On the stability of the stochastic gradient Langevin algorithm with dependent data stream*

Miklós Rásonyi¹, Kinga Tikosi¹²

Abstract

We prove, under mild conditions, that the stochastic gradient Langevin dynamics converges to a limiting law as time tends to infinity, even in the case where the driving data sequence is dependent.

Keywords: stochastic gradient, Langevin dynamics, dependent data

1. Stochastic gradient Langevin dynamics

Sampling from high-dimensional, possibly not even logconcave distributions is a challenging task, with far-reaching applications in optimization, in particular, in machine learning, see Raginsky et al. [1], Chau et al. [2], Barkhagen et al. [3], Brosse et al. [4].

Let \( U : \mathbb{R}^d \rightarrow \mathbb{R}_+ \) be a given function and consider the corresponding Langevin equation

\[
d\Theta_t = -\nabla U(\Theta_t) \, dt + \sqrt{2} \, dW_t,
\]

*Both authors were supported by the “Lendület” grant 2015-6 of the Hungarian Academy of Sciences.

¹Alfréd Rényi Institute of Mathematics, Réaltandóda utca 13-15, 1053 Budapest, Hungary

²During the preparation of this paper the author attended the PhD school of Central European University, Budapest.
where $W$ is a $d$-dimensional standard Brownian motion. Under suitable assumptions, the unique invariant probability $\mu$ for the diffusion process $\pi$ has a density (with respect to the $d$-dimensional Lebesgue measure) that is proportional to $\exp(-U(x))$, $x \in \mathbb{R}^d$.

In practice, Euler approximations of $\pi$ may be used for sampling from $\mu$, i.e. a recursive scheme

$$\vartheta_{t+1}^\lambda = \vartheta_t^\lambda - \lambda \nabla U(\vartheta_t^\lambda) + \sqrt{2\lambda} \xi_{t+1}$$

(2)
is considered for some small $\lambda > 0$ and independent standard $d$-dimensional Gaussian sequence $\xi_i$, $i \geq 1$.

In some important applications, however, $U, \nabla U$ are unknown, one disposes only of unbiased estimates $H(\theta, Y_t)$, $t \in \mathbb{N}$ of $\nabla U(\theta)$, where $Y_t$ is some stationary data sequence. From this point on we switch to rigorous mathematics.

Let us fix integers $d, m \geq 1$ and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $\mathcal{B}(\mathcal{X})$ denotes the $\sigma$-algebra of the Borel-sets of a Polish space $\mathcal{X}$. For a random variable $X$, $\mathcal{L}(X)$ denotes its law. The Euclidean norm on $\mathbb{R}^d$ or $\mathbb{R}^m$ will be denoted by $| \cdot |$, while $|| \cdot ||_{TV}$ stands for the total variation distance of probability measures on $\mathcal{B}(\mathbb{R}^d)$. Let $B_r := \{ \theta \in \mathbb{R}^k : |\theta| \leq r \}$ denote the ball of radius $r$, for $r \geq 0$, for both $k = d$ and $k = m$, depending on the context. The notation Leb($\cdot$) refers to the $d$-dimensional Lebesgue-measure.

For $0 < \lambda \leq 1$, $t = 0, 1, \ldots$ and for a constant initial value $\theta_0 \in \mathbb{R}^d$ consider the recursion

$$\theta_{t+1}^\lambda = \theta_t^\lambda - \lambda H(\theta_t^\lambda, Y_t) + \sqrt{\lambda} \xi_{t+1}, \quad t \in \mathbb{N}, \quad \theta_0^\lambda := \theta_0,$$

(3)

where $\xi_i$, $i \geq 1$ is an i.i.d. sequence of $d$-dimensional random variables with
independent coordinates such that $E[\xi_i] = 0$ and $E[|\xi_i|^2] = \sigma^2$ for some $\sigma^2$. Furthermore, the density function $f$ of $\xi_i$ with respect to Leb is assumed strictly positive on every compact set. Assume that $(Y_t)_{t \in \mathbb{Z}}$ is a strict sense stationary process with values in $\mathbb{R}^m$ and it is independent of the noise process $(\xi_t)_{t \geq 1}$. Finally, $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a measurable function.

