Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

# On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm

**Ery Arias-Castro** eariasca@math.ucsd.edu
*Department of Mathematics*
*University of California, San Diego*
*La Jolla, CA 92093, USA*

**David Mason** davidm@udel.edu
*Department of Applied Economics and Statistics*
*University of Delaware*
*Newark, DE 19717, USA*

**Bruno Pelletier** bruno.pelletier@univ-rennes2.fr
*Département de Mathématiques*
*IRMAR – UMR CNRS 6625*
*Université Rennes II, France*

**Editor:** Mikhail Belkin

## Abstract

We consider the problem of estimating the gradient lines of a density, which can be used to cluster points sampled from that density, for example via the mean-shift algorithm of Fukunaga and Hostetler (1975). We prove general convergence bounds that we then specialize to kernel density estimation.

**Keywords:** mean-shift, gradient lines, density estimation, nonparametric clustering.

## 1. Introduction

Fukunaga and Hostetler (1975) propose clustering points in space according to the gradient ascent flows of the underlying density. Let $f$ be a differentiable density on $\mathbb{R}^d$. Assuming for now that $f$ is known, consider the following scheme. Fix $a > 0$ and, starting at $x_0 \in \mathbb{R}^d$, iteratively define

$$x_\ell = x_{\ell-1} + a\frac{\nabla f(x_{\ell-1})}{f(x_{\ell-1})}, \quad \text{for } \ell \geq 1. \tag{1}$$

When it exists, define $x_\infty = \lim_{\ell\to\infty} x_\ell$. The rationale behind the iterative gradient ascent scheme (1) is to have the sequence $(x_\ell : t \geq 0)$ converge to a local mode of $f$ — representing a cluster center, close in the spirit to Hartigan (1975) — without going through a valley. See Figure 1 and Figure 2 for simple illustrations involving the mixture of two Gaussians in dimensions $d = 1$ and $d = 2$. Now, a sample from $f$, say $X_1, \dots, X_n$, can be clustered by applying the iteration (1) to each $X_i$'s, obtaining a sequence $(X_{i,\ell} : \ell \geq 0)$, and grouping according to the limit $X_{i,\infty}$, meaning that $X_i$ and $X_j$ are grouped together if $X_{i,\infty} = X_{j,\infty}$.

In the same spirit, Cheng et al. (2004) propose to use the gradient ascent lines of $f$, which form gradient trees, to perform a kind of hierarchical clustering of points on the plane. Clustering points according to the local maxima of the underlying density is also advocated

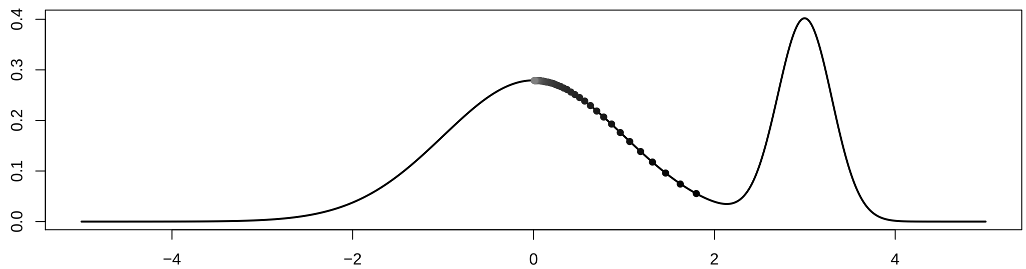


Figure 1: A mixture of two Gaussians in dimension $d = 1$: $f(x) = qg_{0,1}(x) + (1-q)g_{\mu,\sigma}(x)$ where $g_{\mu,\sigma}(x) := e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi}\sigma$, and $q = 0.7$, $\mu = 3$ and $\sigma = 0.3$. The starting point is at $x = 1.8$, and the 50 successive points in the iteration (1) are also plotted. Although the starting point is closer to the peak at $x = 3$, the sequence converges to the peak at $x = 0$.

