A Additional Proof of Main Theories

In this section, we provide additional proofs of Corollaries 3.6 and 3.8 in Section 3.

Proof of Corollary 3.6. We first rewrite the upper bound of $\mathbb{E}[\|\Delta_k\|_2^2]$ as follows,

$$\mathcal{W}_2\big(P(\mathbf{x}_k),\pi\big) \leq (1-\eta\mu/4)^k \mathcal{W}_2(P(\mathbf{x}_0),\pi) + \frac{3\sigma d^{1/2}}{\mu B^{1/2}}\mathbb{1}(B<n) + \frac{2\eta(Ld+M^{3/2}d^{1/2})}{\mu} + \frac{4\eta M(md)^{1/2} \wedge 3\eta^{1/2}d^{1/2}\sigma}{(b\mu)^{1/2}}. \tag{A.1}$$

In order to achieve ϵ-accuracy, we assume that each term on the R.H.S of (A.1) is less than $\epsilon/4$. In order to perform a clean analysis, we will use big $\widetilde{O}$ notation to hide the absolute constants and logarithmic terms. In order to guarantee that the first term on the R.H.S of (A.1) satisfies the requirement, it suffices to set

$$k\eta \leq \frac{4\log\big(\mathcal{W}_2(P(\mathbf{x}_0),\pi)\big)}{\mu} = \widetilde{O}(1/\mu), \tag{A.2}$$

where we use the fact that $\mathcal{W}_2(P(\mathbf{x}_0),\pi)$ is polynomial in d (Cheng et al., 2017). Then by the second term on the R.H.S of (A.1), we can derive that $B = O\big(d\sigma^2/(\mu^2\epsilon^2)\wedge n\big)$. Similarly, from the third and fourth terms on the R.H.S of (A.1), we have

$$\eta = O\bigg(\min\bigg\{\frac{\mu\epsilon}{Ld+M^{3/2}d^{1/2}}, \frac{(b\mu)^{1/2}\epsilon}{M(md)^{1/2}} \vee \frac{b\mu\epsilon^2}{d\sigma^2}\bigg\}\bigg).$$

Note that the gradient complexity is

$$T \leq (k/m+1)\cdot B + kb = \frac{kB}{m} + kb + B,$$

which indicates the optimal choices of m and b should satisfy $mb = B$. Then the optimal step size is

$$\eta = O\bigg(\min\bigg\{\frac{\mu\epsilon}{Ld+M^{3/2}d^{1/2}}, \frac{b\mu^{1/2}\epsilon}{M(Bd)^{1/2}} \vee \frac{b\mu\epsilon^2}{d\sigma^2}\bigg\}\bigg).$$

Now we are ready to compute the gradient complexity, which is given as follows,

$$T \leq kb + B = k\eta b/\eta + B = \widetilde{O}\bigg(B + \frac{bLd + bM^{3/2}d^{1/2}}{\mu^2\epsilon} + \min\bigg\{\frac{M(Bd)^{1/2}}{\mu^{3/2}\epsilon}, \frac{d\sigma^2}{\mu^2\epsilon^2}\bigg\}\bigg).$$

It can be clearly observed that the optimal choice of b should be $b = O(1)$, which further implies that $m = O(B)$. Note that $B = O\big(d\sigma^2/(\mu^2\epsilon^2)\wedge n\big)$, we have the optimal step size

$$\eta = O\bigg(\min\bigg\{\frac{\mu\epsilon}{Ld+M^{3/2}d^{1/2}}, \max\bigg\{\frac{\mu^{3/2}\epsilon^2}{Md\sigma}, \frac{\mu^{1/2}\epsilon}{M(nd)^{1/2}}, \frac{\mu\epsilon^2}{d\sigma^2}\bigg\}\bigg\}\bigg),$$

and the gradient complexity

$$T = \widetilde{O}\bigg(\frac{d\sigma^2}{\mu^2\epsilon^2} \wedge n + \frac{Ld+M^{3/2}d^{1/2}}{\mu^2\epsilon} + \min\bigg\{\frac{d\sigma^2}{\mu^2\epsilon^2}, \frac{Md\sigma}{\mu^{5/2}\epsilon^2}, \frac{Md^{1/2}n^{1/2}}{\mu^{3/2}\epsilon}\bigg\}\bigg). \tag{A.3}$$

