Neural SDE - Information propagation through the lens of diffusion processes

Stefano Peluchetti
Cogent Labs
speluchetti@cogent.co.jp

Stefano Favaro
University of Torino and Collegio Carlo Alberto
stefano.favaro@unito.it

Abstract

Very deep networks are known to suffer from a number of pathologies which limit their expressive power. We consider parameter distributions that shrink as the number of layers increases in order to recover stochastic processes in the limit of infinite total depth and show that the resulting diffusion processes exhibit desirable properties.

1 Introduction

The focus of this work is on the function-space distribution properties of the output (the last layer) of very deep neural networks. We also focus on standard classes of neural networks, as opposed to proposing structural modifications in order to achieve the desired results. When neural networks are also very wide, recent research [11, 12, 3] shows that i.i.d. initializations with constant variance can result in undesirable properties, even when optimally initialized on the edge of chaos (EOC [3]), such as: i) independence of the output from the input; ii) concentration of the output on restrictive families including constant functions. We illustrate this problem in Fig. 1 where we plot output samples (for a given pre-activation) over input values. In the tanh case the input has no discernible impact on the output and the sampled functions are almost constant. This behavior holds for most smooth activation functions [3]. In the ReLU case the input affects the variance of the output and the function samples are piecewise linear. In both cases, the outputs corresponding to any two inputs end up perfectly correlated.

Intuitively, the issue is the constant level of randomness introduced between subsequent layers. In this work we consider initializations where the parameters' distribution shrinks as the number of layers increases and establish the convergence of the output of residual networks (jointly over multiple inputs) to diffusion processes as the number of layers goes to infinity. In this limit, the output satisfies the desiderata: i) it retains dependency from the input; ii) it doesn't suffer from the perfect correlation constraint; iii) it doesn't collapse to a deterministic function nor does it diverge.

2 Neural SDEs Limits

2.1 Diffusion Processes and SDEs

There are many ways to construct continuous-time stochastic processes as limiting dynamics of discrete-time processes, and in this work we consider the simplest case where the limiting process has continuous paths. In all the neural network architectures considered in this work each layer depends exclusively on the previous one. These two conditions identify diffusion processes [13], which are continuous-time Markov processes with continuous paths, as natural candidates for the limiting process. Diffusion processes arise as solutions to SDEs, stochastic versions of ODEs. A SDE is described by:

$$dx_t = \mu_x(x_t)dt + \sigma_x(x_t)dw_t \tag{1}$$

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

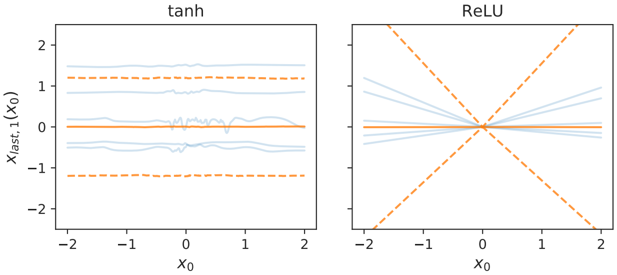

Figure 1: Output samples in blue for 2 fully connected feedforward network with 500 layers of 500 units, input $x_0 \in [-2, 2]$, parameters on EOC; 5%, 50% and 95% quantiles in orange.

where w_t is a driving Brownian motion (BM) [6] of compatible dimensionality. It is easy to give an intuitive characterization of its dynamics. Consider the discretization of (1):

$$x_{t+1} = x_t + \mu_x(x_t)\Delta t + \sigma_x(x_t)\varepsilon_t\sqrt{\Delta t}, \tag{2}$$

where ε_t is a $\mathcal{N}(0, \mathrm{I})$ random vector. The discretization (2) converges (see [7]) to the solution of (1), and we notice the similarity with the Euler discretization of an ODE in the deterministic part of (2).

