Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics

Yee Whye Teh, Alexandre H. Thiery, Sebastian J. Vollmer

Presented by Changyou Chen
January 30, 2015
1. Langevin Dynamics
2. Assumptions
3. Consistency Theorem
4. Fluctuations Theorem
5. Diffusion Limit
6. Experiments
Langevin dynamics

- Given data $X = \{x_1, \cdots, x_N\}$, a generative model $p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$ and prior $p(\theta)$, we want to compute the posterior $\pi(\theta) \triangleq p(\theta|X) \propto p(X|\theta)p(\theta)$.

- In statistical community, the Langevin dynamic are defined by the following SDE:

$$d\theta = \nabla_{\theta} \log \pi(\theta) dt + \sqrt{2B}dW , \quad (1)$$

where $W$ is the standard Brownian motion, $B$ is a constant.
To solve the above SDE, the Euler-Maruyama scheme is used to generate samples that approximate the law of the Langevin dynamics ($\delta$ is the step size):

$$
\theta_{m+1} = \theta_m + \frac{\delta}{2} \nabla_{\theta} \log \pi(\theta_m) + \zeta, \quad \zeta \sim N(0, \delta I)
$$

(2)

In large data setting, stochastic gradient which is evaluated on a subset of data is used instead of the true gradient:

$$
\theta_{m+1} = \theta_m + \frac{\delta}{2} \nabla_{\theta} \log \tilde{\pi}(\theta_m) + \zeta, \quad \zeta \sim N(0, \delta I)
$$

(3)

To enforce convergence, a decreasing step sizes $\{\delta_m\}$ is used:

$$
\theta_{m+1} = \theta_m + \frac{\delta_m}{2} \nabla_{\theta} \log \tilde{\pi}(\theta_m) + \zeta_m, \quad \zeta_m \sim N(0, \delta_m I)
$$

(4)
Outline

1. Langevin Dynamics
2. Assumptions
3. Consistency Theorem
4. Fluctuations Theorem
5. Diffusion Limit
6. Experiments
Two kinds of step sizes and assumptions

- A deceasing step sizes $\delta = (\delta_m)_{m \geq 1}$ satisfying:
  \[
  \lim_{m \to \infty} \delta_m = 0 \quad \text{and} \quad \lim_{m \to \infty} T(m) = \infty, \tag{5}
  \]
  where $T(m) \triangleq \delta_1 + \cdots + \delta_m$.

- A positive step sizes sequence $(\omega_m)_{m \geq 1}$ such that $\omega_m \to 0$ and $\Omega(m) \triangleq \omega_1 + \cdots + \omega_m \to \infty$ and
  \[
  \lim_{m \to \infty} \sum_{m \geq 1} \left| \Delta(\omega_m/\delta_m) \right| \Omega(m) < \infty \tag{6}
  \]
  \[
  \sum_{m \geq 1} \omega^2_m / \left| \delta_m \Omega^2(m) \right| < \infty, \tag{7}
  \]
  where $\Delta(\omega_w/\delta_m) \triangleq \omega_{m+1}/\delta_{m+1} - \omega_m/\delta_m$.
  - basically these constrains $\omega_m$ not go to far away from $\delta_m$
  - e.g., when $\delta_m = (m_0 + m)^{-\alpha} (0 < \alpha < 1)$, a choice of $\omega_m = \delta_m^p$ with $0 < p < 1/\alpha$ satisfies the constraints
Lyapunov function

A function that takes positive values everywhere except at the equilibrium in question, and non-increasing along every trajectory of the ODE:

- used to analyze the stability of ODEs where the actual solution of the ODE is not required

In the analysis of SGLD, the basic assumption is to assume there exists a Lyapunov function such that the gradient is kind of bounded by the function. In this way, properties such as the convergence can be proved.