When assuming $n \gtrsim d\sigma^2/(\mu^2\epsilon^2)$, the above result can be directly simplified as

$$T = \widetilde{O}\bigg(\frac{d\sigma^2}{\mu^2\epsilon^2}\bigg),$$

which completes our proof. □

Proof of Corollary 3.8. We first recall the convergence result of SVRG-LD$^+$ as follows,

$$\mathcal{W}_2\big(P(\mathbf{x}_k),\pi\big) \leq (1-\eta\mu/4)^k \mathcal{W}_2(P(\mathbf{x}_0),\pi) + \frac{3\sigma d^{1/2}}{\mu B^{1/2}}\mathbb{1}(B<n) + \frac{2\eta(Ld+M^{3/2}d^{1/2})}{\mu} + \frac{4\eta M(md)^{1/2} \wedge 3\eta^{1/2}d^{1/2}\sigma}{(b\mu)^{1/2}}.$$

Note that SVRG-LD$^+$ reduces to SVRG-LD when $B = n$, thus the second term on the R.H.S vanishes. Moreover, we do not require Assumption 3.4, then the fourth term on the R.H.S of the above inequality becomes $4\eta M(md)^{1/2}/(b\mu)^{1/2}$. Then, we have the following convergence guarantee of SVRG-LD

$$\mathcal{W}_2\big(P(\mathbf{x}_k),\pi\big) \leq (1-\eta\mu/4)^k \mathcal{W}_2(P(\mathbf{x}_0),\pi) + \frac{2\eta(Ld+M^{3/2}d^{1/2})}{\mu} + \frac{4\eta M(md)^{1/2}}{(b\mu)^{1/2}}.$$

The next step is to prove the gradient complexity of SVRG-LD. Similarly, we follow the proof of Corollary 3.6. Note that we remove the assumption of bounded variance, and consider $B = n$. Under such considerations, the complexity for SVRG-LD$^+$, i.e., (A.3), reduces to

$$T = \widetilde{O}\bigg(n + \frac{Ld+M^{3/2}d^{1/2}}{\mu^2\epsilon} + \frac{Md^{1/2}n^{1/2}}{\mu^{3/2}\epsilon}\bigg),$$

and the step size η should be

$$\eta = O\bigg(\min\bigg\{\frac{\mu\epsilon}{Ld+M^{3/2}d^{1/2}}, \frac{\mu^{1/2}\epsilon}{M(nd)^{1/2}}\bigg\}\bigg).$$

□

B Proof of Technical Lemmas in Section 4

In this section, we prove the technical lemmas used in the proof of our main theory in Section 4. We first state the following necessary lemma which provides an upper bound of $\mathbb{E}[\|\nabla f(\mathbf{x}^\pi)\|]$, where $\mathbf{x}^\pi$ denotes the random variable satisfying the stationary distribution $\pi \propto e^{-f}$.

Lemma B.1. (Dalalyan, 2017) Under Assumption 3.1, consider that $\mathbf{x}^\pi$ satisfies the stationary distribution of (1.1), we have

$$\mathbb{E}[\|\nabla f(\mathbf{x}^\pi)\|_2^2] \leq Md,$$

where M is the smoothness parameter of $f(\mathbf{x})$.

Now we are going to prove Lemmas 4.2 and 4.3.

Proof of Lemma 4.2. Based on the definition of $\mathbf{\Psi}_k$ in (4.2), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{\Psi}_k\|_2^2] &= \mathbb{E}\big[\big\|\nabla f_{\widetilde{\mathcal{I}}_k}(\mathbf{x}_k) - \nabla f_{\widetilde{\mathcal{I}}_k}(\widetilde{\mathbf{x}}_j) - \big(\nabla f(\mathbf{x}_k) - \nabla f(\widetilde{\mathbf{x}}_j)\big)\big\|_2^2\big] \\ &\leq \frac{\mathbb{E}[\|\nabla f_i(\mathbf{x}_k) - \nabla f_i(\widetilde{\mathbf{x}}_j)\|_2^2]}{b} \\ &\leq \frac{M^2}{b}\mathbb{E}\big[\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2\big], \end{aligned} \tag{B.1}$$

where b denotes the minibatch size of $\widetilde{\mathcal{I}}_k$. According to the update form (2.1), we further have