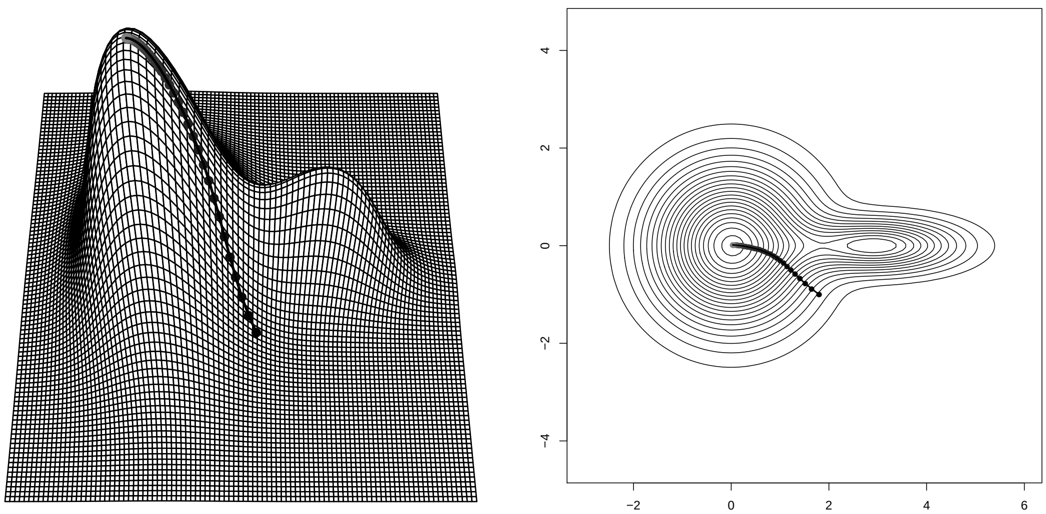


Figure 2: A mixture of two Gaussians in dimension $d = 2$: $f(x, y) = qg_{0,1}(x)g_{0,1}(y) + (1-q)g_{\mu_1,\sigma_1}(x)g_{0,\sigma_2}(y)$ with $q = 0.7$, $\mu_1 = 3$, $\sigma_1 = 1.5$ and $\sigma_2 = 0.5$. The starting point is at $(x, y) = (1.8, -1)$, and the 50 successive points in the iteration (1) are also plotted. Although the starting point is closer to the peak at $(x, y) = (3, 0)$, the sequence converges to the peak at $(x, y) = (0, 0)$.

by Comaniciu and Meer (2002), while an EM-type algorithm for finding the local maxima of the density $f$ is suggested in Carreira-Perpinan and Williams (2003); Carreira-Perpinan (2007); Li et al. (2007).

In practice, the underlying density $f$ is rarely known and has to be estimated. A kernel estimate is used in Fukunaga and Hostetler (1975); Cheng et al. (2004); Li et al. (2007); Comaniciu and Meer (2002). Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a kernel function — an integrable function with $\int_{\mathbb{R}^d} \Phi(x)\mathrm{d}x = 1$ — and for a bandwidth $h > 0$, let $\Phi_h(u) = h^{-d}\Phi(u/h)$. The corresponding kernel estimate for $f$ based on a sample $X_1, \dots, X_n$ is

$$f^{\phi}_{n,h}(x) := \frac{1}{n} \sum_{i=1}^{n} \Phi_h(x - X_i), \tag{2}$$

and if $\Phi$ is differentiable, then we may estimate the gradient of $f$ by

$$\nabla f^{\phi}_{n,h}(x) := \frac{1}{nh} \sum_{i=1}^{n} \nabla \Phi_h(x - X_i).$$

Fukunaga and Hostetler (1975) introduce the term 'mean-shift' when describing the resulting estimate based on the Epanechnikov kernel $\Phi(u) \propto (1 - \|u\|^2)_+$, where $t_+ = \max(t, 0)$ is the positive part of $t \in \mathbb{R}$. Indeed, they show that, in that case,

$$\frac{\nabla f^{\phi}_{n,h}(x)}{f^{\phi}_{n,h}(x)} \propto \frac{1}{|I_{x,h}|} \sum_{i \in I_{x,h}} X_i - x, \quad I_{x,h} := \{i : \|X_i - x\| \le h\}.$$

Cheng (1995) further argues that the gradient ascent algorithm in (1) can be interpreted as a mean-shift when using a spherically symmetric kernel. Indeed, let $\Phi$ be a spherically symmetric kernel on $\mathbb{R}^d$, by which we mean a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ of the form[1] $\Phi(u) = \phi(\|u\|)$, where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-negative function, called the *profile function* in Cheng (1995), that satisfies the following *unit integral* condition

$$\int_{\mathbb{R}^d} \Phi(u) \mathrm{d}u = \omega_d \int_0^{\infty} \phi(r) r^{d-1} \mathrm{d}r = 1, \tag{3}$$

where $\omega_d$ is the surface area of the unit sphere of $\mathbb{R}^d$, and

$$\int_{\mathbb{R}^d} u_i u_j \Phi(u) \mathrm{d}u = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The local average at $x$ is

$$M_{n,h}(x) = \frac{\sum_{i=1}^n X_i \Phi_h(x - X_i)}{\sum_{i=1}^n \Phi_h(x - X_i)} = \frac{1}{n f^{\phi}_{n,h}(x)} \sum_{i=1}^{n} X_i \Phi_h(x - X_i).$$

The *mean shift* at $x$ is defined by

$$T_{n,h}(x) = M_{n,h}(x) - x.$$

This is intimately related to the gradient of another kernel estimate of $f$. To see this, following Cheng (1995), we consider a *shadow kernel* $\Psi$ of $\Phi$, with profile function $\psi$ defined by

$$\Psi(u) = \psi(\|u\|), \qquad \psi(r) = \int_r^{\infty} s \phi(s) \mathrm{d}s. \tag{5}$$

1. Note that Cheng (1995) uses a kernel of the form $\phi(\|u\|^2)$, so the presentation here is little different.

By construction and (3)-(4), $\Psi$ integrates to 1, and is therefore a kernel function; it is also continuously differentiable. Let

$$f_{n,h}^{\psi}(x) = \frac{1}{n}\sum_{i=1}^{n} \Psi_h(x - X_i),$$

which is the kernel estimate of $f$ with kernel $\Psi$ and bandwidth $h$.