2.2 Compatible Architectures, Parameter Distributions and Activation Functions

To obtain diffusion limits we show that it is necessary to consider a specific form of residual architectures. Denote with $x_l \in \mathbb{R}^{d_c}$, $l = 1, \ldots, d_l$, the layers of a neural network of d_l layers, and with x_0 its input. Let $\Delta t = 1/d_l$ define an infinitesimal unit of time and let $\Delta x_l = x_{l+1} - x_l$ define the increments of x_l. Due to the continuity of the paths of the limiting diffusion process, we need $\Pr(\|\Delta x_l\| > \varepsilon | x_l) \downarrow 0$ as $\Delta t \downarrow 0$ for any $\varepsilon > 0$, i.e. we require the increments to vanish eventually. It's easy to see that this cannot be achieved in a feed-forward network where $x_{l+1} = \phi(W_l x_l + b_l)$, unless ϕ is linear or the distribution of (W_l, b_l) depends on x. The same holds for residual network architecture (ResNet) originally introduced in [4]. This leaves us with the identity ResNet [5] where:

$$x_{l+1} = x_l + F_l(x_l)$$

In general, each residual block F_l can be composed of multiple stacked layers. Here we consider the case the simplest implementation with shallow residual blocks:

$$\Delta x_l = F_l(x_l) = \phi(W_l x_l + b_l)$$

For shallow residual blocks, the vanishing increments requirement is satisfied by having the distributions of W_l and b_l both concentrate around 0 provided that $\phi(0) = 0$. We require:

Assumption 2.1 (Parameters Distribution and Scaling)**.** *For $l = 0, 1, \ldots$ let:*

$$W_l \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 \Delta t)$$
$$b_l \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_b^2 \Delta t)$$

Assumption 2.2 (Activation Function Regularity)**.** *The function $\phi : \mathbb{R} \to \mathbb{R}$ satisfies: $\phi(0) = 0$, ϕ is continuously differentiable three times on $\mathbb{R}$, its second and third derivatives have at most exponential tails growth, i.e. for some $k > 0$:*

$$\lim_{|x|\uparrow\infty} \frac{|\phi''(x)|}{e^{k|x|}} + \lim_{|x|\uparrow\infty} \frac{|\phi'''(x)|}{e^{k|x|}} < \infty$$

2.3 Joint Diffusion Limits

We now state the main result, which establishes the convergence in distribution of the ResNet to a diffusion process jointly over 2 inputs as $d_l \uparrow \infty$. We index with t instead of l as we need to introduce a continuous-time interpolation (the limiting-process is a continuous-time one). We consider weak solutions and weak uniqueness as we are interested only in the distributional properties of the limiting SDE. The proof and more details these points are in the Appendix. We use x_t and x'_t to denote the joint evolution of the resent for two inputs x_0 and x'_0.

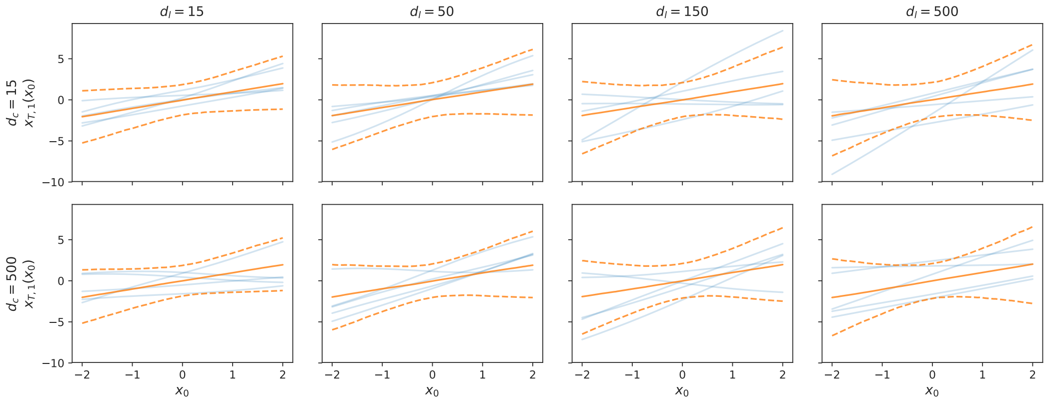

Figure 2: Output samples in blue for a fully connected identity ResNet, different levels of d_l and d_c, input $x_0 \in [-2, 2]$ (i.e. $x_{0,c}$ has the same value x_0 for each $c \in d_c$), parameters according to Assumption 2.1 with $\sigma_b = 1, \sigma_w = 1/\sqrt{d_c}$, $\tanh$ activation, 5%, 50% and 95% quantiles in orange; the scaling on σ_w has a stabilizing effect and has been used to obtain well-defined limits in [8, 11, 12]; we observe a dependency on the input and flexible function shapes (not piecewise linear) compared to Fig. 1 for the $\tanh$ activation; moreover we observe similar distribution properties across different orders of magnitude for d_l and d_c which suggests the existence of a stochastic limit as $d_c \uparrow \infty$.

