Efficient Bayesian Model Selection in PARAFAC via Stochastic Thermodynamic Integration

SUPPLEMENTARY DOCUMENT

Thanh Huy Nguyen, *Student Member, IEEE,* Umut Şimşekli, *Member, IEEE,* Gaël Richard, *Fellow, IEEE*
Ali Taylan Cemgil, *Member, IEEE*

I. Details on the assumption

The update equation for PSGLD from the main text is written as follows.

$$\theta^{(t,l)} = \theta^{(t,l-1)} + \epsilon^{(t,l)}\Big(\mathbf{G}(\theta^{(t,l-1)})\Big(NN_s^{-1}t\sum_{n\in S^{(t,l)}}\nabla_\theta \log p(x_n|\theta^{(t,l-1)}) + \nabla_\theta \log p(\theta^{(t,l-1)})\Big)\Big) + \mathbf{G}^{\frac{1}{2}}(\theta^{(t,l-1)})\eta^{(t,l)}. \quad (1)$$

From [1], we propose to use the preconditioned matrix $\mathbf{G}(\theta)$ defined as follows

$$\mathbf{G}(\theta^{(t,l)}) = \mathrm{diag}\Big(\mathbf{1} \oslash \big(\lambda\mathbf{1} + \sqrt{\mathbf{v}(\theta^{(t,l)})}\big)\Big), \quad (2)$$

where

$$\mathbf{v}(\theta^{(t,l)}) = \alpha\mathbf{v}(\theta^{(t,l-1)}) + (1-\alpha)\bar{\mathbf{g}}(\theta^{(t,l-1)}; S^{(t,l)}) \odot \bar{\mathbf{g}}(\theta^{(t,l-1)}; S^{(t,l)}). \quad (3)$$

In (3), $\alpha \in [0,1]$ and

$$\bar{\mathbf{g}}(\theta^{(t,l-1)}; S^{(t,l)}) = N_s^{-1}t\sum_{n\in S^{(t,l)}}\nabla_\theta \log p(x_n|\theta^{(t,l-1)}).$$

Operators $\odot$ and $\oslash$ denote element-wise product and division, respectively. For simplicity, we will ignore the temperature t in the notation for θ. Now, we introduce the local generator of PSGLD (corresponding to (1))

$$\tilde{\mathcal{L}}_l = \Big(\mathbf{G}(\theta_l)\Big(NN_s^{-1}t\sum_{n\in S^{(t,l)}}\nabla_\theta \log p(x_n|\theta_l) + \nabla_\theta \log p(\theta_l)\Big)\Big)\cdot\nabla_\theta + \frac{1}{2}\mathbf{G}(\theta_l)\big(\mathbf{G}(\theta_l)^\top\big) : \nabla_\theta\nabla_\theta^\top, \quad (4)$$

where $\mathbf{a}\cdot\mathbf{b}$ denotes vector inner product between $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{A}:\mathbf{B}$ by definition is equal to $tr\{\mathbf{A}^\top\mathbf{B}\}$ for some matrices $\mathbf{A}$ and $\mathbf{B}$. Next, let $\mathcal{L}_k$ be the generator of the PSGLD with full gradient, i.e,

$$\mathcal{L}_l = \Big(\mathbf{G}(\theta_l)\Big(t\sum_{n\in S^{(t,l)}}\nabla_\theta \log p(x_n|\theta_l) + \nabla_\theta \log p(\theta_l)\Big) + \Gamma(\theta_l)\Big)\cdot\nabla_\theta + \frac{1}{2}\mathbf{G}(\theta_l)\big(\mathbf{G}(\theta_l)^\top\big) : \nabla_\theta\nabla_\theta^\top,$$

where $\Gamma_u(\theta) = \sum_v \partial\mathbf{G}_{uv}(\theta)/\partial\theta_v$. We have the relation

$$\tilde{\mathcal{L}}_l = \mathcal{L}_l + \Delta V^{(t,l)} + \Gamma(\theta_l)\cdot\nabla_\theta, \quad (5)$$

where $\Delta V^{(t,l)} = (N\bar{g}(\theta_l; S^{(l)}) - g(\theta_l))^\top\mathbf{G}(\theta_l)\nabla_\theta$ is an operator and $g(\theta_l) = t\sum_{n=1}^N \nabla_\theta \log p(x_n|\theta_l)$ is the full gradient.