**Lemma 1** *(Cheng, 1995) At any point $x$ of $\mathbb{R}^d$, we have*

$$T_{n,h}(x) = h^2 \frac{\nabla f_{n,h}^{\psi}(x)}{f_{n,h}^{\phi}(x)}.$$

Assume that $\nabla\Psi$ is bounded in $\mathbb{R}^d$. Then by the Law of Large Numbers, for each fixed $x \in \mathbb{R}^d$, $f_{n,h}^{\phi}(x) \to f_h^{\phi}(x)$ and $\nabla f_{n,h}^{\psi}(x) \to \nabla f_h^{\psi}(x)$, almost surely as $n \to \infty$, where

$$f_h^{\phi}(x) = \int f(y)\Phi_h(x-y)\mathrm{d}y,$$

and $f_h^{\psi}$ is defined similarly. Furthermore, if $f$ is bounded and continuously differentiable on $\mathbb{R}^d$ with bounded gradient, then $f_h^{\phi}(x) \to f(x)$ and $\nabla f_h^{\psi}(x) \to \nabla f(x)$ as $h \to 0$. Hence, for any $x$ fixed such that $f(x) > 0$,

$$T_{n,h}(x) \to T_h(x) \sim h^2 \nabla \log f(x),$$

as $n \to \infty$ first, followed by $h \to 0$. Following this line of thought, the mean-shift algorithm appears to approximate the gradient ascent scheme (1), with $a = h^2$. The convergence results in Cheng (1995) and Comaniciu and Meer (2002) provide only a very partial mathematical backing to this intuition.

Our contribution is a mathematical proof of consistency for the estimation of gradient ascent lines by the original mean-shift algorithm of Fukunaga and Hostetler (1975). We note that the same approach also applies to the more general mean-shift algorithm of Cheng (1995), and applies directly to the algorithm suggested by Cheng et al. (2004). In detail, let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Starting at $x_0 \in \mathbb{R}^d$, we study the convergence as $a \to 0$ of the sequence

$$x_\ell = x_{\ell-1} + a\nabla f(x_{\ell-1}), \quad \text{for } \ell \geq 1, \tag{6}$$

towards the gradient ascent line of $f$ starting at $x_0$. In particular, we characterize the limit $x_\infty$, providing a consistency result for the clustering algorithm based on the local maxima of $f$. Note that (6) includes (1) by replacing $f$ with $\log f$. We note that such convergence results are available in the rich literature on dynamic systems — see, e.g., Stetter (1973, Sec 3.5), Beyn (1987) and Merlet and Pierre (2010, Sec 2) — and in the literature on convex optimization (where $f$ is convex) — see, e.g., Boyd and Vandenberghe (2004, Sec. 9.3) and Bolte et al. (2010). However, for the general case, we could not find a specific rate of convergence as the one we obtain in (14). Although higher-order discretization schemes can be designed (Stetter, 1973), we focus entirely on the first-order scheme (6). We further elaborate on the literature after stating our main results in Section 2.

Then, given another differentiable function $\hat{f}$, meant to approximate $f$, we compare the sequence $(\hat{x}_\ell)$ to $(x_\ell)$, where

$$\hat{x}_\ell = \hat{x}_{\ell-1} + a\nabla\hat{f}(\hat{x}_{\ell-1}), \quad \text{for } \ell \geq 1, \tag{7}$$

starting at the same point $\hat{x}_0 = x_0$. In particular, when estimating the gradient ascent lines of a density $f$ based on a sample $X_1, \ldots, X_n$, $\hat{f}$ can be taken to be some estimate of $f$, and the gradient ascent sequence defined by $\hat{f} = \log\hat{f}$ (starting at some $x_0$) is compared to that of $f = \log f$. Such approximation results are often called perturbation or stability results in the literature on dynamical systems. See, for example, Hirsch and Smale (1974, Chap 6) or Teschl (2012, Sec 2.5). Most of these results are qualitative (e.g., pertaining to the topology of the gradient flow lines), while the bound we obtain in (15) is quantitative.

Finally, we provide an explicit convergence rate for the case where the density is estimated by kernel convolution. This seems to be new in the literature on the mean-shift algorithm and, more generally, on the estimation of the gradient lines of a density.

The rest of the paper is organized as follows. In Section 2, we establish our main results, one on the convergence of the gradient ascent scheme (6), and another on the stability of smooth flows, relating the gradient flows of $f$ and $\hat{f}$ when these functions are close as $C^2$ functions. In Section 3, we deduce convergence rates for the algorithm of Fukunaga and Hostetler (1975) defined in (1). The technical arguments are given in Section 4.

## 2. Main results

Before stating our main results, we introduce some notations. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we let $f^{(\ell)}(x)$ denote the differential form of $f$ of order $\ell$ at a point $x \in \mathbb{R}^d$, and let $H_f(x)$ denote the Hessian matrix of $f$, when they exist. The differential form $f^{(\ell)}(x)$ of $f$ at $x$ is the multilinear map from $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ ($\ell$ times) to $\mathbb{R}$ defined by

$$f^{(\ell)}(x)[u_1, \ldots, u_\ell] = \sum_{i_1,\ldots,i_\ell=1}^{d} \frac{\partial^\ell f(x)}{\partial x_{i_1} \ldots \partial x_{i_\ell}} u_{1,i_1} \ldots u_{\ell,i_\ell},$$

where, for each $1 \leq i \leq \ell$, $u_i$ has components $u_i = (u_{i,1}, \ldots, u_{i,d})$. Given a multilinear map $L$ of order $\ell$ from $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ to $\mathbb{R}$, we denote by $\|L\|$ its operator norm defined by

$$\|L\| = \sup\left\{|L[u_1, \ldots, u_\ell]| : \|u_1\| = \cdots = \|u_\ell\| = 1\right\}, \tag{8}$$

and writing $L$ as

$$L[u_1, \ldots, u_\ell] = \sum_{i_1,\ldots,i_\ell}^{d} L_{i_1,\ldots,i_\ell} u_{1,i_1} \ldots u_{\ell,i_\ell},$$

we denote by $\|L\|_{\max}$ the norm defined by

$$\|L\|_{\max} = \max\{|L_{i_1 \ldots i_\ell}| \,:\, 1 \leq i_1, \ldots, i_\ell \leq d\}. \tag{9}$$