Theorem 2.1. *The continuous time interpolation x_t, x'_t of x_l, x'_l converges to the weakly unique weak solution of:*

$$d \begin{bmatrix} x_t \\ x'_t \end{bmatrix} = \begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix} dt + \begin{bmatrix} \sigma^2(x_t, x_t) & \sigma^2(x_t, x'_t) \\ \sigma^2(x'_t, x_t) & \sigma^2(x'_t, x'_t) \end{bmatrix}^{1/2} dw_t \qquad (t \in [0, T = 1]) \tag{3}$$

$$\mu(x) = \frac{1}{2}\phi''(0)(\sigma_b^2 + \sigma_w^2 \|x\|^2) 1_{d_c}$$

$$\sigma^2(x, y) = \phi'(0)^2 (\sigma_b^2 + \sigma_w^2 \langle x, y \rangle)\, \mathrm{I}_{d_c}$$

where w_t is a $2d_c$-dimensional BM, 1_{d_c} is a d_c-dimensional unit vector, I_{d_c} is a $d_c \times d_c$ identity matrix. When $\phi''(0) = 0$ (for instance $\phi = \tanh$) the solution is guaranteed not to explode.

An immediate consequence of Theorem 2.1 is that the bivariate distribution of the output given two inputs converges to the transition density of the solution of (3), a stochastic distribution. If $\phi''(0) = 0$, the process is non-explosive (i.e. such transition density is always well-defined). As the integration time is finite, the dependency on the input doesn't vanish in the limit of infinite depth and can be controlled via σ_w and σ_b. As the diffusion matrix squared is non-singular we also don't suffer from the perfect-correlation problem. This satisfies our desiderata i),ii),iii) in the introduction.

We empirically investigate the distribution properties of (3) in Fig. 2 where we plot sample outputs (for a given component $c \in d_c$) over input values. Fig. 2 can be contrasted with Fig. 1. In both figures, each sample is generated by sampling all parameters once and computing the outputs corresponding to each input in a range.

3 Related Work and Conclusions

[14] considers initializations for residual networks which are not encompassed yet by our analysis. Conversely, the residual blocks in [14] cannot be shallow. Some parameters are initialized with a similar scaling, but the residual blocks are multiplied by parameters initialized at 0, hence our desiderata iii) is not satisfied.

The desire of obtaining flexible distributions in function space is especially relevant for Bayesian inference. For instance, a prior model that puts all the probability mass on constant functions cannot fit non-constant functions. While our results are a "pre-requisite" to construct infinitely deep models, in order to obtain competitive performance more attention needs to be paid to architectural choices, also at the level of input and output layers. Our results can be extended without requiring novel theoretical developments to convolutional networks and non i.i.d. parameter distributions.

References

[1] P. Billingsley. *Convergence of Probability Measures*. Wiley-Interscience, 2nd edition, 1999.

[2] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. Wiley-Interscience, 2009.

[3] S. Hayou, A. Doucet, and J. Rousseau. On the impact of the activation function on deep neural networks training. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2672–2680, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[6] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1999.

[7] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, corrected edition, 1992.

[8] R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

[9] D. B. Nelson. Arch models as diffusion approximations. *Journal of econometrics*, 45(1-2):7–38, 1990.

[10] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition, 2003.

[11] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pages 3360–3368, 2016.

[12] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

[13] D. W. Stroock and S. S. Varadhan. *Multidimensional diffusion processes*. Springer, 2006 edition, 2006.

[14] H. Zhang, Y. N. Dauphin, and T. Ma. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019.

A Diffusions and Diffusion Limits

We review results which are useful in establishing convergence to diffusion limits. Let x_l, $l = 0, 1, \dots$ be a generic d-dimensional discrete-time Markov process. Let $\Delta t > 0$ define an infinitesimal unit of time and $\Delta x_l = x_{l+1} - x_l$ define the increments of x_l. We will rely on the following condition where it's implicit that the distribution $p(x_l|x_{l-1})$ depends on Δt.

Assumption A.1 (Infinitesimal Coefficients)**.** *Let x_l, $l = 0, 1, \dots$ be a d-dimensional discrete-time Markov process, and assume that there exist $\mu_x(x) : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma_x^2(x) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ such that:*

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{E}[\Delta x_l | x_l]}{\Delta t} = \mu_x(x_l) \tag{4}$$

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{V}[\Delta x_l | x_l]}{\Delta t} = \sigma_x^2(x_l) \tag{5}$$

$$\lim_{\Delta t \downarrow 0} \frac{\mathbb{E}[\Delta x_l^{2+\delta} | x_l]}{\Delta t} = 0 \tag{6}$$

for some $\delta > 0$, where all convergences are uniform on compacts of $\mathbb{R}^d$ for each vector and matrix component. Moreover $\mu_x(x)$ and $\sigma_x^2(x)$ are continuous and $\sigma_x^2(x)$ admits a Cholesky factorization $\sigma_x(x)$, i.e. $\sigma_x(x)\sigma_x(x)^\top = \sigma_x^2(x)$.