Let $\bar{\phi} = \int \phi(\theta)p(\theta)\mathrm{d}\theta$ and $\hat{\phi} = \frac{1}{S_L}\sum_{l=1}^L \epsilon_l\phi(\theta_l)$, where $\phi(\theta)$ is a test function, $S_L = \sum_{l=1}^L \epsilon_l$ and $(\theta_l)_l$ is sampled according to (1), (2), (3) (by fixing t and replacing $\theta^{(t,l)}$ and $\theta^{(t,l-1)}$ by θ_l and θ_{l-1}, respectively).

In the rest of this section, we will denote $\langle\cdot\rangle$ expectation with respect to θ. We suppose that the two following assumptions are satisfied.

A 1. *The step-size $\{\epsilon_l\}$ are decreasing, i.e. $0 < \epsilon_{l+1} < \epsilon_l$, with 1) $\sum_{l=1}^\infty \epsilon_l = \infty$; and 2) $\sum_{l=1}^\infty \epsilon_k^2 < \infty$.*

A 2. *For each l there exists a functional ψ such that the Poisson equation $\mathcal{L}_l\psi(\theta_l) = \phi(\theta_l) - \bar{\phi}$ is satisfied, and there exists a function $\mathcal{V}$ such that $\|\mathcal{D}^d\psi\| \le C_d\mathcal{V}^{p_d}$ for $d \in \{0,1,2,3\}$, $C_d, p_d > 0$. In addition, $\sup_l\langle\mathcal{V}^p(\theta_k)\rangle < \infty$, and $\mathcal{V}$ is smooth such that $\sup_{s\in(0,1)}\mathcal{V}^p(s\theta + (1-s)Y) \le C(\mathcal{V}^p(\theta) + \mathcal{V}^p(Y))$ $\forall\theta, Y, p \le \max\{2p_d\}$ and for some $C > 0$.*

 T. H. Nguyen, U. Şimşekli and Gaël Richard are with the Image, Data, Signal department, Télécom ParisTech, Paris 75013, France. A. T. Cemgil is with the Department of Computer Engineering, Boğaziçi University, Bebek, Istanbul, 34342, Turkey. E-mail: {thanh.nguyen, umut.simsekli, gael.richard}@telecom-paristech.fr, taylan.cemgil@boun.edu.tr

Manuscript received November 17, 2017.

II. The proof for main theorem

With assumptions A 1 and A 2, we have the following lemma.

Lemma 1. *Let $\bar{\phi} = \int \phi(\theta)p(\theta)d\theta$ and $\hat{\phi} = \frac{1}{S_L}\sum_{l=1}^{L}\epsilon_l\phi(\theta_l)$, where $\phi(\theta)$ is a test function, $S_L = \sum_{l=1}^{L}\epsilon_l$ and $(\theta_l)_l$ is sampled according to* (1), (2), (3) *(by fixing t and replacing $\theta^{(t,l)}$ and $\theta^{(t,l-1)}$ by θ_l and θ_{l-1}, respectively). Then for some $C > 0$*

$$\left|\langle\hat{\phi} - \bar{\phi}\rangle\right| \leq C\left(\frac{1}{S_L} + \sum_{l=1}^{L}\frac{\epsilon_l}{S_L}\langle\|\Delta V^{(t,l)}\|\rangle + \frac{1}{S_L}\sum_{l=1}^{L}\epsilon_l^2 + \frac{1-\alpha}{\alpha^{3/2}}\right).$$

Proof We will follow the proof techniques of Theorem 1 in [1].

From the assumptions, there exists a functional ψ that satisfies the following Poisson equation

$$\mathcal{L}_l\psi(\theta_l) = \phi(\theta_l) - \bar{\phi}. \tag{6}$$

and the expectation of $\psi(\theta_l)$ can be decomposed as follows

$$\begin{aligned}\langle\psi(\theta_l)\rangle &= e^{\epsilon_l\tilde{\mathcal{L}}_l}\psi(\theta_{l-1}) + O(\epsilon_l^2)\\ &= (\mathbb{I} + \epsilon_l\tilde{\mathcal{L}}_l)\psi(\theta_{l-1}) + O(\epsilon_l^2),\end{aligned} \tag{7}$$

where $\mathbb{I}$ is the identity map. Sum over $l = 1, \ldots, L$ both sides of (7) and use (5), we obtain

$$\sum_{l=1}^{L}\langle\psi(\theta_l)\rangle = \sum_{l=1}^{L}\psi(\theta_{l-1}) + \sum_{l=1}^{L}\epsilon_l\mathcal{L}_l\psi(\theta_{l-1}) + \sum_{l=1}^{L}\epsilon_l\Delta V^{(t,l)}\psi(\theta_{l-1}) + \sum_{l=1}^{L}\epsilon_l\Gamma(\theta_l)\cdot\nabla_\theta\psi(\theta_{l-1}) + C\sum_{l=1}^{L}\epsilon_l^2.$$