A particular case of (3) is the stochastic gradient Langevin dynamics (SGLD), introduced in Welling and Teh [5], designed to learn from large datasets. See more about different versions of SGLD and their connections in Brosse et al. [4]. Note that in the present setting, unlike in SGLD, we do not assume that $H$ is the gradient of a function and we do not assume $\xi_i$ to be Gaussian.

A setting similar to ours was considered in Lovas and Rásonyi [6] under different assumptions. We will compare our results to those of Lovas and Rásonyi [6] at the end of Section 2 below.

The sampling error of $\theta^{\lambda}_t$ has been thoroughly analysed in the literature: $d(\theta^{\lambda}_t, \mu)$ has been estimated for various probability metrics $d$, see Chau et al. [2], Barkhagen et al. [3], Raginsky et al. [1], Brosse et al. [4]. The ergodic behaviour of $\theta^{\lambda}_t$, however, has eluded attention so far. If $Y_t$ are i.i.d. then $\theta^{\lambda}_t$ is a homogeneous Markov chain and standard results of Markov chain theory apply. In the more general, stationary case (considered in Barkhagen et al. [3], Chau et al. [2]), however, that machinery is not available. In the present note we study scheme (3) with stationary $Y_t$ and establish that its law converges to a limit in total variation.
2. Main results

Assumption 1. There is a constant $\Delta > 0$ and a measurable function $b : \mathbb{R}^m \to \mathbb{R}_+$ such that, for all $\theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$

$$\langle H(\theta, y), \theta \rangle \geq \Delta |\theta|^2 - b(y). \quad (4)$$

Assumption 2. There exist constants $K_1, K_2, K_3 > 0$ and $\beta \geq 1$ such that

$$|H(\theta, y)| \leq K_1|\theta| + K_2|y|^{\beta} + K_3. \quad (5)$$

Assumption 3. There exist (finite) constants $M_y, M_b > 0$ such that $\mathbb{E}[|Y_0|^{2\beta}] \leq M_y$ and $\mathbb{E}[b(Y_0)] \leq M_b$.

Theorem 2.1. Let Assumptions 1, 2 and 3 hold. Then, for $\lambda$ small enough, the law $\mathcal{L}(\theta^\lambda_0)$ of the iteration defined by (3) converges in total variation as $t \to \infty$ and the limit does not depend on the initialization $X_0$.

In Lovas and Rásonyi [6], $\Delta$ in (4) was allowed to depend on $y$ but $b$ in (4) had to be constant, the process $Y$ was assumed bounded and the process $\xi$ Gaussian. Furthermore, in Assumption 2, $\beta$ had to be 1. Under these conditions the conclusion of Theorem 2.1 was obtained, together with a rate estimate.

Theorem 2.1 above complements the results of Lovas and Rásonyi [6]: $\Delta$ must be constant in our setting but the restrictive boundedness hypothesis on $Y$ could be removed, $\xi$ need not be Gaussian, $\beta$ in (4) can be arbitrary and $b$ in (4) may depend on $y$. The examples in Section 5 demonstrate that our present results cover a wide range of relevant applications where the obtained generalizations are crucial.
3. Markov chains in random environment

The rather abstract Theorem 3.1 below, taken from Gerencsér and Rásonyi [7], is the key result we use in this paper. Let us first recall the related terminology and the assumptions.

Let $X$ and $Y$ be Polish spaces and let $(X_n)_{n \in \mathbb{N}}$ (resp. $(Y_n)_{n \in \mathbb{N}}$) be a non-decreasing sequence of (non-empty) Borel-sets in $X$ (resp. $Y$). Consider a parametric family of transition kernels, i.e. a map $Q : Y \times X \times \mathcal{B}(X) \to [0,1]$ such that for all $B \in \mathcal{B}(X)$ the function $(x,y) \mapsto Q(x,y,B)$ is measurable and for every $(x,y) \in X \times Y$ $Q(x,y,\cdot)$ is a probability.