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2] &= \mathbb{E}\bigg[\bigg\|\eta\sum_{i=jm}^{k}\mathbf{g}_i + \sqrt{2}\sum_{i=jm}^{k}\boldsymbol{\epsilon}_k\|_2^2\bigg\|\bigg] \\ &= \mathbb{E}\bigg[\eta^2\bigg\|\sum_{i=jm}^{k}\mathbf{g}_i\bigg\|_2^2 + 2ld\eta\bigg],\end{aligned}$$

where we denote $l = k - jm$, and the second inequality follows from that $\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_2^2] = \eta d$. Regarding to the first term on the R.H.S of the above equation, we have

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\sum_{i=jm}^{k}\mathbf{g}_i\Big\|_2^2\Big] &= \mathbb{E}\Big[\Big\|\sum_{i=jm}^{k}\nabla f_{\widetilde{\mathcal{I}}_i}(\mathbf{x}_i) - \nabla f_{\widetilde{\mathcal{I}}_i}(\widetilde{\mathbf{x}}_j) + \nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\Big\|_2^2\Big] \\
&\leq \mathbb{E}\Big[2\Big\|\sum_{i=jm}^{k}\nabla f_{\widetilde{\mathcal{I}}_i}(\mathbf{x}_i) - \nabla f_{\widetilde{\mathcal{I}}_i}(\widetilde{\mathbf{x}}_j)\Big\|_2^2 + 2\Big\|\sum_{i=jm}^{k}\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\|_2^2\Big] \\
&\leq \mathbb{E}\Big[2lM^2\sum_{i=jm}^{k}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_j\|_2^2 + 2l^2\|\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\|_2^2\Big].
\end{aligned}$$

By combining the above results, we obtain

$$\mathbb{E}[\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2] \leq 2lM^2\eta^2\sum_{i=jm}^{k}\mathbb{E}[\|\mathbf{x}_i - \widetilde{\mathbf{x}}_j\|_2^2] + 2l^2\eta^2\mathbb{E}[\|\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\|_2^2] + 2ld\eta.$$

Note that $l \leq m$, and using discrete Grönwall lemma (Clark, 1987), we have

$$\mathbb{E}[\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2] \leq \big(2m^2\eta^2\mathbb{E}[\|\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\|_2^2] + 2ld\eta\big)e^{2m^2M^2\eta^2}. \tag{B.2}$$

Regarding to $\mathbb{E}[\|\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\|_2^2]$, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j)\|_2^2] &\leq 3\mathbb{E}[\|\nabla f_{\mathcal{I}_j}(\widetilde{\mathbf{x}}_j) - \nabla f(\widetilde{\mathbf{x}}_j)\|_2^2] + 3\mathbb{E}[\|\nabla f(\widetilde{\mathbf{x}}_j) - \nabla f(\mathcal{L}_{jm\eta}\mathbf{x}^\pi)\|_2^2] + 3\mathbb{E}[\|\nabla f(\mathcal{L}_{jm\eta}\mathbf{x}^\pi)\|_2^2] \\
&\leq 3\|e_j\|_2^2 + 3M^2\Delta_{jm} + 3Md,
\end{aligned}$$

where the last inequality follows from Lemma B.1 and Assumption 3.1. Moreover, by plugging the above result into (B.2) and (B.1), we obtain

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\Psi}_k\|_2^2] &\leq \frac{M^2}{b}\mathbb{E}[\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2] \\
&\leq \frac{M^2}{b}\Big(6m^2\eta^2\big(\mathbb{E}[\|\boldsymbol{e}_j\|_2^2] + M^2\mathbb{E}[\|\Delta_{jm}\|_2^2] + Md\big) + 2ld\eta\Big)e^{2m^2M^2\eta^2}.
\end{aligned}$$