Divide both sides by S_L then use (6), we get

$$\hat{\phi} - \bar{\phi} = \frac{\langle\psi(\theta_L)\rangle - \psi(\theta_0)}{S_L} + \frac{1}{S_L}\sum_{l=1}^{L-1}(\langle\psi(\theta_l)\rangle - \psi(\theta_l)) - \sum_{l=1}^{L}\frac{\epsilon_l}{S_L}\Delta V^{(t,l)}\psi(\theta_{l-1}) - \sum_{l=1}^{L}\frac{\epsilon_l}{S_L}\Gamma(\theta_l)\cdot\nabla_\theta\psi(\theta_{l-1}) - C\frac{\sum_{l=1}^{L}\epsilon_l^2}{S_L}. \tag{8}$$

We note that the term $|\langle\psi(\theta_L)\rangle - \psi(\theta_0)|$ is bounded by the assumptions. By Lemma 4 of [1] and assumption A 2, we also have

$$\left|\sum_{l=1}^{L}\frac{\epsilon_l}{S_L}\Gamma(\theta_l)\cdot\nabla_\theta\psi(\theta_{l-1})\right| = O\left(\frac{1-\alpha}{\alpha^{3/2}}\right).$$

Further, the expectation of the second term of the right side of (8) is equal to 0. Hence, it implies from (8) that there exists some constant $C > 0$ such that

$$\left|\langle\hat{\phi} - \bar{\phi}\rangle\right| \leq C\left(\frac{1}{S_L} + \sum_{l=1}^{L}\frac{\epsilon_l}{S_L}\langle\|\Delta V^{(t,l)}\|\rangle + \frac{1}{S_L}\sum_{l=1}^{L}\epsilon_l^2 + \frac{1-\alpha}{\alpha^{3/2}}\right).$$

□

We recall that, for STI, we evaluate $\log p(x)$ via the following identification:

$$\log p(x) = \int_0^1 \langle\log p(x|\theta)\rangle_{p(\theta|t)}\mathrm{d}t \tag{9}$$

where $\langle\log p(x|\theta)\rangle_{p(\theta|t)}$ denotes the expectation of $\log p(x|\theta)$ under $p(\theta|t)$, $p(\theta|t) = (1/z(t))p(\theta)p(x|\theta)^t$ with $z(t) = \int p(\theta)p(x|\theta)^t\mathrm{d}\theta$.

The first step of the evaluation is to estimate the expectation under the integral sign using SG-MCMC:

$$\langle\log p(x|\theta)\rangle_{p(\theta|t)} \approx \frac{1}{L}\frac{N}{N_s}\sum_{l=1}^{L}\sum_{n\in S^{(t,l)}}\log p(x_n|\theta^{(t,l)}) \tag{10}$$

where $\theta^{(t,l)}$ denotes samples drawn from $p(\theta|t)$, $S^{(t,l)}$ denotes random subsets of $\{1, 2, \ldots, N\}$ and N_s is the size of each $S^{(t,l)}$. Then a trapezoidal rule is used for numerically approximating the integration over t:

$$\log p(x) \approx \sum_{i=0}^{T-1}\Delta t_i\frac{\langle\log p(x|\theta)\rangle_{p(\theta|t_i)} + \langle\log p(x|\theta)\rangle_{p(\theta|t_{i+1})}}{2} \tag{11}$$

where $0 = t_0 < t_1 < \ldots < t_T = 1$ and $\Delta t_i = t_{i+1} - t_i$. The proof for the main theorem is as follows.

Theorem 1. *Let $\mathcal{L} = \int_0^1 f(t)dt$ be the log-marginal likelihood (Eq. (9)) with $f(t) = \langle \log p(x|\theta) \rangle_{p(\theta|t)}$ and $\hat{\mathcal{L}}$ be the estimator of $\mathcal{L}$ by STI (Eq. (10), (11)) using PSGLD as the sampling method for $\theta^{(t,l)}$ with constant stepsize ϵ. Assume that $\{x_n\}_{n=1}^N$ is i.i.d, $\log p(x|\theta)$ satisfies Assumption 2 in [1] (assumption A 2 above), $f(t)$ is twice differentiable and $|f''(t)| \leq U$ for $t \in [0,1]$ and for some $U > 0$, a uniform partition of $[0,1]$ is choosen, i.e. $\Delta t_i = 1/T$ for all $i = 0, \ldots, T-1$. Then we have for some constant $C > 0$*

$$\left| \langle \hat{\mathcal{L}} \rangle - \mathcal{L} \right| \leq C \left(\frac{1}{L\epsilon} + \max_t \left\{ \frac{1}{L} \sum_{l=1}^{L} \langle \| \Delta V^{(t,l)} \| \rangle \right\} + \epsilon + \frac{1}{T^2} + \frac{1-\alpha}{\alpha^{3/2}} \right).$$