An $X$ valued stochastic process $(X_t)_{t \in \mathbb{N}}$ is called a Markov chain in a random environment with transition kernel $Q$ if $X_0 \in X$ is deterministic (for simplicity) and

$$\mathbb{P}(X_{t+1} \in A | \mathcal{F}_t) = Q(X_t, Y_t, A), \text{ for } t \in \mathbb{N},$$

where we use the filtration $\mathcal{F}_t = \sigma(Y_k, k \in \mathbb{Z}; X_j, 0 \leq j \leq t)$.

For a parametric family of transition kernels $Q$ and a bounded (or non-negative) function $V : X \to \mathbb{R}$ define

$$[Q(y)V](x) = \int_X V(z)Q(x,y,dz), \text{ for } x \in X. \quad (7)$$

**Assumption 4.** Let the process $(X_t)_{t \in \mathbb{N}}$ started from $X_0$ with be such that

$$\sup_{t \in \mathbb{N}} \mathbb{P}(X_t \notin X_n) \to 0, n \to \infty. \quad (8)$$

**Assumption 5.** (Minorization condition) Let $\mathbb{P}(Y_0 \notin Y_n), n \to \infty$. Assume that there exists a sequence of probability measures $(\nu_n)_{n \in \mathbb{N}}$ and a non-decreasing sequence $(\alpha_n)_{n \in \mathbb{N}}$ with $\alpha_n \in (0,1]$ such that for all $n \in \mathbb{N}, x \in$
Theorem 3.1. (Theorem 2.11. of Gerencsér and Rásonyi [7]) Let Assumptions \( 4 \) and \( 5 \) hold. Then there exists a probability \( \mu^* \) on \( \mathcal{B}(\mathcal{X} \times \mathcal{Y}) \) such that

\[
||\mathcal{L}(X_t, (Y_{t+k})_{k \in \mathbb{Z}}) - \mu^*||_{TV} \to 0, \quad \text{as } t \to \infty.
\]

If \((X'_t)_{t \in \mathbb{N}}\) is another such Markov chain started from a different \( X'_0 \) satisfying Assumption \( 4 \) then

\[
||\mathcal{L}(X_t, (Y_{t+k})_{k \in \mathbb{Z}}) - \mathcal{L}(X'_t, (Y_{t+k})_{k \in \mathbb{Z}})||_{TV} \to 0, \quad \text{as } t \to \infty. \qquad \Box
\]

4. Proofs

Define the Markov chain associated to the recursive scheme (3) as

\[
Q(\theta, y, A) = \mathbb{P}(\theta - \lambda H(\theta, y) + \sqrt{\lambda} \xi_{n+1} \in A),
\]

for all \( y \in \mathcal{Y} := \mathbb{R}^m, \theta \in \mathcal{X} := \mathbb{R}^d \) and \( A \in \mathcal{B}(\mathbb{R}^d) \).

Lemma 4.1. For small enough \( \lambda \), under Assumptions \( 1 \) and \( 2 \), the process \((\theta^k_t)_{t \in \mathbb{N}}\) given by recursion (3) satisfies Assumption \( 4 \) with \( \mathcal{X}_n := B_n \) (the ball of radius \( n \)).
PROOF. Choose $V(\theta) = |\theta|^2$. Then, since $E\xi_1 = 0$,

$$[Q(y)V](\theta) = \mathbb{E}[V(\theta - \lambda H(\theta, y) + \sqrt{\lambda}\xi_1)]$$

$$= |\theta|^2 + \lambda^2|H(\theta, y)|^2 + \lambda\mathbb{E}|\xi_1|^2 - 2\lambda\langle \theta, H(\theta, y) \rangle$$

$$\leq (1 - 2\lambda\Delta)|\theta|^2 + \lambda(\sigma^2 + 2b(y)) + 3\lambda^2(K_1^2|\theta|^2 + K_2^2|y|^{2\beta} + K_3^2)$$

$$= (1 - 2\lambda\Delta + 3\lambda^2K_1^2)V(\theta) + \lambda(\sigma^2 + 2b(y)) + 3\lambda^2(K_2^2|y|^{2\beta} + K_3^2)$$

$$= \gamma V(\theta) + K(y),$$

with $K(y) = \lambda(\sigma^2 + 2b(y)) + 3\lambda^2[K_2^2|y|^{2\beta} + K_3^2]$ and $\gamma = (1 - 2\lambda\Delta + 3\lambda^2K_1^2)$.