In the following, $a \lesssim b$ means there exists a $C > 0$ such that $a \leq Cb$. 
Assumptions

Assumption

The drift term $\theta \rightarrow \frac{1}{2} \nabla \log \pi(\theta)$ continuous; there exists a Lyapunov function $V: \mathbb{R}^d \rightarrow [1, \infty)$ such that $V \rightarrow \infty$ as $\|\theta\| \rightarrow \infty$, twice differentiable with bounded derivatives, satisfying:

1. There are constants $\alpha, \beta > 0$ such that $\forall \theta \in \mathbb{R}^d$:

$$\langle \nabla V(\theta), \frac{1}{2} \nabla \log \tilde{\pi}(\theta) \rangle \leq -\alpha V(\theta) + \beta .$$

2. There exists $p_H \geq 2$ such that:

$$\mathbb{E} \left[ \|\nabla \log \tilde{\pi}(\theta) - \nabla \log \pi(\theta)\|^{2p_H} \right] \lesssim V^{p_H}(\theta) .$$

3. $\forall \theta \in \mathbb{R}^d$, $\|\nabla V(\theta)\|^2 + \|\nabla \log \pi(\theta)\|^2 \lesssim V(\theta)$

(8) ensure $\nabla \log \tilde{\pi}(\theta) = \mathbb{E}(\theta_{m+1} - \theta_m)$ points towards the center of the state space, (9) and (10) control the magnitude of the drift term.
**Assumptions**

**Property of $V(\theta)$**

- For any test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, define ($\theta_m$’s are samples from the SGLD)

\[
\pi_m(\varphi) \triangleq \left\{ \delta_1 \varphi(\theta_0) + \cdots + \delta_m \varphi(\theta_{m-1}) \right\} / T(m) \\
\pi^\omega_m(\varphi) \triangleq \left\{ \omega_1 \varphi(\theta_0) + \cdots + \omega_m \varphi(\theta_{m-1}) \right\} / \Omega(m)
\]  

(10) \hspace{1cm} (11)

**Lemma**

*Under the above assumption, for any exponent $0 \leq p \leq p_H$, the following holds almost surely:*

\[
\pi(V^p) < \infty
\]  

(12)

\[
\sup_{m>1} \pi_m(V^{p/2}) < \infty
\]  

(13)

\[
\sup_{m>1} \mathbb{E} [V^p(\theta_m)] < \infty
\]  

(14)

\[
\sup_{m>1} \pi^\omega_m(V^{p/2}) < \infty
\]  

(15)
Outline

1. Langevin Dynamics
2. Assumptions
3. Consistency Theorem
4. Fluctuations Theorem
5. Diffusion Limit
6. Experiments
Consistency theorem

For any test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, define

$$\pi_m(\varphi) \triangleq \{ \delta_1 \varphi(\theta_0) + \cdots + \delta_m \varphi(\theta_{m-1}) \} / T(m) \quad (16)$$

$$\pi^\omega_m(\varphi) \triangleq \{ \omega_1 \varphi(\theta_0) + \cdots + \omega_m \varphi(\theta_{m-1}) \} / \Omega(m) \quad (17)$$

$$\pi(\varphi) \triangleq \int_{\mathbb{R}^d} \varphi(\theta) \pi(d\theta) \quad (18)$$

Theorem (Consistency)

Let the step-sizes and the Lyapunov function $V : \mathbb{R}^d \to [1, \infty)$ satisfy the above assumptions. For $0 \leq p \leq p_H/2$ and a test function $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfying $|\varphi(\theta)/V^p(\theta)|$ being globally bounded, the followings hold almost surely:

$$\lim_{m \to \infty} \pi_m(\varphi) = \pi(\varphi) \quad (19)$$

$$\lim_{m \to \infty} \pi^\omega_m(\varphi) = \pi(\varphi) \quad (20)$$
Main ideas for the proof