Proof We will follow the steps of the proof of Theorem 1 in [2]. First, we specify the formula of the estimator $\hat{\mathcal{L}}$. By definition:

$$\hat{\mathcal{L}} = \sum_{i=0}^{T-1} \Delta t_i \frac{\hat{f}(t_i) + \hat{f}(t_{i+1})}{2},$$

where

$$\hat{f}(t) = \frac{1}{L} \sum_{l=1}^{L} \frac{N}{N_s} \sum_{n \in S^{(t,l)}} \log p(x_n | \theta^{(t,l)}).$$

Next, we define

$$\tilde{f}(t) = \frac{1}{L} \sum_{l=1}^{L} \sum_{n=1}^{N} \log p(x_n | \theta^{(t,l)}).$$

Now, we write the true log-marginal likelihood in the form

$$\mathcal{L} = \sum_{i=0}^{T-1} \int_{t_i}^{t_{i+1}} f(t) \mathrm{d}t.$$

By applying integration by parts, each integrals in the above sum can be written as

$$\int_{t_i}^{t_{i+1}} f(t) \mathrm{d}t = \Delta t \frac{f(t_i) + f(t_{i+1})}{2} + g(t_i),$$

where $\forall i \ \Delta t = \Delta t_i = 1/T$ and

$$g(t) = \int_0^{\Delta t} \left(\frac{(y - \Delta t/2)^2}{2} - \frac{(\Delta t)^2}{8} \right) f''(y+t) \mathrm{d}y.$$

If S is a random data subsample, we denote $\langle \cdot \rangle$ expectation with respect to θ and $\langle \langle \hat{f}(t) | S \rangle_{\theta|S} \rangle_S$ expectation with respect to S of expectation with respect to $\theta|S$ of $\hat{f}(t)$. Then, we have

$$
\begin{aligned}
\Big|\langle\hat{\mathcal{L}}\rangle - \mathcal{L}\Big| &= \Big|\langle\mathcal{L} - \hat{\mathcal{L}}\rangle\Big| \\
&= \Big|\langle\sum_{i=0}^{T-1}\Big(\Delta t\frac{f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})}{2} + g(t_i)\Big)\rangle\Big| \\
&= \Big|\langle\sum_{i=0}^{T-1}\Delta t\frac{f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})}{2}\rangle + \sum_{i=0}^{T-1} g(t_i)\Big| \\
&\leq \Big|\langle\sum_{i=0}^{T-1}\Delta t\frac{f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})}{2}\rangle\Big| + \Big|\sum_{i=0}^{T-1} g(t_i)\Big| \\
&= \frac{\Delta t}{2}\Big|\sum_{i=0}^{T-1}\langle f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})\rangle\Big| + \Big|\sum_{i=0}^{T-1} g(t_i)\Big| \\
&\leq \frac{\Delta t}{2}\sum_{i=0}^{T-1}\Big|\langle f(t_i)\rangle - \langle\hat{f}(t_i)\rangle\Big| + \frac{\Delta t}{2}\sum_{i=0}^{T-1}\Big|\langle f(t_{i+1})\rangle - \langle\hat{f}(t_{i+1})\rangle\Big| + \Big|\sum_{i=0}^{T-1} g(t_i)\Big| \\
&= \frac{\Delta t}{2}\sum_{i=0}^{T-1}\Big|\langle f(t_i)\rangle - \langle\langle\hat{f}(t_i)|S\rangle_{\theta|S}\rangle_S\Big| + \frac{\Delta t}{2}\sum_{i=0}^{T-1}\Big|\langle f(t_{i+1})\rangle - \langle\langle\hat{f}(t_{i+1})|S\rangle_{\theta|S}\rangle_S\Big| + \Big|\sum_{i=0}^{T-1} g(t_i)\Big| \\
&= \frac{\Delta t}{2}\sum_{i=0}^{T-1}\Big|\langle f(t_i)\rangle - \langle\tilde{f}(t_i)\rangle\Big| + \frac{\Delta t}{2}\sum_{i=0}^{T-1}\Big|\langle f(t_{i+1})\rangle - \langle\tilde{f}(t_{i+1})\rangle\Big| + \Big|\sum_{i=0}^{T-1} g(t_i)\Big|.
\end{aligned}
$$