Note that for small enough $\lambda$, $\gamma \in (0, 1)$, independent of $y$.

Now using Lemma 4.2 below and setting $\theta = \theta_0$ and $y_k = Y_k$ for $k \geq 1$ we get, for each $t \geq 1$,

$$\mathbb{E}|\theta_1^\lambda|^2 = \mathbb{E}[Q(Y_t)Q(Y_{t-1}) \ldots Q(Y_1)V](\theta_0) \leq \gamma^t V(\theta_0) + \sum_{i=1}^{t} \gamma^i \mathbb{E}K(Y_i)$$

$$= \gamma^t|\theta_0|^2 + \sum_{i=1}^{t} \gamma^i[\lambda(\sigma^2 + 2\mathbb{E}[b(Y_i)]) + 3\lambda^2(K_2^2\mathbb{E}|Y_i|^{2\beta} + K_3^2)]$$

$$\leq |\theta_0|^2 + \frac{\gamma}{1 - \gamma}[(\sigma^2 + 2M_b) + 3(K_2^2M_y + K_3^2)] < \infty,$$

by Assumption 3. Then, using Markov’s inequality, we arrive at

$$\mathbb{P}(\theta_1^\lambda \notin \mathcal{X}_n) = \mathbb{P}(|\theta_1^\lambda| > n) \leq \frac{\sup_1 \mathbb{E}|\theta_1^\lambda|^2}{n^2} \to 0, \text{ as } n \to \infty. \quad (11)$$

**Lemma 4.2.** Assume $[Q(y)V](\theta) \leq \gamma V(\theta) + K(y)$. Then

$$[Q(y_k)Q(y_{k-1}) \ldots Q(y_1)V](\theta) \leq \gamma^k V(\theta) + \sum_{i=1}^{k} \gamma^{i-1} K(y_i). \quad (12)$$

**Proof.** We prove the statement by induction. For $k = 1$, it is true by assumption. Using that

$$[Q(y_2)Q(y_1)V](x) = \int_{\mathcal{X}} Q(x, y_2, dr) \int_{\mathcal{X}} V(z)Q(r, y_1, dz), \text{ for } r \in \mathcal{X}, \quad (13)$$
for \( k > 1 \) we get

\[
[Q(y_k)Q(y_{k-1}) \cdots Q(y_1)V](\theta) = \int_{\mathcal{X}} Q(\theta, y_k, dx) \left[ Q(y_{k-1})Q(y_{k-2}) \cdots Q(y_1)V \right](x)
\]

\[
\leq \int_{\mathcal{X}} \left( \gamma^{k-1}V(x) + \sum_{i=1}^{k-1} \gamma^{i-1} K(y_i) \right) Q(\theta, y_k, dx)
\]

\[
= \gamma^{k-1} \int_{\mathcal{X}} V(x)Q(\theta, y_k, dx) + \sum_{i=1}^{k-1} \gamma^{i-1} K(y_i)
\]

\[
\leq \gamma^k V(\theta) + \sum_{i=1}^{k} \gamma^{i-1} K(y_i). \quad \square
\]

**Lemma 4.3.** Define \( \mathcal{X}_n = B_n, \mathcal{Y}_n := B_n, n \in \mathbb{N} \) and let Assumptions 7 and 2 hold. Then Assumption 5 is satisfied, for all \( \lambda \).