On the other hand, we are able to derive a different upper bound of $\mathbb{E}[\|\boldsymbol{\Psi}_k\|_2^2]$ in the other way. Note that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\Psi}_k\|_2^2] &= \mathbb{E}\big[\big\|\nabla f_{\widetilde{\mathcal{I}}_k}(\mathbf{x}_k) - \nabla f_{\widetilde{\mathcal{I}}_k}(\widetilde{\mathbf{x}}_j) - \big(\nabla f(\mathbf{x}_k) - \nabla f(\widetilde{\mathbf{x}}_j)\big)\big\|_2^2\big] \\
&\leq 2\mathbb{E}[\|\nabla f_{\widetilde{\mathcal{I}}_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|_2^2] + 2\mathbb{E}[\|\nabla f_{\widetilde{\mathcal{I}}_k}(\widetilde{\mathbf{x}}_j) - \nabla f(\widetilde{\mathbf{x}}_j)\|] \\
&\leq \frac{4d\sigma^2}{b},
\end{aligned}$$

where the last inequality follows from Lemma 4.3, which is also applicable for $\mathbf{x}_k$ and $\widetilde{\mathcal{I}}_k$. Finally, combining the above two upper bounds on $\mathbb{E}[\|\boldsymbol{\Psi}_k\|_2^2]$, we are able to complete the proof. □

Proof of Lemma 4.3. Note that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}_j\|_2^2] &= \mathbb{E}[\|\nabla f_{\mathcal{I}_j}(\mathbf{x}_0^{(j)}) - \nabla f(\mathbf{x}_0^{(j)})\|_2^2] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{i\in\mathcal{I}_j}\nabla f_i(\mathbf{x}_0^{(j)}) - \nabla f(\mathbf{x}_0^{(j)})\Big\|_2^2\Big].
\end{aligned}$$

Let $\mathbf{u}_i = \nabla f_i(\mathbf{x}_0^{(j)}) - \nabla f(\mathbf{x}_0^{(j)})$, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}_j\|_2^2] &= \mathbb{E}\Big[\Big\|\frac{1}{B}\mathbf{u}_i\Big\|_2^2\Big] \\ &= \frac{1}{B^2}\mathbb{E}\Big[\sum_{i\neq i'\in\mathcal{I}_j}\mathbf{u}_i^\top\mathbf{u}_{i'} + \sum_{i\in\mathcal{I}_j}\|\mathbf{u}_i\|_2^2\Big] \\ &= \frac{B-1}{Bn(n-1)}\sum_{i\neq i'}\mathbf{u}_i^\top\mathbf{u}_{i'} + \frac{d\sigma^2}{B}.\end{aligned}$$

Note that $\mathbb{E}\big[\sum_{i,i'}\mathbf{u}_i^\top\mathbf{u}_i'\big] = \mathbb{E}\big[\big(\sum_i\mathbf{u}_i\big)^\top\sum_i\mathbf{u}_i\big] = 0$, thus

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}_j\|_2^2] &= \frac{B-1}{Bn(n-1)}\sum_{i,i'}\mathbf{u}_i^\top\mathbf{u}_{i'} - \frac{B-1}{Bn(n-1)}\sum_{i=i'}\mathbf{u}_i^\top\mathbf{u}_{i'} + \frac{d\sigma^2}{B} \\ &= \frac{(n-B)d\sigma^2}{B(n-1)} \\ &\leq \frac{\mathbb{1}(B<n)d\sigma^2}{B}.\end{aligned}$$

□

C Additional Experiments

In this additional experiment, we would like to study the empirical performance of SVRG-LD$^+$ for sampling from non-log-concave distributions, even though our theoretical results are only applicable to strongly log-concave distributions. We apply the SVRG-LD$^+$ algorithm to a Bayesian independent component analysis (ICA) model, and compare its performance with SGLD, SAGA-LD and SVRG-LD. In the ICA model, we observe n examples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\ldots,n}$. Following Welling & Teh (2011); Dubey et al. (2016), we have the following probability of observing example $\mathbf{x}_i$ given the model matrix $\mathbf{W}$,