By hypotheses, $f''(t)$ is uniformly bounded, hence we can bound the last term as follows

$$
\Big|\sum_{i=0}^{T-1} g(t_i)\Big| \leq \frac{C_g}{12T^2}.
$$

By applying Lemma 1 with $\epsilon_l = \epsilon$ for all $l = 1, \ldots, L$, we get

$$
\Big|\langle f(t_i)\rangle - \langle\tilde{f}(t_i)\rangle\Big| \leq C_i\left(\frac{1}{L\epsilon} + \frac{1}{L}\sum_{l=1}^{L}\langle\|\Delta V^{(t_i,l)}\|\rangle + \epsilon + \frac{1-\alpha}{\alpha^{3/2}}\right)
$$

for some $C_i > 0$. Thus,

$$
\begin{aligned}
\Big|\langle\hat{\mathcal{L}}\rangle - \mathcal{L}\Big| &\leq \frac{\Delta t}{2}\left(\frac{1}{L\epsilon} + \max_t\Big\{\frac{1}{L}\sum_{l=1}^{L}\langle\|\Delta V^{(t,l)}\|\rangle\Big\} + \epsilon + \frac{1-\alpha}{\alpha^{3/2}}\right)\sum_{i=0}^{T-1} C_i \\
&\quad + \frac{\Delta t}{2}\left(\frac{1}{L\epsilon} + \max_t\Big\{\frac{1}{L}\sum_{l=1}^{L}\langle\|\Delta V^{(t,l)}\|\rangle\Big\} + \epsilon + \frac{1-\alpha}{\alpha^{3/2}}\right)\sum_{i=0}^{T-1} C_{i+1} + \frac{C_g}{12T^2} \\
&\leq \Delta t\left(\frac{1}{L\epsilon} + \max_t\Big\{\frac{1}{L}\sum_{l=1}^{L}\langle\|\Delta V^{(t,l)}\|\rangle\Big\} + \epsilon + \frac{1-\alpha}{\alpha^{3/2}}\right)T\max_{0\leq i\leq T} C_i + \frac{C_g}{12T^2} \\
&= \left(\frac{1}{L\epsilon} + \max_t\Big\{\frac{1}{L}\sum_{l=1}^{L}\langle\|\Delta V^{(t,l)}\|\rangle\Big\} + \epsilon + \frac{1-\alpha}{\alpha^{3/2}}\right)\max_{0\leq i\leq T} C_i + \frac{C_g}{12T^2} \\
&\leq C\left(\frac{1}{L\epsilon} + \max_t\Big\{\frac{1}{L}\sum_{l=1}^{L}\langle\|\Delta V^{(t,l)}\|\rangle\Big\} + \epsilon + \frac{1}{T^2} + \frac{1-\alpha}{\alpha^{3/2}}\right),
\end{aligned}
$$

where $C = \max\{\max_{1\leq i\leq T} C_i, \frac{C_g}{12}\}$. □

III. Formal definitions of the blocks and the parts

Let us formally define the proposed procedure. We first denote the *partition* $\mathcal{P}_B(\mathcal{S})$ as a collection of B different non-empty disjoint subsets of a set $\mathcal{S}$, and the union of these subsets is equal to $\mathcal{S}$. Now, let us formally define a part and a block.

Definition 1. *A block $\Lambda \subset [I] \times [J] \times [K]$ is the Cartesian product of three sets which belong to $\mathcal{P}_B([I])$, $\mathcal{P}_B([J])$ and $\mathcal{P}_B([K])$, respectively.*

Definition 2. *A part $\Pi^{(l)} \subset [I] \times [J] \times [K]$, at iteration l, is a collection of mutually disjoint blocks and is defined as follows:*

$$\Pi^{(l)} = \cup_{b=1}^{B} \Lambda_b^{(l)} = \cup_{b=1}^{B} I_b^{(l)} \times J_b^{(l)} \times K_b^{(l)},$$

where $I_b^{(l)} \in \mathcal{P}_B([I])$, $J_b^{(l)} \in \mathcal{P}_B([J])$, $K_b^{(l)} \in \mathcal{P}_B([K])$ and $\forall b \neq b'$, $I_b^{(l)} \cap I_{b'}^{(l)} = \emptyset$, $J_b^{(l)} \cap J_{b'}^{(l)} = \emptyset$, $K_b^{(l)} \cap K_{b'}^{(l)} = \emptyset$.