We note for future reference that

$$\|L\|_{\max} \leq \|L\| \leq d^{\frac{\ell}{2}} \|L\|_{\max}. \tag{10}$$

For a set $S \subset \mathbb{R}^d$, we also define

$$\kappa_\ell(f, S) = \sup_{x \in S} \|f^{(\ell)}(x)\|. \tag{11}$$

Note that $\kappa_\ell(f, S)$ is well-defined and is finite when $f$ is of class $C^\ell$ and $S$ is compact. The *upper level set* of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $b \in \mathbb{R}$ is defined as

$$\mathcal{L}_f(b) = \{x \in \mathbb{R}^d : f(x) \geq b\}. \tag{12}$$

We suppress the dependence on $f$ whenever no confusion is possible.

Recall that a *critical point* of $f$ is a point $x$ at which the gradient of $f$ vanishes, that is, such that $\nabla f(x) = 0$. A *flow line* or *integral curve* of the positive gradient flow of $f$ is a curve $x$ such that $x'(t) = \nabla f(x(t))$. Note that, along any flow line, the value of $f$ increases, that is, the function $t \mapsto f(x(t))$ is increasing with $t$. By the theory of ordinary differential equation, through any point $x_0 \in \mathbb{R}^d$ passes a unique flow line $x(t)$ defined for $t \in [0, t_0)$, where $t_0 > 0$, such that $x(0) = x_0$ (Hirsch et al., 2004, Section 7.2); we say that $x(t)$ is the flow line starting at $x_0$. Let $x^\star$ be a critical point of $f$. We say that $x_0$ is in the attraction basin of $x^\star$ if the flow line $x(t)$ starting at $x_0$ is defined for all $t \geq 0$ and $\lim_{t\to\infty} x(t) = x^\star$. An accumulation point of a sequence of points through an integral curve $x$, i.e., a sequence of the form $\{x(t_n) : t_1 < t_2 < \dots\}$, is called a limit point of $x$. Any limit point of a gradient flow line of $f$ is necessarily a critical point of $f$; see Hirsch et al. (2004, Section 9.3, Proposition, p. 206) and Hirsch et al. (2004, Section 9.3, Theorem, p. 205).

We start by establishing the convergence of the gradient ascent scheme (6) towards the flow lines of the underlying function $f$. Starting from a point $x_0$ in the attraction basin of the location of a stable local maximum $x^\star$, under some conditions stated below, the iteration (6) converges to $x^\star$. In fact, the polygonal line defined by the sequence $(x_\ell)$ is uniformly close to the flow line starting at $x_0$ and ending at $x^\star$. For the definition of a stable equilibrium of a dynamical system, we refer to Hirsch et al. (2004, Section 8.4).

**Theorem 1 (Convergence of gradient ascent)** *Let $f$ be a function of class $C^3$. Let $(x(t) : t \geq 0)$ denote the flow line of $f$ starting at $x_0$ and ending at a local maxima $x^\star$ of $f$. Let $(x_\ell)$ be the sequence defined in* (6) *starting at $x_0$. Then there exists $A = A(x_0, f) > 0$ such that, whenever $0 < a < A$,*

$$\lim_{\ell\to+\infty} x_\ell = x^\star. \tag{13}$$

*Denote by $x_a(t)$ the following polygonal line*

$$x_a(t) = x_{\ell-1} + (t/a - \ell + 1)(x_\ell - x_{\ell-1}), \quad \forall t \in [(\ell-1)a, \ell a).$$

*Assume $H_f(x^\star)$ has all eigenvalues in $(-\overline{\nu}, -\underline{\nu})$ for some $0 < \underline{\nu} < \overline{\nu}$. Then, there exists a $C = C(x_0, f, \underline{\nu}, \overline{\nu}) > 0$ such that, for any $0 < a < A$,*

$$\sup_{t\geq 0} \|x_a(t) - x(t)\| \leq C a^\delta, \quad \delta := \frac{\underline{\nu}}{\underline{\nu} + \overline{\nu}}. \tag{14}$$

We mention the convergence result (Comaniciu and Meer, 2002, Th 1), which essentially says that, when $f$ is a kernel density estimator with bandwidth $h$ as in (2), the sequence $(x_\ell)$

in (6) with choice $a = h^2$ converges and $(f(x_\ell))$ is monotone nondecreasing. In the literature on dynamical systems, the convergence result (13) is proved in (Merlet and Pierre, 2010, Sec 2), together with convergence rates, but under slightly different conditions; in particular, $f$ is assumed to have compact upper level sets. Beyn (1987) compares the discrete and continuous trajectories under milder conditions, but only at a discrete grid of time points, and does so assuming that the starting point $x_0$ is sufficiently close to the corresponding stationary point $x^\star$. Moreover, the starting point of the discrete and continuous trajectories in Beyn (1987) are potentially different. In fact, Beyn (1987) refers the reader to (Stetter, 1973) — which we mentioned earlier — for the case where the starting points may be taken to be the same.