**Proof.** For all \( A \in \mathcal{B}(\mathcal{X}) \),

\[
Q(\theta, y, A) = \mathbb{P}(\theta - \lambda H(\theta, y) + \sqrt{\lambda} \xi_1 \in A)
\]

\[
\geq \int_{\mathbb{R}^d} 1_{\{\theta - \lambda H(\theta, y) + \sqrt{\lambda} \xi_1 \in A \cap B_n\}} f(w) \, dw
\]

\[
= \frac{1}{\lambda^{d/2}} \int_{A \cap B_n} f \left( \frac{z - \theta + \lambda H(\theta, y)}{\sqrt{\lambda}} \right) \, dz
\]

\[
\geq \frac{\text{Leb}(A \cap B_n)}{\lambda^{d/2}} C(n) = \frac{\text{Leb}(A \cap B_n)}{\lambda^{d/2}} \frac{C(n) \text{Leb}(B_n)}{\text{Leb}(B_n)},
\]

where we use that for \( \theta, z \in B_n \) and \( y \in B_n \) we have

\[
\left| z - \theta + \lambda H(\theta, y) \right| \leq n + \lambda (K_1 n + K_2 n^\beta + K_3) =: R(n).
\]

Therefore the integrand can be bounded from below by

\[
C(n) := \inf_{x \in B_R(n)} f(x) > 0.
\]

Then define \( \nu_n(A) := \frac{\text{Leb}(A \cap B_n)}{\text{Leb}(B_n)} \) and \( \alpha_n := \frac{C(n) \text{Leb}(B_n)}{\lambda^{d/2}} \), which proves that Assumption 5 holds. \( \square \)

**Proof (of Theorem 2.1).** Follows from Lemmas 4.1, 4.3 and Theorem 3.1 \( \square \)
5. Examples

5.1. Nonlinear regression

Let us consider a nonlinear regression problem which can also be seen as a one layer neural network in a supervised learning setting, where only one trainable layer connects the input and the output vectors. The training set consists of entries \( Y_t = (Z_t, L_t) \) with the features \( Z_t \in \mathbb{R}^{d_0} \) and the corresponding labels \( L_t \in \mathbb{R}^{d_1} \) for \( t \in 1, \ldots, N \). We assume that \( Y_t \) is a stationary process. Set \( m := d_0 + d_1 \), the dimension of \( Y_t \).

The trainable parameters will be a matrix \( W \in \mathbb{R}^{d_0 \times d_1} \) and a vector \( g \in \mathbb{R}^{d_1} \), therefore the dimension of \( \theta := (W, g) \) will be \( d = d_0 d_1 + d_1 \). The prediction function \( h : \mathbb{R}^{d_0} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_1} \) is defined by \( h(z, \theta) := s(W z + g) \), where \( s = (s_1, \ldots, s_{d_1}) \) is a collection of nonlinear activation functions \( s_i : \mathbb{R} \rightarrow \mathbb{R} \) for \( i = 1, \ldots, d_1 \). We will assume that each \( s_i \) and their derivatives \( s_i' \) are all bounded by some constant \( M_s \) for \( i = 1, \ldots, d_1 \).

Choosing the loss function to be mean-square error, one aims to minimize the empirical risk, that is

\[
\min \{ \mathbb{E}[|h(Z_t, \theta) - L_t|^2] + \kappa |\theta|^2 \},
\]

with some \( \kappa > 0 \), where the second term is added for regularization.

It is standard to solve this optimization step using gradient-based methods. For \( y = (z, l) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \) denote \( U(\theta, y) = |h(z, \theta) - l|^2 + \kappa |\theta|^2 \) and the updating function to be used in the algorithm will be

\[
H(\theta, y) = \nabla U(\theta, y) = \frac{\partial}{\partial \theta} |h(z, \theta) - l|^2 + 2\kappa \theta. \tag{15}
\]

**Lemma 5.1.** The function \( H(\theta, y) \) defined as above satisfies Assumptions \( \square \) and \( \square \)
Proof. Using the chain rule, a short calculation gives
\[
\left| \frac{\partial}{\partial \theta} h(z; \theta) - l \right|^2 = \sum_{i=1}^{d_0+1} \sum_{j=1}^{d_1} (2(h(z; \theta))_j - l_j) s_j'(\langle W_j, z \rangle + g_j) z_i^2,
\]
where we define \( z_{d_0+1} = 1 \) and \( W_j \) stands for the \( j \)th row of \( W \). Notice that by the boundedness of \( s' \) and \( s \) this is bounded in \( \theta \) and at most quadratic in \( y \). Then Assumption \( 2 \) is satisfied with \( \beta = 2 \).