In the parallelization scheme, the update equation of the STI with SGLD for the latent variable $\mathbf{A}$ is written as follows.

$$\mathbf{A}_b^{(t,l)} = \mathbf{A}_b^{(t,l-1)} + \epsilon^{(t,l)} \Big(N|\Pi^{(l)}|^{-1} t \sum_{(i,j,k)\in\Lambda_b^{(l)}} \nabla_{\mathbf{A}_b} \log p(x_{ijk}|\mathbf{A}_b^{(t,l-1)}, \mathbf{B}_b^{(t,l-1)}, \mathbf{C}_b^{(t,l-1)}) + \nabla_{\mathbf{A}_b} \log p(\mathbf{A}_b^{(t,l-1)}) \Big) + \eta_{\mathbf{A}_b}^{(t,l)}. \tag{12}$$

where

$$\begin{aligned} \mathbf{A}_b^{(t,l)} &\equiv \{a_{ir}^{(t,l)} | i \in I_b^{(t,l)}, r \in [R]\}, \\ \mathbf{B}_b^{(t,l)} &\equiv \{b_{jr}^{(t,l)} | j \in J_b^{(t,l)}, r \in [R]\}, \\ \mathbf{C}_b^{(t,l)} &\equiv \{c_{kr}^{(t,l)} | k \in K_b^{(t,l)}, r \in [R]\}. \end{aligned}$$

Using (12), we update $\mathbf{A}_b^{(t,l)}$, for $b = 1, \ldots, B$, in parallel. A similar process works for $\mathbf{B}$ and $\mathbf{C}$.

IV. Parallel implementation of the PSGLD: update rules

Formally, for STI with preconditioned SGLD, the latent variables $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are updated as follows. First, we rewrite the (1) for the update of $\mathbf{A}$ in the form:

$$\begin{aligned} \mathbf{A}_b^{(t,l)} =& \mathbf{A}_b^{(t,l-1)} + \epsilon^{(t,l)} \Big(\mathbf{G}(\mathbf{A}_b^{(t,l-1)}) \Big(N|\Pi^{(l)}|^{-1} t \sum_{(i,j,k)\in\Lambda_b^{(l)}} \nabla_{\mathbf{A}_b} \log p(x_{ijk}|\mathbf{A}_b^{(t,l-1)}, \mathbf{B}_b^{(t,l-1)}, \mathbf{C}_b^{(t,l-1)}) + \nabla_{\mathbf{A}_b} \log p(\mathbf{A}_b^{(t,l-1)}) \Big) \Big) \\ &+ \mathbf{G}^{\frac{1}{2}}(\mathbf{A}_b^{(t,l-1)}) \eta^{(t,l)}. \end{aligned}$$

where

$$\mathbf{G}(\mathbf{A}_b^{(t,l)}) = \mathrm{diag}\Big(\mathbf{1} \oslash \big(\lambda \mathbf{1} + \sqrt{\mathbf{v}(\mathbf{A}_b^{(t,l)})} \big) \Big),$$

$$\mathbf{v}(\mathbf{A}_b^{(t,l)}) = \alpha \mathbf{v}(\mathbf{A}_b^{(t,l-1)}) + (1-\alpha) \bar{\mathbf{g}}(\mathbf{A}_b^{(t,l-1)}, \Lambda_b^{(l)}) \odot \bar{\mathbf{g}}(\mathbf{A}_b^{(t,l-1)}, \Lambda_b^{(l)}),$$

$$\bar{\mathbf{g}}(\mathbf{A}_b^{(t,l-1)}, \Lambda_b^{(l)}) = |\Pi^{(l)}|^{-1} t \times \sum_{(i,j,k)\in\Lambda_b^{(l)}} \nabla_{\mathbf{A}_b} \log p(x_{ijk}|\mathbf{A}_b^{(t,l-1)}, \mathbf{B}_b^{(t,l-1)}, \mathbf{C}_b^{(t,l-1)}),$$

and

$$\begin{aligned} \mathbf{A}_b^{(t,l)} &\equiv \{a_{ir}^{(t,l)} | i \in I_b^{(t,l)}, r \in [R]\}, \\ \mathbf{B}_b^{(t,l)} &\equiv \{b_{jr}^{(t,l)} | j \in J_b^{(t,l)}, r \in [R]\}, \\ \mathbf{C}_b^{(t,l)} &\equiv \{c_{kr}^{(t,l)} | k \in K_b^{(t,l)}, r \in [R]\}. \end{aligned}$$

Then we update $\mathbf{A}_b^{(t,l)}$, for $b = 1, \ldots, B$, in parallel. A similar process works for $\mathbf{B}$ and $\mathbf{C}$.

