# STOCHASTIC THERMODYNAMIC INTEGRATION: EFFICIENT BAYESIAN MODEL SELECTION VIA STOCHASTIC GRADIENT MCMC

## SUPPLEMENTARY DOCUMENT

*Umut Şimşekli*[1], *Roland Badeau*[1], *Gaël Richard*[1], *Ali Taylan Cemgil*[2]

1: LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
2: Dept. of Computer Engineering, Boğaziçi University, 34342, Bebek, İstanbul, Turkey

**ABSTRACT**

This is a supplementary document for the paper "Stochastic Thermodynamic Integration: Efficient Bayesian Model Selection via Stochastic Gradient MCMC", published in the ICASSP 2016 proceedings.

## 1. BOUNDING THE BIAS INDUCED BY STOCHASTIC THERMODYNAMIC INTEGRATION

In this section we show that the bias induced by STI can be bounded. Our proof construction is similar to the ones that are used for analyzing the error in conventional numerical integration methods, where we further make use of recent results presented in SG-MCMC literature.

Before proceeding to the theorem, we first define the following quantities in order to simplify the notation.

- We first define $\mathcal{L}$ to be the log-marginal likelihood, the quantity that is to be estimated:

$$\mathcal{L} \triangleq \log p(x) = \int_0^1 f(t)dt, \tag{1}$$

where

$$f(t) \triangleq \big\langle \log p(x|\theta) \big\rangle_{p(\theta|t)}. \tag{2}$$

- We define $\hat{\mathcal{L}}$ to be the estimator of the log-marginal likelihood that is obtained via STI:

$$\hat{\mathcal{L}} \triangleq \sum_{i=0}^{T-1} \Delta t_i \frac{\hat{f}(t_i) + \hat{f}(t_{i+1})}{2}, \tag{3}$$

where

$$\hat{f}(t) \triangleq \frac{1}{K} \sum_{k=1}^{K} \frac{N}{N_s} \sum_{n \in S^{(t,k)}} \log p(x_n|\theta^{(t,k)}). \tag{4}$$

- Finally, we define

$$\tilde{f}(t) \triangleq \frac{1}{K} \sum_{k=1}^{K} \sum_{n=1}^{N} \log p(x_n|\theta^{(t,k)}). \tag{5}$$

**Theorem 1.** *Assume that* $\{x_n\}_{n=1}^N$ *is i.i.d.,* $\log p(x,\theta)$ *is differentiable,* $f(t)$ *is twice differentiable and its second derivative is uniformly bounded, i.e.,* $|f''(t)| < U$ *for* $t \in [0,1]$ *and for some* $U > 0$*. The domain of the temperature variable* $t$ *is uniformly discretized, i.e.,* $\Delta t_i = \Delta t \triangleq (1/T)$ *for all* $i = 0, 1, \ldots, T-1$*, and* $\theta^{(t,k)}$ *is generated by an SG-MCMC method (see [1, 2]) with constant step-size* $\epsilon$*. We further assume that* $\log p(x|\theta)$ *satisfies the conditions given in Assumption 1 described in [2] . Then, the bias of STI can be bounded as:*

$$\Big| \langle \hat{\mathcal{L}} \rangle - \mathcal{L} \Big| = \mathcal{O}\Big( \frac{1}{K\epsilon} + \epsilon + \frac{1}{T^2} \Big), \tag{6}$$

*where* $K$ *is the number of samples (after burn-in) that are generated by the SG-MCMC method for a single power posterior.*

*Proof.* Let us first rewrite the true log-marginal likelihood as a sum of the areas under disjoint intervals, given as follows:

$$\mathcal{L} = \sum_{i=0}^{T-1} \int_{t_i}^{t_{i+1}} f(t)dt, \tag{7}$$

where $t_i \triangleq i\Delta t$. Now, let us consider the individual integrals in the summation. By applying integration by parts twice, these integrals can be written as follows [3]:

$$\int_{t_i}^{t_{i+1}} f(t)dt = \Delta t \frac{f(t_i) + f(t_{i+1})}{2} + g(t_i) \tag{8}$$

where

$$g(t) \triangleq \int_0^{\Delta t} \left( \frac{(y - \Delta t/2)^2}{2} - \frac{\Delta t^2}{8} \right) f''(y + t)dy. \tag{9}$$