- **Weak convergence of \((\pi_m)_{m \geq 1}\):**
  - prove \(\pi_\infty(\mathcal{A} \tilde{\phi}) = 0\), where \(\mathcal{A}\) is the generator of the LD and \(\tilde{\phi}\) is smooth and compactly supported
  - \(\frac{\sum_{k=1}^{m} \mathbb{E}_{k-1} [\tilde{\phi}(\theta_k) - \tilde{\phi}(\theta_{k-1})]}{T(m)} - \left\{ \frac{\sum_{k=1}^{m} \mathbb{E}_{k-1} [\tilde{\phi}(\theta_k) - \phi(\theta_{k-1})]}{T(m)} - \pi_m(\mathcal{A} \tilde{\phi}) \right\}\) converges to 0 when \(m \to \infty\)
  - first term \(\to 0\) with a martingale argument, second term \(\to 0\) using properties of the generator

- **Convergence for general test functions \(\varphi\) such that \(|\varphi(\theta)/V^p(\theta)|\) is globally bounded:**
  - by Tietze’s extension theorem there exists a continuous function \(\tilde{\varphi}\) with compact support agreed with \(\varphi\)
  - need to show:

\[
|\pi_m(\varphi) - \pi(\varphi)| \leq |\pi_m(\varphi) - \pi_m(\tilde{\varphi})| + |\pi(\varphi) - \pi(\tilde{\varphi})| + |\pi_m(\tilde{\varphi}) - \pi(\tilde{\varphi})| \to 0
\]

  - first two achieved by the almost surely finiteness of \(\sup_m \pi_m(V^{PH}/2)\), the last one by weak convergence of \(\pi_m\).
Fluctuations Theorem

Outline

1. Langevin Dynamics
2. Assumptions
3. Consistency Theorem
4. **Fluctuations Theorem**
5. Diffusion Limit
6. Experiments
Previous slides show the asymptotic value of $\pi_m(\varphi)$ is $\pi(\varphi)$ under certain assumptions.

What is the asymptotic fluctuation of $\pi_m(\varphi)$?
- what is the variance of $\pi_m(\varphi) - \pi(\varphi)$?
For this, a fourth order Taylor series expansion of $\pi_m(\varphi) - \pi(\varphi)$ is required:

$$
\pi_m(\varphi) - \pi(\varphi) = \frac{h(\theta_m) - h(\theta_0)}{T(m)} + \frac{1}{T(m)} \sum_{k=0}^{m-1} \left( \frac{1}{2} \nabla h(\theta_k) \nabla \log \pi(\theta_k) \delta_{k+1} + \frac{1}{2} \nabla^2 h(\theta_k) \delta_{k+1} - \sum_{n=1}^{k} \sum_{i=0}^{n} C_{n,i}^{(k)} \delta_{k+1}^{(n+i)/2} - R_5^{(k)} \right), \quad (22)
$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is the solution of the Poisson equation:

$$
\varphi - \pi(\varphi) = \mathcal{A} h, \quad (23)
$$

$$
C_{n,i}^{(k)} \triangleq \frac{1}{2i!(n-i)!} \nabla^n h(\theta_k) \nabla \log \tilde{\pi}(\theta_k)^i \eta_{n+1} \delta_{k+1}^{(n+i)/2}, \quad R_5^{(k)} \text{ is the left five ordered term.}
$$
It can be shown that all terms in the above expansion approaches to 0 except for the following grouped *Fluctuations* term and *Bias* term:

**Fluctuations:**
\[
\frac{1}{T(m)} \sum_{k=0}^{m-1} \nabla h(\theta_k) \eta_{k+1} \delta_{k+1}^{1/2} \Rightarrow O(T(m)^{-1/2}) \quad (24)
\]

**Bias:**
\[
\frac{1}{T(m)} \sum_{k=0}^{m-1} \left( \frac{1}{8} \nabla^2(\theta_k) \nabla \log \tilde{\pi}(\theta_k)^2 + \frac{1}{4} \nabla^3 h(\theta_k) \nabla \log \tilde{\pi}(\theta_k) \eta_{k+1}^2 \\
+ \frac{1}{24} \nabla^4 h(\theta_k) \eta_{k+1}^4 \right) \delta_{k+1}^2 \Rightarrow O(T(m)^{-1} \sum_{k=0}^{m-1} \delta_k^2) \quad (25)
\]