Next, we establish a stability result for flows of smooth functions. In words, under some conditions made precise below, when $f$ and $\hat{f}$ are close as $C^2$ functions, then their flow lines are also close. Denote by $B(x,r)$ the open ball of radius $r$ centered at $x$ and by $\bar{B}(x,r)$ its closure.

**Theorem 2 (Stability of smooth flows)** *Suppose $f$ and $\hat{f}$ are of class $C^3$. Let $(x(t) : t \geq 0)$ be a flow line of $f$ starting at $x_0$ and ending at $x^\star$ where $H_f(x^\star)$ has all eigenvalues in $(-\overline{\nu}, -\underline{\nu})$ for some $0 < \underline{\nu} < \overline{\nu}$. Let $\hat{x}(t)$ be the flow line of $\hat{f}$ starting at $x_0$. Let $S = \mathcal{L}(f(x_0)/2) \cap \bar{B}(x_0, 3r_0)$ where $r_0 = \max_t \|x(t) - x_0\|$, and define*

$$\eta_m = \sup_{x \in S} \|f^{(m)}(x) - \hat{f}^{(m)}(x)\|.$$

*Then there is a constant $C = C(f, x_0, \underline{\nu}, \overline{\nu}) \geq 1$ such that, when $\max(\eta_0, \eta_1, \eta_2) \leq 1/C$ and $\eta_3 \leq C$, $\hat{x}(t)$ is defined for all $t \geq 0$ and*

$$\sup_{t \geq 0} \|x(t) - \hat{x}(t)\| \leq C \max\left\{\sqrt{\eta_0}, \eta_1^\delta\right\}, \tag{15}$$

*where $\delta$ is defined in (14).*

Stability results tend to be qualitative in the literature on dynamical systems. However, to establish the bound above, we do use a well-known quantitative result. See Lemma 7, which we took from Hirsch et al. (2004, Sec 17.5).

Combining Theorems 1 and 2, we arrive at the following bound for approximating the flow lines of a function $f$ by the polygonal line obtained from the gradient ascent algorithm (7) based on an approximation $\hat{f}$ to $f$.

**Corollary 1** *In the context of Theorem 2, for $a > 0$, define*

$$\hat{x}_a(t) = \hat{x}_{\ell-1} + (t/a - \ell + 1)(\hat{x}_\ell - \hat{x}_{\ell-1}), \quad \forall t \in [(\ell-1)a, \ell a), \tag{16}$$

*where $(\hat{x}_\ell)$ is defined in (7). Then there is a constant $C = C(f, x_0, \underline{\nu}, \overline{\nu}) \geq 1$ such that, when $\max(\eta_0, \eta_1, \eta_2) \leq 1/C$ and $\eta_3 \leq C$,*

$$\sup_{t \geq 0} \|\hat{x}_a(t) - x(t)\| \leq C\left[a^\delta + \max\left\{\sqrt{\eta_0}, \eta_1^\delta\right\}\right], \tag{17}$$

*where $\delta$ is defined in (14).*

Note that the exponent $\delta$ which appears in these results only depends on the ratio $\overline{\nu}/\underline{\nu}$ which is a lower bound on the condition number of $H_f(x^\star)$. But the constants in Theorems 1 and 2 depend on $\underline{\nu}$ and $\overline{\nu}$ not only through their ratio.

We note that Beyn (1987) establishes a result similar to Corollary 1 under milder assumptions. Indeed, just as we do here, he studies how the discrete system (7) approximates the continuous system

$$x'(t) = \nabla f(x(t)),$$

when the functions $\hat{f}$ and $f$ may differ. He bounds the difference between the discrete and continuous trajectories, with possibly different starting points, over a discrete grid of time points, assuming the starting point $x_0$ is close enough to $x^\star$. He also assumes that $\nabla \hat{f}(x^\star) = 0$, which simplifies the analysis a fair amount. With these working assumptions, his bound is in $\kappa_2 a + \eta_1$ — see Equation (3.5) there. His method of proof is based on the theory of stable (solution) manifolds (Irwin, 1980, Chap 4). Our approach is more elementary and we do not know whether this more sophisticated approach has the potential to improve on ours.

We emphasize that Theorems 1 and 2, and their combined fruit in Corollary 1, are designed to establish our result on the uniform consistency of gradient line estimators based on kernel density estimators as stated in Theorem 3 in the next section.

## 3. The estimation of gradient lines of a density

Let $\hat{f}_{n,h}$ be the kernel density estimate of $f$ in (2) with kernel $\Phi$ and bandwidth $h$. Sharp almost-sure convergence rates in the uniform norm of kernel density estimates have been obtained by several authors, for example Einmahl and Mason (2000); Giné and Guillou (2002); Einmahl and Mason (2005). Using the recent results of Mason and Swanepoel (2011) and Mason (2012), we derive strong uniform norm convergence rates for $\hat{f}_{n,k}$ and its derivatives.

We first control the bias component.