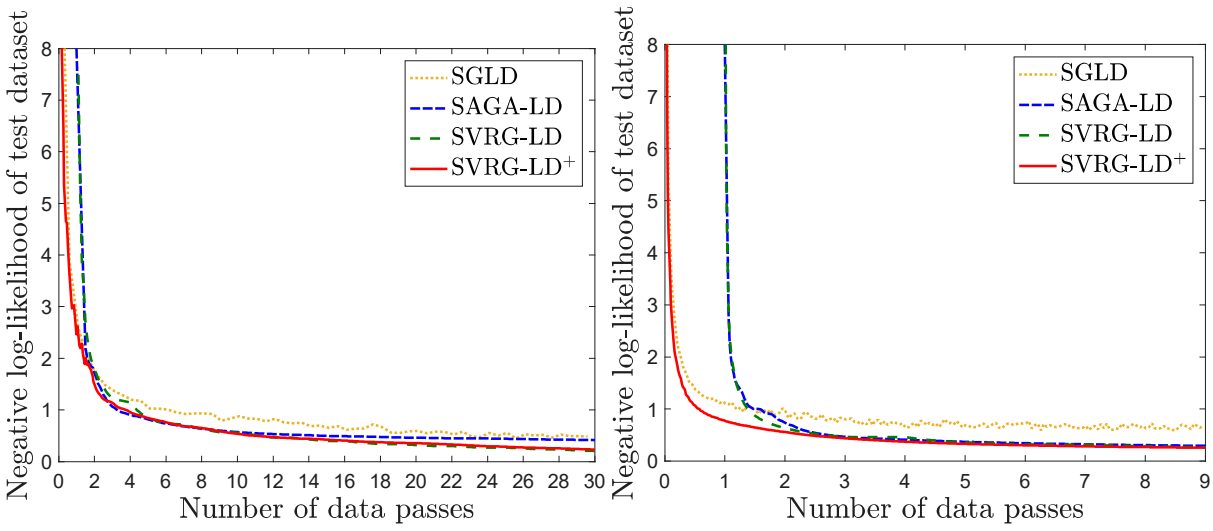
$$p(\mathbf{x}_i|\boldsymbol{W}) = |\det(\mathbf{W})|\prod_i p(\mathbf{w}_i^\top\mathbf{x}_i),$$

where $p(\mathbf{w}_i^\top\mathbf{x}_i) = 1/(4\cosh^2(\mathbf{w}_i^\top\mathbf{x}_i/2))$. We consider Gaussian prior over $\mathbf{W}$, i.e., $p(\mathbf{W}) \sim \mathcal{N}(0, \lambda^{-1}\mathbf{I})$. Then we can write the log-posterior as the average of n component functions, i.e., $\sum_{i=1}^n f_i(\mathbf{W})/n$, where

$$f_i(\mathbf{W}) = -n\big[\log(|\det(\mathbf{W})|) + 2\sum_{i=1}^d\log\big(\cosh(\mathbf{w}_i^\top\mathbf{x}_i/2)\big)\big] + \lambda\|\mathbf{W}\|_F^2.$$

It is worth noting that the target distribution is no longer log-concave. We use EEG dataset[6] to perform the ICA algorithm, which contains 125337 examples with 34 channels. In this experiment, we consider two regimes with different sample size n. To achieve this, we randomly extract two subsets with size 1000 and 10000 from the original dataset for training, and randomly extract 10000 examples from the rest of the dataset for test. We again compare the performance of SVRG-LD$^+$ with three baseline algorithms: SGLD, SAGA-LD and SVRG-LD, where the mini-batch size b is set to be 1, and the batch size B in SVRG-LD$^+$ is chosen to obtain the fastest convergence rate. Similar to Bayesian logistic regression, We plot the negative log-likelihood on the test dataset with respect to the number of effective data passes in Figure 5(a)-5(b). It should be noted that in the first epoch, SVRG-LD and SAGA-LD compute the full gradient using all the n examples, thus the curves of these two algorithms should start from the first data pass instead of the origin. It can be clearly observed that SVRG-LD$^+$ converges faster than the baseline algorithms. This observation reveals the potential of SVRG-LD$^+$ for sampling from non-log-concave distributions.

[6] https://mmspg.epfl.ch/cms/page-58322.html



(a) $n = 1000$ (b) $n = 10000$

Figure 5: Comparison of different algorithms for Bayesian independent component analysis, where y axis shows the negative log-likelihood on the test data, and x axis is the number of data passes. (a) sample size $n = 1000$; (b) sample size $n = 10000$.