The infinitesimal evolution of diffusion processes is characterized by its infinitesimal mean vector (4) and infinitesimal covariance matrix (5), so the first two limits pinpoint the form of the limiting stochastic evolution. Condition (6) is a technical one in the sense that it allows us to consider the limits (4) and (5) instead of their truncated version [9].

Under additional assumptions, the following result establishes that in the limit x_l can be embedded in a diffusion process.

Theorem A.1. *Let x_l, $l = 0, 1, \dots$ be a d-dimensional discrete-time Markov process, and define the continuous-time process $\overline{x}_t$ on $t \in [0, T]$ by continuous-on-right step-wise-constant interpolation of x_l:*

$$\overline{x}_t = x_l \mathbb{1}_{l\Delta t \le t < (l+1)\Delta t} \quad (l = 0, \dots, d_l, \;\; \Delta t = T/d_l) \tag{7}$$

for some $T > 0$.

Consider the d-dimensional stochastic differential equation (SDE) with initial value x_0, drift vector $\mu_x(x)$ given by (4), and diffusion matrix $\sigma_x(x)$ obtained taking the square root of (5):

$$dx_t = \mu_x(x_t)dt + \sigma_x(x_t)dw_t, \tag{8}$$

where w_t is a d-dimensional Brownian motion (BM) with independent components. Equation (8) is short-hand notation for:

$$x_T = x_0 + \int_0^T \mu_x(x_t)dt + \int_0^T \sigma_x(x_t)dw_t,$$

where T is the integration interval, the first integral is a standard (Riemann) integral, and the second integral is an Ito integral.

Assume that Assumption A.1 holds and that SDE (8) admits an weakly unique and non-explosive weak solution. Then the stochastic process defined by (7) with initial value x_0 converges in law to such solution. Moreover, this result continues to hold when x_0 is an independent and square integrable random variable, in which case the driving BM is independent of x_0. In both cases, the convergence in law is on $D([0,\infty), \mathbb{R}^d)$ the space of functions from $[0,\infty)$ that are continuous from the right with finite left limits endowed with the Skorohod metric [1].

Proof of Theorem A.1. This is [9, Theorem 2.2]: Assumption A.1 and the postulated weakly unique and non-explosive weak solution satisfy all the conditions required for the application of [9, Theorem 2.2]. Note that we use a stronger non-explosivity condition [10]. Alternatively, for this standard result the reader can refer to the monograph [13] on which [9] is based; yet another reference is [2]. □

The reader is referred to the monograph [10] for a gentle introduction to SDEs and Ito integration theory. In Theorem A.1, the continuous-time interpolation $\overline{x}_t$ of x_l is introduced because we are seeking a continuous-time limiting process from a discrete-time process. Observe that the convergence established in Theorem A.1 is strong in the sense that it concerns with the convergence of the distribution of the stochastic process $\overline{x}_t$ as a stochastic object on the whole time interval $[0,T]$ to the distribution of the diffusion limit. For instance, this convergence implies the joint convergence of $\overline{x}_{t_1},\ldots,\overline{x}_{t_n}$ for any collection of times $t_1,\ldots,t_n$, and not only the convergence of the terminal value $\overline{x}_T$.

In Theorem A.1 we postulate the existence and weak uniqueness of a weak solution of the limiting SDE, and its non-explosive behavior. We consider weak solutions and weak (i.e. in law) uniqueness, instead of strong solutions and strong (pathwise) uniqueness, as we are interested exclusively in distributional aspects of the limiting process [9, 10]. Several assumptions exist in the literature in order to guarantee that the additional assumptions of Theorem A.1 are satisfied. The following assumptions suffice for our goals:

Assumption A.2 (Weak Existence and Uniqueness on Compacts)**.** *The functions $\mu_x(x)$ and $\sigma_x(x)$ are twice continuously differentiable.*

Assumption A.3 (Non-explosive Solution)**.** *There exist a finite $C > 0$ such that for each $x \in \mathbb{R}^d$:*

$$\|\mu_x(x)\| + \|\sigma_x(x)\| \leq C(1 + \|x\|)$$

When Assumption A.1 and Assumption A.2 hold (as it will be the case in all the ResNet models considered) but Assumption A.3 doesn't, we still obtain convergence to the unique solution of (8) but x_t might diverge to infinity with positive probability as $d_l \uparrow \infty$.