Define the asymptotic ratio of the bias and fluctuations as:
\[
\mathcal{B}(m) \triangleq \frac{1}{\sqrt{T(m)}} \sum_{k=0}^{m-1} \delta_{k+1}^2 . \quad (26)
\]
Theorem (Fluctuations)

Given the above assumptions, the followings hold a.s.:

- **In case the fluctuation dominate**, i.e., \( \mathbb{B}(m) \to 0 \):

  \[
  \lim_{m \to \infty} T^{1/2}(m) \{ \pi_m(\phi) - \pi(\phi) \} = N(0, \sigma^2(\phi)),
  \]  
  \( (27) \)

  where \( \sigma^2(\phi) = \lim_{T \to \infty} \text{Var}(T^{-1/2} \int_0^T [\phi - \pi(\phi)](\theta_s) \, ds) \).

- **In case the fluctuations and bias are on the same scale**, i.e., \( \mathbb{B}(m) \to \mathbb{B}_\infty \in (0, \infty) \):

  \[
  \lim_{m \to \infty} T^{1/2}(m) \{ \pi_m(\phi) - \pi(\phi) \} = N(\mu(\phi), \sigma^2(\phi)),
  \]  
  \( (28) \)

  \[
  \mu(\phi) = -\mathbb{B}_\infty \mathbb{E} \left[ \frac{1}{8} \nabla^2 h(\theta) \nabla \log \hat{\pi}(\theta, \mathcal{U})^2 + \frac{1}{4} \nabla^3 h(\theta) \nabla \log \hat{\pi}(\theta) + \frac{1}{24} \nabla^4 h(\theta) \right].
  \]

- **In case the bias dominate**, i.e., \( \mathbb{B}(m) \to \infty \):

  \[
  \lim_{m \to \infty} \frac{\pi_m(\phi) - \pi(\phi)}{T(m)^{-1} \sum_{k=1}^m \delta_k^2} = \mu(\phi)
  \]  
  \( (29) \)
In the choice of step-size sequence $\delta_m = (m_0 + m)^{-\alpha}$, the bias dominate when $0 < \alpha < 1/3$, fluctuations dominate when $1/3 < \alpha < 1$:

- the optimal rate of convergence is obtained for $\alpha = 1/3$, i.e., the largest decreasing coefficient we can choose to guarantee the solution is unbiased is when $\alpha = 1/3$.
Outline

1. Langevin Dynamics
2. Assumptions
3. Consistency Theorem
4. Fluctuations Theorem
5. Diffusion Limit
6. Experiments
Constructing continuous sample paths

Let \( \{ (\theta_1^1, \theta_2^1, \cdots), \cdots, (\theta_1^r, \theta_2^r, \cdots), \cdots \} \) be samples obtained from running multiple times of the SGLD algorithm. We construct piecewise affine continuous time sample paths \( (S^{(r)})_{r \geq 1} \) as:

\[
S^{(r)} \left( x T^{(r)}_{k-1} + (1 - x) T^{(r)}_k \right) = x \theta^{(r)}_{k-1} + (1 - x) \theta^{(r)}_k , \tag{30}
\]

where \( x \in [0, 1] \), \( T^{(r)}_0 = 0 \) and \( T^{(r)}_k = \delta^{(r)}_1 + \cdots + \delta^{(r)}_k \).
Diffusion Limit

Theorem

*Under certain assumptions, if*

\[
\text{mesh}(\delta^{(r)}) \triangleq \max \left\{ \delta_k^{(r)} : 1 \leq k \leq m(r) \right\} \to 0 \text{ as } r \to \infty, \text{ then}
\]

\[(S^{(r)})_{r \geq 1} \text{ defined above converges weakly on } (\mathcal{C}([0,T], \mathbb{R}^d), \| \cdot \|_{\infty}) \text{ to the Langevin diffusion below started at } S_0 = \theta_0:\]

\[
dS_t = \frac{1}{2} \nabla \log \pi(S_t) dt + dW_t \tag{31}
\]
Main ideas for the proof