Using the same argument about the boundedness of \( s \) and \( s' \)
\[
\left| \langle \frac{\partial}{\partial \theta} h(z; \theta) - l \rangle^2, \theta \rangle \right| = \left| \sum_{i=1}^{d_0+1} \sum_{j=1}^{d_1} 2(h(z; \theta))_j - l_j \right| s_j'(W_j z + g_j) z_i \theta_{i,j} \right| \leq C_0 d M_s \left( |y|^2 + 1 \right) |\theta|,
\]
for some \( C_0 > 0 \). Using that \( \langle \frac{\partial}{\partial \theta} \kappa|\theta|^2, \theta \rangle = 2 \kappa |\theta|^2 \), we get that \( \langle \nabla U(\theta), \theta \rangle \geq c |\theta|^2 - C (|y|^4 + 1) \) with some \( c, C \) therefore Assumption \( 1 \) is satisfied with \( b(y) \) being of degree 4 in \( y \).

5.2. A tamed algorithm for neural networks

It has been observed that in multi-layer neural networks quadratic regularization is not sufficient to guarantee dissipativity, while adding a higher order term would violate Lipschitz continuity. So the standard SGLD algorithm diverges anyway. To remedy this, certain “tamed” schemes have been suggested in Lovas et al. \[8\].

In contrast to the previous case now we will hidden layers between the input and output: layer 0 is the input, layer \( n \) is the output and 1, \ldots, \( n-1 \) are the hidden layers of the neural network for some \( n > 1 \). The prediction function \( h \) will be defined as the composition of a sequence of \( n+1 \) linear transformations and activation functions, i.e. \( h(z, \theta) = s_n(W_n s_{n-1}(W_{n-1} \ldots s_0(W_0 z))) \)
where $\theta$ is the collection of all parameters $W_i \in \mathbb{R}^{d_{i-1} \times d_i}$, $i = 1, \ldots, n$ and $s_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ is a componentwise non-linear activation function, assumed bounded together with its derivatives by some constant $M_s$. Therefore $h : \mathbb{R}^{d_0} \times \mathbb{R}^{d} \to \mathbb{R}^{d_n}$, where $d = \sum_{i=1}^{n} d_{i-1}d_i$ is the dimension of $\theta$. For the case of simplicity in this case we assumed that there is no bias term $g$. The training set consists of entries $Y_t = (Z_t, L_t)$ with the features $Z_t \in \mathbb{R}^{d_0}$ and the corresponding labels $L_t \in \mathbb{R}^{d_n}$, the dimension of each $Y_t$ is $m = d_0 + d_n$.

We assume that $Y_t$ is a stationary process.

As in the previous subsection, the regularized empirical risk has the form

$$U(\theta, y) = |h(z, \theta) - l|^2 + \frac{\eta}{2(r+1)}|\theta|^{2(r+1)}$$

with some $r \geq 0$, $\eta > 0$. Denoting $G(\theta, y) = \nabla U(\theta, y)$, the “tamed” updating function we use will be defined as $H(\theta, y) := \frac{G(\theta, y)}{1 + \sqrt{\lambda} |\theta|^{2r}}$, for every $\theta \in \mathbb{R}^d$, $y \in \mathbb{R}^m$. Note that this function depends on $\lambda$.

We will use the following.

**Lemma 5.2.** (Proposition 4 of Lovas et al. [8])

$$\left| \frac{\partial}{\partial \theta} h(z, \theta) - l \right|^2 \leq C (1 + |y|)^2 (1 + |\theta|^{n+1}),$$

(17)

where $C > 0$ depends on $D = \max_{j=1, \ldots, n} d_j$, $n$ and $M_s$. \qed

**Lemma 5.3.** For $\lambda$ small enough, the conclusions of Theorem [2.7] hold for the scheme [3] with $H(\theta, y)$ defined as above, provided that $r \geq \frac{n+2}{2}$ and Assumption [2] holds.