**Lemma 2** *Assume $\Phi$ is nonnegative, $C^3$ on $\mathbb{R}^d$ with all partial derivatives up to order 3 vanishing at infinity, and satisfies*

$$\int_{\mathbb{R}^d} \Phi(x)\mathrm{d}x = 1, \quad \int_{\mathbb{R}^d} x\Phi(x)\mathrm{d}x = 0 \quad \textit{and} \quad \int_{\mathbb{R}^d} \|x\|^2 \Phi(x)\mathrm{d}x < \infty. \tag{18}$$

*Then for any $C^3$ density $f$ on $\mathbb{R}^d$ with bounded derivatives up to order 3, there is a constant $C > 0$ such that*

$$\sup_{x \in \mathbb{R}^d} \left\| \mathbb{E}\big[\hat{f}^{(\ell)}_{n,h}(x)\big] - f^{(\ell)}(x) \right\| \le C h^{(3-\ell)\wedge 2}, \quad \forall 0 \le \ell \le 3. \tag{19}$$

Next, we control the variance component. For this, we apply the main result of Mason and Swanepoel (2011). See also Theorem 4.1 with Remark 4.2 in Mason (2012).

**Lemma 3** *Suppose that $\Phi$ is of the form $\Phi : (x_1, \dots, x_d) \mapsto \prod_{k=1}^d \phi_k(x_k)$, and that each $\phi_k$ is nonnegative, integrates to 1, and is $C^3$ on $\mathbb{R}$ with derivatives up to order 3 being of* *bounded variation and in $L_1(\mathbb{R}^d)$. Then, for any bounded density $f$ on $\mathbb{R}^d$, there exists a $0 < b_0 < 1$ such that*

$$\limsup_{n\to\infty} \sup_{\frac{\log n}{n} \le h^d \le b_0} \sup_{x\in\mathbb{R}^d} \sqrt{\frac{nh^{d+2\ell}}{\log n}} \left\| \hat{f}^{(\ell)}_{n,h}(x) - \mathbb{E}\big[\hat{f}^{(\ell)}_{n,h}(x)\big] \right\| < \infty, \quad \forall 0 \le \ell \le 3, \quad a.s. \tag{20}$$

It is straightforward to design a kernel that satisfies the conditions of Lemmas 2 and 3. In fact, the Gaussian kernel $\Phi(x) = (2\pi)^{-d/2} \exp(-\|x\|^2/2)$ is such a kernel.

Assuming that (20) holds and applying Corollary 1, we deduce a convergence result for the mean-shift algorithm of Fukunaga and Hostetler (1975). We note that a similar result holds for the simpler gradient ascent method of Cheng et al. (2004).

**Theorem 3** *Consider a density $f$ satisfying the conditions of Lemma 2. Suppose $\hat{f}_{n,h}$ is a kernel estimator of $f$ of the form (2), where $\Phi$ satisfies the conditions of Lemmas 2 and 3. Let $(x(t) : t \ge 0)$ be the flow line of $f$ starting at a point $x_0$ with $f(x_0) > 0$, ending at a point $x^\star$ where $H_f(x^\star)$ has all eigenvalues in $(-\overline{\nu}, -\underline{\nu})$ for some $0 < \underline{\nu} < \overline{\nu}$. For $a > 0$, define $(\hat{x}_a(t) : t \ge 0)$ by*

$$\hat{x}_a(t) = \hat{x}_{\ell-1} + (t/a - \ell + 1)(\hat{x}_\ell - \hat{x}_{\ell-1}), \quad \forall t \in [(\ell-1)a, \ell a),$$

*where*

$$\hat{x}_\ell = \hat{x}_{\ell-1} + a \frac{\nabla \hat{f}_{n,h}(\hat{x}_{\ell-1})}{\hat{f}_{n,h}(\hat{x}_{\ell-1})}, \quad \textit{for } \ell \ge 1.$$

*Suppose that $h \to 0$ and $\frac{nh^{d+6}}{\log n} \to \infty$. Then there exists a constant $C > 0$ such that, with probability one, for all $n$ large enough,*

$$\sup_{t\ge 0} \|\hat{x}_a(t) - x(t)\| \le C \left[a + h^2\right]^\delta, \quad \delta := \frac{\underline{\nu}}{\underline{\nu} + \overline{\nu}}. \tag{21}$$

The approximation error decreases as the discretization step $a$ gets smaller, simply because it controls the precision of the (discrete) gradient ascent scheme (7). We made this precise in Theorem 1. However, as $a$ gets smaller, the computational burden of running this gradient ascent scheme to its limit becomes heavier. So there is a compromise between (statistical and numerical) estimation and computational complexity. That said, choosing $a$ smaller (in order of magnitude) than $h^2$ does not improve our bound (21). When $a$ is that small, the main source of error comes from estimating the density, rather than the accuracy of the gradient ascent scheme, and the resulting rate is

$$\sup_{t\ge 0} \|\hat{x}_a(t) - x(t)\| \le C\gamma_n \left(\frac{\log n}{n}\right)^{2\delta/(d+6)},$$

for any choice of sequence $(\gamma_n)$ with $\gamma_n \to \infty$. We note that faster rates are possible for densities that are $C^k$ for $k > 3$, since they can be estimated more accurately by a higher order kernel (Devroye and Gyorfi, 1985). We also mention that the curse of dimensionality is at play here since we are estimating a nonparametric density.

## 4. Proofs

We start in Section 4.1 with some auxiliary results that will be used in the proofs of our main results. Theorem 1 and Theorem 2 are proved in Sections 4.2 and 4.3 respectively. We prove Lemma 2 and Lemma 3 in Sections 4.4 and 4.5, and then Theorem 3 in Section 4.6.

### 4.1 Preliminary results

The following is a discrete version of Gronwall's lemma. The proof is straightforward and left to the reader.