For any time $T_{k-1}^{(r)} \leq t \leq T_k^{(r)}$, we have

$$S^{(r)}(t) = S^{(r)}(T_{k-1}^{(r)}) + \left( \int_{T_{k-1}^{(r)}}^{t} \frac{1}{2} \nabla \log \pi(S^{(r)}(T_{k-1}^{(r)}))\,du + \tilde{W}^{(r)}(t) - \tilde{W}(T_{k-1}^{(r)}) \right)$$

$$+ \int_{T_{k-1}^{(r)}}^{t} \frac{1}{2} \left( \nabla \log \tilde{\pi}(S^{(r)}(T_{k-1}^{(r)})) - \nabla \log \pi(S^{(r)}(T_{k-1}^{(r)})) \right)\,du,$$

(32)

where $\tilde{W}^{(r)}$ is the continuous piecewise affine processes that agree with $W$ in $T_k^{(r)}$ and is affine between $T_k^{(r)}$ and $T_{k+1}^{(r)}$, i.e., $\tilde{W}^{(k)}$ weakly converges to $W$. Some algebra yields

$$S^{(r)}(t) = \theta_0 + \left( \int_{0}^{t} \frac{1}{2} \nabla \log \pi(S^{(r)}(u))\,du + \tilde{W}^{(r)}(t) \right)$$

plus some terms that converge to 0.
Outline

1. Langevin Dynamics
2. Assumptions
3. Consistency Theorem
4. Fluctuations Theorem
5. Diffusion Limit
6. Experiments
Linear Gaussian model

\[
x_i | \theta \sim N(\theta, \sigma_x^2) \tag{33}
\]
\[
\theta | \sigma_\theta \sim N(0, \sigma_\theta^2) \tag{34}
\]

where \(\sigma_\theta\) and \(\theta_x^2\) are known.

- Satisfy those assumptions: there exists a Lyapunov function \(V(\theta) = 1 + \frac{(\theta - \mu_p)^2}{2\sigma_p^2}\), where \(\mu_p\) and \(\sigma_p^2\) are the posterior mean and variance of the model respectively, such that:

\[
\langle \nabla V(\theta), \frac{1}{2} \nabla \log \pi(\theta) \rangle \leq -\alpha V(\theta) + \beta \tag{35}
\]

\[
\mathbb{E} \left[ \| \nabla \log \tilde{\pi}(\theta) - \nabla \log \pi(\theta) \|^2_{PH} \right] \lesssim V^{PH}(\theta) \tag{36}
\]

\[
\| \nabla V(\theta) \|^2 + \| \nabla \log \pi(\theta) \|^2 \lesssim V(\theta) \tag{37}
\]
Linear Gaussian model

Figure 1: Decay of the MSE for step sizes $\delta_m \approx m^{-\alpha}$, $\alpha \in \{0.1, 0.2, 0.3, 0.33, 0.4, 0.5\}$. The MSE decays algebraically for all step sizes, with fastest decay at approximately $\alpha = 0.33$.

7.1.1 Verification of Assumption 4

We verify in this section that Assumption (4) is satisfied for the following choice of Lyapunov function,

$$V(\theta) = 1 + (\theta - \mu_p)^2 \sigma_p^2.$$ 

Since the error term $H(\theta, U)$ is globally bounded, the drift $(1/2)\nabla \log \pi$ and the Lyapunov function $V$ are linear, Assumptions (4).1 and (4).2 are satisfied. Finally, to verify Assumption (4).3, it suffices to note that since $\nabla \log \pi(\theta) = -(\theta - \mu_p)/\sigma_p^2$ we have

$$\langle \nabla V(\theta), 1/2 \nabla \log \pi(\theta) \rangle = - (\theta - \mu_p)^2 \sigma_p^4 = 1 - V(\theta) \sigma_p^2.$$ 

In other words, Assumption (4).3 holds with $\alpha = \beta = 1/\sigma_p^2$.