B Proof of Main Theorem

Lemma B.1. *If ϕ satisfies Assumption 2.2, $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ with $\sigma^2 \leq \sigma_*^2$, $\alpha > 0$, then we can find $M_2(\alpha,\sigma_*^2) < \infty$ and $M_3(\alpha,\sigma_*^2) < \infty$ such that:*

$$\mathbb{E}\left[|\phi''(\varepsilon)|^\alpha\right] \leq M_2(\alpha,\sigma_*^2)$$
$$\mathbb{E}\left[|\phi'''(\varepsilon)|^\alpha\right] \leq M_3(\alpha,\sigma_*^2)$$

Proof. We prove the result only for $\phi''(\varepsilon)$, the case for $\phi'''(\varepsilon)$ being identical. Let L large enough such that $|\phi''(x)| \leq K_1 e^{K_2|x|}$ for $|x| \geq L$ then:

$$\begin{aligned}\mathbb{E}\left[|\phi''(\varepsilon)|^\alpha\right] &= \mathbb{E}\left[|\phi''(\varepsilon)|^\alpha \mathbb{1}_{|\varepsilon|\leq L}\right] + \mathbb{E}\left[|\phi''(\varepsilon)|^\alpha \mathbb{1}_{|\varepsilon|> L}\right] \\ &\leq \sup_{|x|\leq L} |\phi''(x)|^\alpha + K_1^\alpha\, \mathbb{E}[e^{K_2\alpha|\varepsilon|}]\end{aligned}$$

The first term is finite. The fact that the second term can be bounded by a finite and increasing function in σ^2 follows from the symmetry in law of ε and the form of its movement generating function. □

Proof of Theorem 2.1. Let $d = 2d_c$ and

$$x = \begin{bmatrix} x_l \\ x_l' \end{bmatrix} \in \mathbb{R}^d$$
$$b = \begin{bmatrix} b_l \\ b_l \end{bmatrix} \in \mathbb{R}^d$$
$$W = \begin{bmatrix} W_l & 0\,\mathrm{I}_{d_c} \\ 0\,\mathrm{I}_{d_c} & W_l \end{bmatrix} \in \mathbb{R}^{d\times d}$$
$$h = Wx + b \in \mathbb{R}^d$$
$$\mu_x(x) = \begin{bmatrix} \mu(x_l) \\ \mu(x_l') \end{bmatrix}$$
$$\sigma_x^2(x) = \begin{bmatrix} \sigma^2(x_t, x_t) & \sigma^2(x_t, x_t') \\ \sigma^2(x_t', x_t) & \sigma^2(x_t', x_t') \end{bmatrix}$$

where we dropped the dependency on l for notational convenience and reserve subscripts to indexing. A second order Taylor expansion of ϕ around 0 yields for $i = 1, \dots, d$:

$$\frac{\Delta x_i}{\Delta t} = \frac{\phi(h_i\sqrt{\Delta t})}{\Delta t} = \phi'(0)h_i\Delta t^{-1/2} + \frac{1}{2}\phi''(0)h_i^2 + \frac{1}{6}\phi'''(\varepsilon_i)h_i^3\Delta t^{1/2}$$

with $\varepsilon_i \in (-h_i\sqrt{\Delta t}, h_i\sqrt{\Delta t})$. To prove (4) we want to show that $\forall R > 0$:

$$\lim_{\Delta t \downarrow 0} \sup_{\|x\|<R} \left| \mathbb{E}\left[\phi'(0)h_i\Delta t^{-1/2} + \frac{1}{2}\phi''(0)h_i^2 + \frac{1}{6}\phi'''(\varepsilon)h_i^3\Delta t^{1/2}\right] - \mu_x(x)_i \right| = 0.$$

Now, $h_i = W_i x + b_i$ and the distribution assumptions on W and b lead to

$$\mathbb{E}\left[\phi'(0)h_i\Delta t^{-1/2} + \frac{1}{2}\phi''(0)h_i^2\right] = \frac{1}{2}\phi''(0)\,\mathbb{V}[Wx+b]_{i,i} = \mu_x(x)_i$$

