty) \leq \frac{1}{\sqrt{M}} \mathcal{W}_2(\Lambda_T, \Gamma_\infty) \tag{27}$$

To further proceed, we need to make a mild assumption that $\mathcal{W}_2(\mu_T, \nu_\infty) \leq \frac{1}{M^{1/2+\alpha}} \mathcal{W}_2(\Lambda_T, \Gamma_\infty)$. We wish to make some comments on the additional assumption. This assumption is reasonable. With this assumption, we can make the claim that the improvement of SVRG-POS over SVRG-LD is more significant than that of SAGA-POS over SAGA-LD, which is actually verified by our experiments, implying the reasonability of our assumption. Moreover, this assumption does not conflict with our result, because $\mathcal{W}_2(\mu_T, \nu_\infty) \leq \frac{1}{M^{1/(2+\alpha)}} \mathcal{W}_2(\Lambda_T, \Gamma_\infty) \leq \frac{1}{\sqrt{M}} \mathcal{W}_2(\Lambda_T, \Gamma_\infty)$. Furthermore, this assumption can be supported from a theoretical perspective. Consider the continuous function $\log_M \left( W_2(\mathbf{\Theta}_1, \mathbf{\Theta}_2) M / \sum_{i=1}^M W_2(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}) \right) - 1/2$. In a bounded space considered in practice, the above function is bounded from below. Since in practice we cannot use infinite particles, the required $\alpha$ does exist within the positive minima for every M mentioned above. Although we do not aim at giving an explicit expression for it, the existence is enough to explain the experiment results in our paper. Last, this assumption is supported in the algorithm itself. Please notice the fact that SPOS can be viewed as the combination of SVGD and SGLD. The SVGD part can constrain our algorithm to maintain some good properties that SGLD does not endow.

**Proof of Theorem 1, Theorem 2 and Theorem 3** Applying the results for $\mathcal{W}_2(\Lambda_T, \Gamma_\infty)$ in Theorem 8, Theorem 9 and Theorem 10, we can get the corresponding results for $\mathcal{W}_2(\mu_T, \nu_\infty)$ in the SAGA-POS, SVRG-POS and SVRG-POS$^+$. Then we can bound $\mathcal{W}_2(\mu_T, \mu^*)$,which is what we desire, with the following fact

$$\mathcal{W}_2(\mu_T, \mu^*) \leq \mathcal{W}_2(\mu_T, \nu_\infty) + \mathcal{W}_2(\nu_\infty, \mu^*) \tag{28}$$

Note that from the proof of Theorem 3 and Remark 1 in (Zhang et al., 2020), we can get that

$$\mathcal{W}_2(\nu_\infty, \mu^*) \leq \frac{C_1}{\sqrt{M}} \tag{29}$$

Apply the results in Theorem 8, Theorem 9 and Theorem 10 above, we can prove Theorem 1, Theorem 2 and Theorem 3.

## E. Comparison between SPOS and its Variance-Reduction Counterpart

In (Zhang et al., 2020), the authors use the distance $\tilde{\mathcal{B}}_T$ defined as $\tilde{\mathcal{B}}_T \triangleq \sup |\mathbb{E}_{\mu_T}[f(\boldsymbol{\theta})] - \mathbb{E}_{\mu^*}[f(\boldsymbol{\theta})]|$. When $\|f\|_{lip} \leq 1$, $\tilde{\mathcal{B}}_T$ is equivalent to $\mathcal{W}_1(\mu_T, \mu^*)$. According the proof in (Zhang et al., 2020), the authors did give a bound in terms of $\mathcal{W}_1(\mu_T, \mu^*)$. With the results in (Zhang et al., 2020), we can get the following theorem:

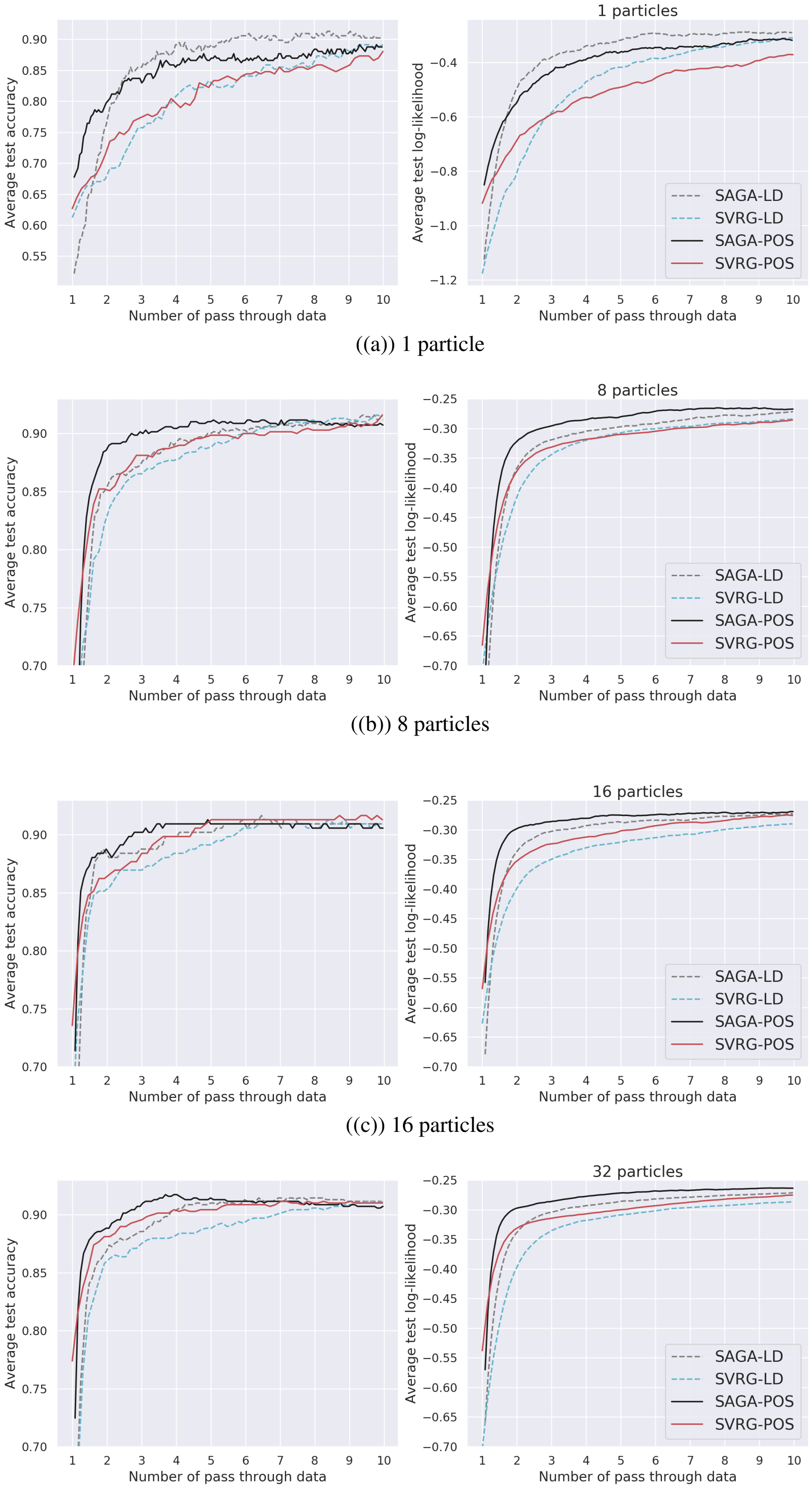
**Theorem 12 (Fixed Stepsize)** *Under Assumption 1, there exit some positive constants $(c_1, c_2, c_3, c_4, c_5, c_6)$ such that the bound for $\mathcal{W}_1(\mu_T, \mu^*)$ in the SPOS algorithm satisfies:*

$$\mathcal{W}_1(\mu_T, \mu^*) \leq \frac{c_1}{\sqrt{M}(\beta^{-1} - c_2)} + c_3 \exp\left\{-2\left(\beta^{-1} m_F - 2L_K - L_F\right) Th\right\} + c_6 M d^{\frac{3}{2}} \beta^{-3} (c_4 \beta^2 B^{-1} + c_5 h)^{\frac{1}{2}} T^{\frac{1}{2}} h^{\frac{1}{2}} \ .$$

Firstly, we should notice that the third term $c_3 M d^{\frac{3}{2}} \beta^{-3} (c_4 \beta^2 B^{-1} + c_5 h)^{\frac{1}{2}} T^{\frac{1}{2}} h^{\frac{1}{2}}$ on the right side increases with $T$ and $M$. However, the bound for SAGA-POS, SVRG-POS and SVRG-POS$^+$ in our paper decrease with both $T$ and $M$, which means that the bound for SAGA-POS, SVRG-POS and SVRG-POS$^+$ are tighter than the bound of SPOS. Furthermore, the convergence of SPOS is characterized in $\mathcal{W}_1(\mu_T, \mu^*)$. But the convergence of SAGA-POS, SVRG-POS and SVRG-POS$^+$ are characterized by $\mathcal{W}_2(\mu_T, \mu^*)$. Due to the well-known fact that $\mathcal{W}_1(\mu_T, \mu^*) \leq \mathcal{W}_1(\mu_T, \mu^*)$, we can verify that SAGA-POS, SVRG-POS and SVRG-POS$^+$ can outperform SPOS in theory. Although the result for SPOS in (Zhang et al., 2020) may be improved in the future, we believe that SAGA-POS, SVRG-POS and SVRG-POS$^+$ still can perform better, which has been verified in experiments in our paper.

## F. More Experiments Results

We further examine the impact of the number of particles to the convergence rates of variance-reduced SGLD and SPOS. As indicated by Theorems 1-3 (discussed in Remark 1 and 2), when the number of particles are large enough, the convergence rates of SAGA-POS and SVRG-POS would both outperform their SGLD counterparts. In addition, the performance gap would increase with increasing $M$, as indicated in Remark 4. We conduct experiments on the $Australian$ dataset by varying the particle numbers among $\{1, 8, 16, 32\}$. The results are plotted in Figure 5, which are roughly aligned with our theory.



((a)) 1 particle

((b)) 8 particles

((c)) 16 particles

((d)) 32 particles

*Figure 5.* Testing accuracy and log-likelihood vs the number of data pass for SPOS with varying number of particles.