7.1.2 Simulations

We chose $\sigma_\theta = 1, \sigma_x = 5$ and created a data set consisting of $N = 100$ data points simulated from the model. We used $n = 10$ as the size of subsets used to estimate the gradients. We evaluated the convergence behaviour of SGLD using the test function $A\phi$ with

$$\phi = \sin (x - \mu_p - 0.5\sigma_p)$$

$$A\phi = - x - \mu_p \frac{\sigma_p^2}{2} \cos (x - \mu_p - 0.5\sigma_p) + \frac{1}{2} \sin (x - \mu_p - 0.5\sigma_p)^2.$$
Logistic regression

\[ p(y_i = 1|x_i, \theta) = \logit(\theta^T x_i) \quad (38) \]
\[ \theta \sim \mathcal{N}(0, I) \quad (39) \]

There exists a Lyapunov function \( V(\theta) = 1 + \|\theta\|^2 \) such that the assumptions are satisfied:

\[ \langle \nabla V(\theta), \frac{1}{2} \nabla \log \pi(\theta) \rangle \leq -\alpha V(\theta) + \beta \quad (40) \]
\[ \mathbb{E} [\|\nabla \log \tilde{\pi}(\theta) - \nabla \log \pi(\theta)\|_{2PH}^2] \lesssim V^{PH}(\theta) \quad (41) \]
\[ \|\nabla V(\theta)\|^2 + \|\nabla \log \pi(\theta)\|^2 \lesssim V(\theta) \quad (42) \]
Figure 5: Behaviour of the mean squared error for different subsample sizes $n$ for 3-dimensional logistic regression of the step-sizes sequence $(\delta_m)_{m \geq 0}$. The consistency of the algorithm is mainly due to the decreasing step-sizes procedure that asymptotically removes the bias from the discretisation and ultimately mitigates the use of an unbiased estimate of the gradient instead of the exact value. Additionally, we have proved a diffusion limit result that establishes that, when observed on the right (inhomogeneous) time scale, the sample paths of the SGLD can be approximated by a Langevin diffusion.

The CLT and bias-variance decomposition can be leveraged to show that it is optimal to choose a step-sizes sequences $(\delta_m)_{m \geq 0}$ that scales as $\delta_m \approx m^{-1/3}$; the resulting algorithm converges at rate $m^{-1/3}$. Note that this recommendation is different from the previously suggested Welling and Teh (2011) choice of $\delta_m \approx m^{-1/2}$.

Our theory suggests that an optimally tuned SGLD method converges at rate $O(m^{-1/3})$, and is thus asymptotically less efficient than a standard MCMC procedure. We believe that this result does not necessarily preclude SGLD to be more efficient in the initial transient phase, a result hinted at in Figure 4; the detailed study of this (non-asymptotic) phenomenon is an interesting venue of research. The asymptotic convergence rate of SGLD depends crucially on the decreasing step sizes, which is required to reduce the effect of the discretisation bias due to the lack of a Metropolis-Hastings correction. Another avenue of exploration is to determine more precisely the bias resulting from the discretisation of the Langevin diffusion, and to study the effect of the choice of step sizes in terms of the trade-off between bias, variance, and computation.

Appendix A. Proof of Lemma 5

For clarity, the proof is only presented in the scalar case $d = 1$; the multidimensional setting is entirely similar. Before embarking on the proof, let us first mention some consequences of Assumptions 4 that will be repeatedly used in the sequel. Since the second derivative $V''$ is globally bounded and $(V')^2$ is upper bounded by a multiple of $V$, we have that
\[
\left| \left( V^p \right)'(\theta) \right| \lesssim V^p(\theta)^{-1} \tag{38}
\]

\[
\alpha = 0.35 \quad N = 100
\]

\[
\text{MSE} \quad 10^3 \quad 10^5
\]

\[
\text{Likelihood evaluations}
\]
Thanks for your attention!!!