**Lemma 4** *Let $(y_\ell : \ell \geq 0)$ be a sequence of non-negative real numbers such that*

$$y_{\ell+1} \leq Q_1 + (1 + Q_2) y_\ell.$$

*Then*

$$y_\ell \leq y_0 e^{Q_2 \ell} + \frac{e^{Q_2 \ell} - 1}{Q_2} Q_1.$$

The result below is on the behavior of the upper level set near a stable local maximum.

**Lemma 5** *Suppose that $f$ is of class $C^3$. Let $x^\star$ be the location of a stable local maxima of $f$ where $H_f(x^\star)$ has all eigenvalues in $(-\overline{\nu}, -\underline{\nu})$ with $\overline{\nu} > \underline{\nu} > 0$. For $\epsilon > 0$, let $\mathcal{C}(\epsilon)$ be the connected component of $\mathcal{L}_f(f(x^\star) - \epsilon)$ that contains $x^\star$. Then there is a constant $C_5 = C_5(f, x^\star)$ such that*

$$\bar{B}(x^\star, \sqrt{2\epsilon/\overline{\nu}}) \subset \mathcal{C}(\epsilon) \subset \bar{B}(x^\star, \sqrt{2\epsilon/\underline{\nu}}), \quad \textit{for all } \epsilon \leq C_5, \tag{22}$$

*and*

$$f(x^\star) - f(x) \leq \frac{\overline{\nu}}{2} \|x - x^\star\|^2, \quad \textit{for all } x \textit{ such that } \|x - x^\star\| \leq \sqrt{2C_5/\overline{\nu}}. \tag{23}$$

**Proof** Fix $r > 0$. Let $\mathbf{H}$ and $\kappa_3$ be short for $H_f(x^\star)$ and $\kappa_3(f, \bar{B}(x^\star, r))$, respectively. Let $\underline{\nu} < \underline{\nu}' < \overline{\nu}' < \overline{\nu}$ be such that $H_f(x^\star)$ has all eigenvalues in $[-\overline{\nu}', -\underline{\nu}']$. First, we prove (22). A Taylor development of $f$ at $x \in \bar{B}(x^\star, r)$ gives

$$f(x) = f(x^\star) + \frac{1}{2}\mathbf{H}[x - x^\star, x - x^\star] + R(x, x^\star), \quad \text{with } |R(x, x^\star)| \leq \frac{\kappa_3}{6}\|x - x^\star\|^3. \tag{24}$$

When $x \in \bar{B}\left(x^\star, r\right)$, using the Taylor expansion (24), we get that

$$\begin{aligned} f(x) &\leq f(x^\star) - \frac{\underline{\nu}'}{2}\left\|x^\star - x\right\|^2 + \frac{\kappa_3}{6}\left\|x^\star - x\right\|^3 \\ &\leq f(x^\star) - \frac{\underline{\nu}}{2}\left\|x^\star - x\right\|^2 \end{aligned}$$

when $\|x^\star - x\| \leq \xi_1 := \frac{3(\underline{\nu}' - \underline{\nu})}{\kappa_3} \wedge r$. Fix $0 < \epsilon < \frac{\underline{\nu}\xi_1^2}{2}$ so that $\sqrt{\left(\frac{2\epsilon}{\underline{\nu}}\right)} < \xi_1$. We then have $f(x) < f(x^\star) - \epsilon$ when $\sqrt{\left(\frac{2\epsilon}{\underline{\nu}}\right)} < \|x^\star - x\| \leq \xi_1$. This implies that

$$\mathcal{L}_f(f(x^\star) - \epsilon) \subset \bar{B}\big(x^\star, \sqrt{\left(\tfrac{2\epsilon}{\underline{\nu}}\right)}\big) \cup \bar{B}\big(x^\star, \xi_1\big)^{\mathsf{c}},$$

and since the two sets on the right-hand side are disconnected, while $\mathcal{C}(\epsilon)$ is connected and contains $x^\star$, necessarily, $\mathcal{C}(\epsilon) \subset B\big(x^\star, \sqrt{(\frac{2\epsilon}{\underline{\nu}})}\big)$.

We also get using (24) that

$$\begin{aligned} f(x) &\geq f(x^\star) - \frac{\overline{\nu}'}{2}\left\|x^\star - x\right\|^2 - \frac{\kappa_3}{6}\left\|x^\star - x\right\|^3 \\ &\geq f(x^\star) - \frac{\overline{\nu}}{2}\left\|x^\star - x\right\|^2 \end{aligned}$$

when $\|x^\star - x\| \leq \xi_2 := \frac{3(\overline{\nu} - \overline{\nu}')}{\kappa_3} \wedge r$. Fix $0 < \epsilon < \frac{\overline{\nu}\xi_2^2}{2}$ so that $\sqrt{(\frac{2\epsilon}{\overline{\nu}})} < \xi_2$. Then whenever $\|x^\star - x\| \leq \sqrt{(\frac{2\epsilon}{\overline{\nu}})}$, we have $f(x) \geq f(x^\star) - \epsilon$. Reasoning as above, we obtain $B\big(x^\star, \sqrt{(\frac{2\epsilon}{\overline{\nu}})}\big) \subset \mathcal{C}\left(\epsilon\right)$.

Therefore, by choosing $C_5 < \xi_1 \wedge \xi_2$, we see that (22) holds. Note that $\xi_1$ and $\xi_2$ depend on $r$. Since we do not need an explicit value for the constant $C_5$, we leave $r > 0$ arbitrarily fixed.

The bound (23) is a direct consequence of (22). ■

Next is a result establishing exponential convergence rates for the gradient flow of a smooth function ending at a stable local maximum.