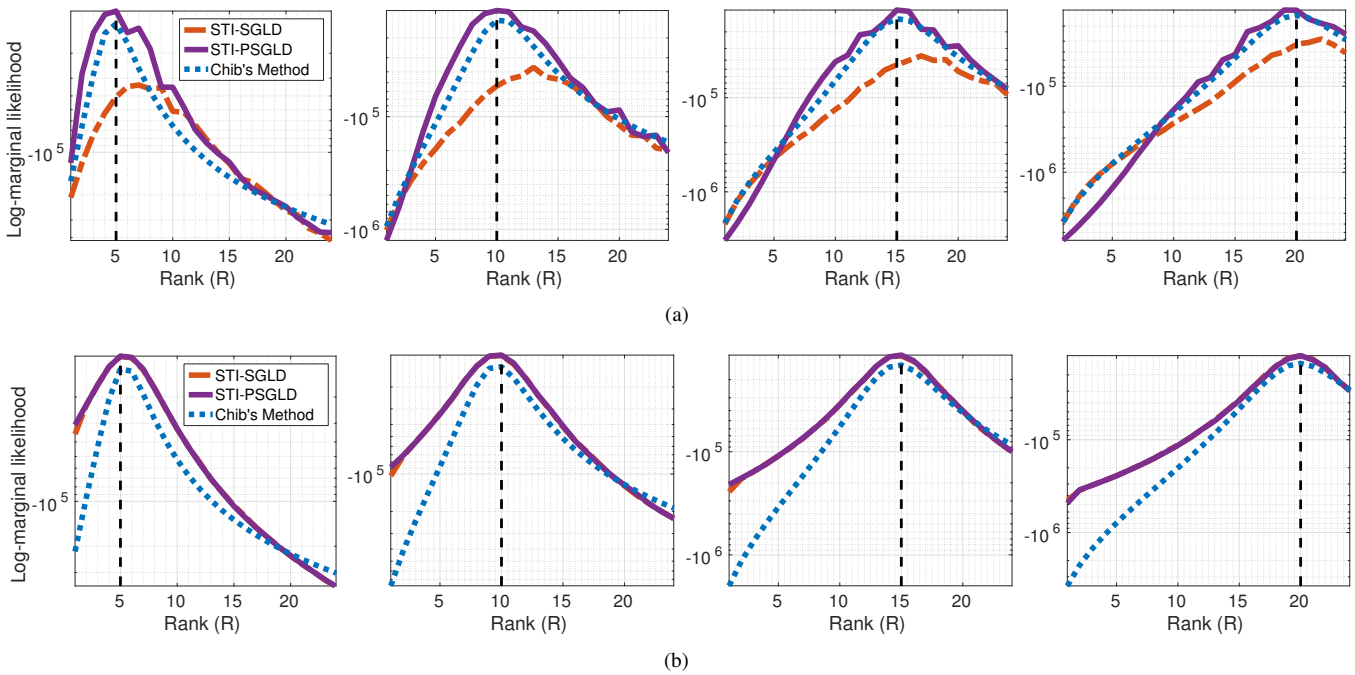

Fig. 1: Simulation results on a Gaussian model with 1(a) small number of iterations and 1(b) large number of iterations.

V. Additional Experiments on a Gaussian additive model

In this section, we use SGLD and PSGLD on a simple model in order to show the differences between the two methods. In this model, the log-marginal likelihood can be calculated explicitly, this is a good criterion to have the first perspectives of the two methods. The model is given as follows:

$$\theta_r \sim \mathcal{N}(\theta_r; \mu_\theta, \sigma_\theta^2), \qquad x_n|\theta \sim \mathcal{N}(x_n; \sum_{r=1}^{R} \theta_r, \sigma_x^2),$$

where $\{\theta_r\}_{r=1}^R$ is the set of the latent variables and $\{x_n\}_{n=1}^N$ is the set of the observations drawn from a Gaussian distribution whose mean is the sum of R i.i.d Gaussian latent variables θ_r. In this model, R is unknown a-priori. Our task is to estimate the marginal likelihood $p(x|R)$ of the data for M different values of R to determine which R is the most suitable for the model.