It remains to show that

$$\lim_{\Delta t \downarrow 0} \sup_{\|x\|<R} \left|\mathbb{E}\left[\phi'''(\varepsilon_i)h_i^3\right]\right| \Delta t^{1/2} = 0,$$

for which it suffices to show that $\sup_{\|x\|<R} \left|\mathbb{E}\left[\phi'''(\varepsilon_i)h_i^3\right]\right|$ can be bounded by $M(R) < \infty$ uniformly in Δt. By Cauchy–Schwarz:

$$\sup_{\|x\|<R} \left|\mathbb{E}\left[\phi'''(\varepsilon_i)h_i^3\right]\right| \leq \sup_{\|x\|<R} \mathbb{E}\left[\phi'''(\varepsilon_i)^2\right]^{1/2} \sup_{\|x\|<R} \mathbb{E}\left[h_i^6\right]^{1/2}. \tag{9}$$

The constraint $\sup_{\|x\|<R}$ corresponds to a constraint on the variance of h_i hence the second sup is finite. By Lemma B.1 the first sup is finite too and not increasing in Δt as $|\varepsilon_i| \leq \sqrt{\Delta t}|h_i|$ which allows us to produce the desired bound $M(R)$.

Regarding (6), following a first order Taylor expansion of ϕ around 0 we need to show that for $i = 1, \dots, d$ and $R > 0$:

$$\lim_{\Delta t \downarrow 0} \sup_{\|x\|<R} \left| \mathbb{E}\left[\frac{\left(\phi'(0)h_i\Delta t^{1/2} + \frac{1}{2}\phi''(\varepsilon_i)h_i^2\Delta t\right)^4}{\Delta t}\right] \right| = 0$$

with $\varepsilon_i \in (-h_i\sqrt{\Delta t}, h_i\sqrt{\Delta t})$. Note that The term inside the expectation is composed of a sum of terms of the form $kh_i^n\phi''(\varepsilon_i)^m\Delta t^\alpha$ for integers $n, m \geq 0$ and reals $\alpha > 0$, $k \in \mathbb{R}$. This results from repeated applications of the Cauchy–Schwarz inequality and Lemma B.1 as we did previously to prove (4).

Regarding (5), we can compute $\mathbb{E}[\Delta x(\Delta x)']/\Delta t$ instead of $\mathbb{V}[\Delta x]/\Delta t$ as in the infinitesimal limit of $\Delta t \downarrow 0$ the two quantities have to agree due to the convergence of the infinitesimal mean that we have already established. Hence following two first order Taylor expansions of ϕ around 0 we need to show that for $i, j = 1, \dots, d$ and $R > 0$:

$$\lim_{\Delta t \downarrow 0} \sup_{\|x\|<R} \left| \mathbb{E}\left[\frac{\left(\phi'(0)h_i\Delta t^{1/2} + \frac{1}{2}\phi''(\varepsilon_i)h_i^2\Delta t\right)\left(\phi'(0)h_j\Delta t^{1/2} + \frac{1}{2}\phi''(\varepsilon_j)h_j^2\Delta t\right)}{\Delta t}\right] \right.$$
$$\left. - \sigma_x^2(x)_{i,j} \right| = 0$$

with $\varepsilon_i \in (-h_i\sqrt{\Delta t}, h_i\sqrt{\Delta t})$, $\varepsilon_j \in (-h_j\sqrt{\Delta t}, h_j\sqrt{\Delta t})$. The only term inside the expectation not vanishing in Δt is

$$\mathbb{E}[\phi'(0)^2 h_i h_j] = \phi'(0)^2\,\mathbb{V}[Wx+b]_{i,j} = \sigma_x^2(x)_{i,j}$$

The (uniform on compacts) convergence of all terms aside from $\sigma_x(x)_{i,j}^2$ to 0 once again follows from repeated applications of the Cauchy–Schwarz inequality and Lemma B.1.

Finally, the continuity of $\mu_x(x)$ and $\sigma_x(x)$ are a consequence of the continuity of the conditional covariance $\mathbb{V}[Wx+b]$, and as $\mathbb{V}[Wx+b]$ is positive semi-definite so is $\sigma_x^2(x)$ which satisfies the existence of its square root matrix requirement. Again from the properties of $\mathbb{V}$ Assumption A.2 follows, and it can be easily verified that when $\phi''(0) = 0$, i.e. there is no drift, Assumption A.3 is satisfied too. This completes the proof. □