Then, we can bound the bias as follows: (recall that $S$ is the data subsample that is chosen with replacement, $\langle \cdot \rangle$ denotes expectation with respect to $\theta$, and $\big\langle \big\langle \cdot | y \big\rangle_{x|y} \big\rangle_y$ denotes conditional expectation)

$$\begin{aligned}
\Big| \langle \hat{\mathcal{L}} \rangle - \mathcal{L} \Big| =& \Big| \langle \mathcal{L} - \hat{\mathcal{L}} \rangle \Big| && (10)\\
=& \Big| \langle \sum_{i=0}^{T-1} \Big( \Delta t \frac{f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})}{2} + g(t_i) \Big) \rangle \Big| && (11)\\
=& \Big| \langle \sum_{i=0}^{T-1} \Delta t \frac{f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})}{2} \rangle + \sum_{i=0}^{T-1} g(t_i) \Big| && (12)\\
\leq& \Big| \langle \sum_{i=0}^{T-1} \Delta t \frac{f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1})}{2} \rangle \Big| + \Big| \sum_{i=0}^{T-1} g(t_i) \Big| && (13)\\
=& \frac{\Delta t}{2} \Big| \sum_{i=0}^{T-1} \langle f(t_i) - \hat{f}(t_i) + f(t_{i+1}) - \hat{f}(t_{i+1}) \rangle \Big| + \Big| \sum_{i=0}^{T-1} g(t_i) \Big| && (14)\\
\leq& \frac{\Delta t}{2} \sum_{i=0}^{T-1} \Big| \langle f(t_i) \rangle - \langle \hat{f}(t_i) \rangle \Big| + \frac{\Delta t}{2} \sum_{i=0}^{T-1} \Big| \langle f(t_{i+1}) \rangle - \langle \hat{f}(t_{i+1}) \rangle \Big| + \Big| \sum_{i=0}^{T-1} g(t_i) \Big| && (15)\\
=& \frac{\Delta t}{2} \sum_{i=0}^{T-1} \Big| \langle f(t_i) \rangle - \big\langle \langle \hat{f}(t_i) | S \rangle_{\theta|S} \big\rangle_S \Big| + \frac{\Delta t}{2} \sum_{i=0}^{T-1} \Big| \langle f(t_{i+1}) \rangle - \big\langle \langle \hat{f}(t_{i+1}) | S \rangle_{\theta|S} \big\rangle_S \Big| + \Big| \sum_{i=0}^{T-1} g(t_i) \Big| && (16)\\
=& \frac{\Delta t}{2} \sum_{i=0}^{T-1} \Big| \langle f(t_i) \rangle - \langle \tilde{f}(t_i) \rangle \Big| + \frac{\Delta t}{2} \sum_{i=0}^{T-1} \Big| \langle f(t_{i+1}) \rangle - \langle \tilde{f}(t_{i+1}) \rangle \Big| + \Big| \sum_{i=0}^{T-1} g(t_i) \Big| && (17)
\end{aligned}$$

Since we assume $f''(t)$ is uniformly bounded, the last term can be bounded as follows [3]:

$$\Big| \sum_{i=0}^{T-1} g(t_i) \Big| \leq \frac{U}{12T^2}. \tag{18}$$

By using Theorem 2 of [2], we can bound the individual summands of the first two terms, as follows:

$$|\langle f(t_i) \rangle - \langle \tilde{f}(t_i) \rangle \Big| \leq C_i \Big( \frac{1}{K\epsilon} + \epsilon \Big) \tag{19}$$

for some $C_i > 0$. By defining $C \triangleq \max_{0 \leq i \leq T} C_i$ and using $\Delta t = 1/T$, we have

$$\begin{aligned}
\Big| \langle \hat{\mathcal{L}} \rangle - \mathcal{L} \Big| \leq& \frac{\Delta t}{2} \Big( \frac{1}{K\epsilon} + \epsilon \Big) \sum_{i=0}^{T-1} C_i + \frac{\Delta t}{2} \Big( \frac{1}{K\epsilon} + \epsilon \Big) \sum_{i=0}^{T-1} C_{i+1} + \frac{U}{12T^2} && (20)\\
\leq& \frac{\Delta t}{2} \Big( \frac{1}{K\epsilon} + \epsilon \Big) CT + \frac{\Delta t}{2} \Big( \frac{1}{K\epsilon} + \epsilon \Big) CT + \frac{U}{12T^2} && (21)\\
=& \Big( \frac{1}{K\epsilon} + \epsilon \Big) C + \frac{U}{12T^2}. && (22)
\end{aligned}$$

Finally, by defining $L \triangleq \max\{C, \frac{U}{12}\}$, we obtain the desired bound:

$$\Big|\langle\hat{\mathcal{L}}\rangle - \mathcal{L}\Big| \leq L\Big(\frac{1}{K\epsilon} + \epsilon + \frac{1}{T^2}\Big) \implies \Big|\langle\hat{\mathcal{L}}\rangle - \mathcal{L}\Big| = \mathcal{O}\Big(\frac{1}{K\epsilon} + \epsilon + \frac{1}{T^2}\Big). \tag{23}$$

□

## 2. REFERENCES


[1] Y. A. Ma, T. Chen, and E. Fox, "A complete recipe for stochastic gradient MCMC," in *Advances in Neural Information Processing Systems*, 2015, pp. 2899–2907.

[2] C. Chen, N. Ding, and L. Carin, "On the convergence of stochastic gradient MCMC algorithms with high-order integrators," in *Advances in Neural Information Processing Systems*, 2015, pp. 2269–2277.

[3] K.E. Atkinson, *An Introduction to Numerical Analysis*, Wiley India Pvt. Limited, New Delhi, 2 edition, 2008.