First, we set $\mu_\theta = 5$, $\sigma_\theta^2 = 3$, $\sigma_x^2 = 3$, $N = 5000$, and we choose $T = 10$, $N_s = 250$. For each run of STI (for PSGLD or SGLD as well), we generate $L = 20$ samples, i.e we go through the observed data just once, but we use only the last 10 samples for evaluating the log-likelihood. For the step-size, we choose $a_\epsilon = 10^{-8}$, $b_\epsilon = 0.51$, and keep the step-size fixed after first 10 generating samples. For PSGLD, we set $\alpha = 0.99$ and $\lambda = 10^{-5}$. The result is shown in Fig. 1(a).

After running the experiment for $R\_true = 5, 10, 15, 20$, Fig. 1(a) shows that STI-PSGLD predicts well the true value of R, and the values of the log-marginal likelihood predicted by STI-PSGLD fits very well their true values. For STI-SGLD, the estimated values for log-marginal likelihood is far from the true values and the prediction of the true rank R of the model is inexact. To understand why STI-PSGLD performs well while STI-SGLD has poor performance, we increase the number of iterations by setting $L = 3000$, use the last 1000 samples for evaluating the log-likelihood. The other parameters are kept unchanged. We obtain the result as in the Fig. 1(b).

In this experiment, both methods predict well the true value of R and the estimated values of the log-marginal likelihood of both methods are close to the true ones. In Fig. 1(b), the log-marginal likelihood curve of STI-pSLGD is almost unchanged compared with the previous experiment and this curve for STI-SGLD in fact coincides the one of STI-PSGLD when we use a large number of iterations. The STI-SGLD needs more iterations to converge since it does not have flexible step-sizes like STI-PSGLD. Hence if we use PSGLD, we can reduce the computational cost.

VI. Experiments on real data: computation time by STI-PSGLD

In this section, we analyze the computational time of STI-PSGLD within the experiments conducted on the Facebook dataset. Fig. 2 shows the duration of the STI-PSGLD run for each rank. We compare STI-PSGLD (using a Dell desktop with 3.2 GHz Quad-core Intel Xeon, 12 GB of memory) with CORCONDIA [3] (using a much more powerful computer with 1 TB of

memory): the computational times by using STI-PSGLD for $R = 3, R = 5$ and $R = 11$, for example, are $260, 430$ and 810 seconds, respectively, while these times for CORCONDIA are $390, 570$ and 1200 seconds, respectively. Apparently, our algorithm runs faster than CORCONDIA and it can be verified that the total time consumed by STI-PSGLD for this experiment is 30% less than the time consumed by CORCONDIA.

The computational cost of STI-PSGLD grows approximately linearly with rank R: the computational time is 103 seconds for $R = 1$, then increases gradually with average increment of 70 seconds between consecutive R's, and ends with 812 seconds for $R = 11$. This increment of CORCONDIA tends to increase faster when R is large, as can be seen from the computational times that are reported above.

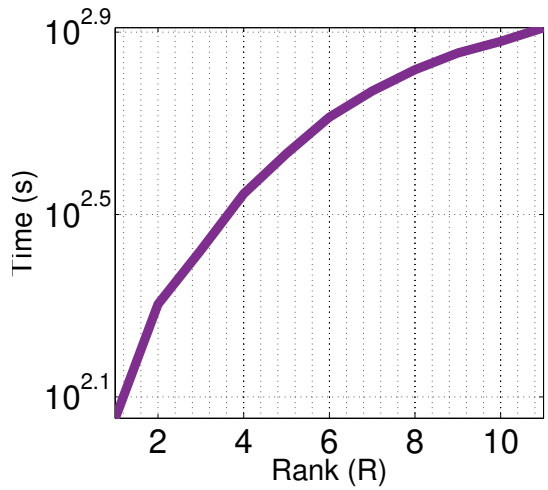

Fig. 2: Computational cost of PSGLD on Facebook dataset.

REFERENCES

[1] C. Li, C. Chen, D. E. Carlson, and L. Carin, "Preconditioned stochastic gradient Langevin dynamics for deep neural networks." in *AAAI*, vol. 2, no. 3, 2016, p. 4.

[2] U. Şimşekli, R. Badeau, G. Richard, and A. T. Cemgil, "Stochastic thermodynamic integration: efficient Bayesian model selection via stochastic gradient MCMC," in *ICASSP*. IEEE, 2016, pp. 2574–2578.

[3] E. E. Papalexakis and C. Faloutsos, "Fast efficient and scalable core consistency diagnostic for the parafac decomposition for big sparse tensors," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5441–5445.