**Proof.** Using Lemma 5.2, Assumption 2 can be checked as follows:

$$|H(\theta, y)| = \frac{\left| \frac{\partial}{\partial \theta} h(z, \theta) - l \right|^2 + \eta |\theta|^{2r}}{1 + \sqrt{\lambda} |\theta|^{2r}} \leq \left| \frac{C (1 + |y|)^2 (1 + |\theta|^{n+1})}{1 + \sqrt{\lambda} |\theta|^{2r}} \right| + \frac{\eta |\theta|^{2r}}{1 + \sqrt{\lambda} |\theta|^{2r}} \leq K_1 |\theta| + K_2 |y|^\beta + K_3.$$

11
where $K_1 = \frac{\eta}{\sqrt{\lambda}}$, $\beta = 2$ and the constants $K_2$ and $K_3$ depend on $\lambda, \eta, n$ and $C$.

Let us check Assumption 1. For the regularization term we have

$$\langle \eta\theta|\theta|^{2r} \rangle = \frac{\eta|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}} \geq \min \left\{ \frac{\eta^2}{2\sqrt{\lambda}}, \frac{\eta}{2} \right\} |\theta|^2 \geq \frac{\eta}{2} |\theta|^2$$

(18)

for $\lambda$ small enough.

The Cauchy inequality, Lemma 5.2 and the choice of $r$ ensures that

$$\left| \left\langle \frac{\partial}{\partial \theta} (|h(z; \theta) - l|^2) \right\rangle \right| \leq C(1 + |\theta|^{n+2})(1 + |y|^2) \leq K'(1 + |y|^2),$$

(19)

for some $K' > 0$. Now combining these estimates, we get

$$\langle H(\theta, y), \theta \rangle \geq \frac{\eta}{2} |\theta|^2 - K'(1 + |y|^2),$$

(20)

therefore Assumption 1 is satisfied with $\Delta = \frac{\eta}{2}$ and $b(y)$ is quadratic in $y$.

We can check that $\gamma = (1 - \eta\sqrt{\lambda} + \lambda\eta^2) < 1$ in Lemma 4.1 for $\lambda$ small enough so the proof of Theorem 2.1 goes through for this choice of $H$.

Allowing $b$ to be of degree 4, $\frac{n+2}{2}$ in Lemma 5.3 could be decreased to $\frac{n+1}{2}$, as easily seen.

References

[1] M. Raginsky, A. Rakhlin, M. Telgarsky, Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis, Proceedings of Machine Learning Research 65 (2017) 1674–1703.

[2] N. H. Chau, Éric Moulines, M. Rásonyi, S. Sabanis, Y. Zhang, On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case, Preprint, arXiv:1905.13142 (2021).
[3] M. Barkhagen, N. H. Chau, Éric. Moulines, M. Rásonyi, S. Sabanis, Y. Zhang, On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case, Bernoulli 27 (2021) 1–33.

[4] N. Brosse, A. Durmus, E. Moulines, The promises and pitfalls of stochastic gradient Langevin dynamics, in: Advances in Neural Information Processing Systems, 2018, pp. 8268–8278.

[5] M. Welling, Y. W. Teh, Bayesian learning via stochastic gradient Langevin dynamics, in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 681–688.

[6] A. Lovas, M. Rásonyi, Markov chains in random environment with applications in queueing theory and machine learning, To appear in Stochastic Processes and their Applications, arXiv:1911.04377 (2021).

[7] B. Gerencsér, M. Rásonyi, Invariant measures for fractional stochastic volatility models, Preprint, arXiv:2002.04832v1 (2020).

[8] A. Lovas, I. Lytras, M. Rásonyi, S. Sabanis, Taming neural networks with TUSLA: Non-convex learning via adaptive stochastic gradient Langevin algorithms, Preprint, arXiv:2006.14514 (2021).