**Lemma 6** *Suppose that $f$ is of class $C^3$. Let $\{\gamma(t) : t \geq 0\}$ be the flow line of $f$ starting at $x_0$ and ending at $x^\star$ where $H_f(x^\star)$ has all its eigenvalues in $(-\infty, -\underline{\nu})$, with $\underline{\nu} > 0$. Then, there is $C_6 = C_6(f, x_0)$ such that, for all $t \geq 0$,*

$$\|\gamma(t) - x^\star\| \leq C_6 e^{-\underline{\nu} t}, \tag{25}$$

*and*

$$f(x^\star) - f(\gamma(t)) \leq C_6 e^{-2\underline{\nu} t}. \tag{26}$$

**Proof** Note that since $\gamma$ has beginning and ending points, $\{\gamma(t) : t \geq 0\}$ is bounded. Let $r_0 > 0$ be such that $\{\gamma(t) : t \geq 0\}$ is contained in the ball $\bar{B}(x^\star, r_0)$. Let $\mathbf{H}$ and $\kappa_3$ be short for $H_f(x^\star)$ and $\kappa_3(f, \bar{B}(x^\star, r_0))$, respectively. A Taylor development of $\nabla f$ at $x \in \bar{B}(x^\star, r_0)$ gives

$$\nabla f(x) = \mathbf{H}(x - x^\star) + R(x, x^\star),$$

with

$$\|R(x, x^\star)\| \leq \kappa_3 \tfrac{\sqrt{d}}{2} \|x - x^\star\|^2.$$

Therefore, we have,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\gamma(t) - x^\star\right) - \mathbf{H}\left(\gamma(t) - x^\star\right) = R\left(\gamma(t), x^\star\right),$$

and so, since $\gamma(0) = x_0$, $\gamma$ satisfies the relation

$$\gamma(t) - x^\star = e^{t\mathbf{H}}(x_0 - x^\star) + \int_0^t e^{(t-s)\mathbf{H}} R\left(\gamma(s), x^\star\right) \mathrm{d}s.$$

Since all the eigenvalues of $\mathbf{H}$ are in $(-\infty, -\underline{\nu})$, there is $\nu > \underline{\nu}$ such that we have

$$\left\|e^{\alpha \mathbf{H}}\right\| \leq e^{-\nu\alpha}, \quad \text{for all } \alpha > 0.$$

Then,

$$\|\gamma(t) - x^\star\| \leq e^{-\nu t}\|x_0 - x^\star\| + \kappa_3 \tfrac{\sqrt{d}}{2} \int_0^t e^{-\nu(t-s)} \|\gamma(s) - x^\star\|^2 \mathrm{d}s.$$

Set $u(t) = e^{\nu t}\|\gamma(t) - x^\star\|$ and $U(t) = \|x_0 - x^\star\| + \kappa_3 \frac{\sqrt{d}}{2} \int_0^t e^{\nu s}\|\gamma(s) - x^\star\|^2 \mathrm{d}s$. Then $u(t) \leq U(t)$ and $U'(t) = \kappa_3 \frac{\sqrt{d}}{2} e^{-\nu t} u^2(t)$, so

$$\frac{U'(t)}{U(t)} = \kappa_3 \tfrac{\sqrt{d}}{2} e^{-\nu t} u(t) \frac{u(t)}{U(t)} \leq \kappa_3 \tfrac{\sqrt{d}}{2} e^{-\nu t} u(t) = \kappa_3 \tfrac{\sqrt{d}}{2} \|\gamma(t) - x^\star\|.$$

But since $\gamma(t) \to x^\star$ as $t \to \infty$, there exists $t_0 > 0$ such that $\|\gamma(t) - x^\star\| \leq \frac{2(\nu - \underline{\nu})}{\kappa_3 \sqrt{d}}$ for all $t \geq t_0$. By integrating between $t_0$ and $t$, we deduce that

$$\log U(t) \leq \log U(t_0) + (\nu - \underline{\nu})(t - t_0),$$

and so

$$\|\gamma(t) - x^\star\| = e^{-\nu t} u(t) \leq e^{-\nu t} U(t) \leq Q_0 e^{-\underline{\nu} t}, \quad \text{for all } t \geq t_0,$$

with $Q_0 := U(t_0) e^{-(\nu - \underline{\nu}) t_0}$. For $t < t_0$, we simply have $\|\gamma(t) - x^\star\| \leq Q_1 e^{-\underline{\nu} t}$, where $Q_1 = \max_{0 \leq t \leq t_0} \|\gamma(t) - x^\star\| e^{\underline{\nu} t}$. Therefore (25) holds with the constant $Q_2 = \max\{Q_0, Q_1\}$.

We now turn to proving (26). For any $x$ in $\bar{B}(x^\star, r_0)$, we have

$$f(x) = f(x^\star) + \frac{1}{2}\mathbf{H}[x - x^\star, x - x^\star] + R(x, x^\star),$$

for all $x$ in $\bar{B}(x^\star, r_0)$, where $R$ is a different function (now real valued) satisfying

$$|R(x, x^\star)| \leq \tfrac{\kappa_3}{6} \|x - x^\star\|^3.$$

Then

$$\begin{aligned} f(x^\star) - f(\gamma(t)) &\leq \tfrac{1}{2}\|\mathbf{H}\| \, \|\gamma(t) - x^\star\|^2 + \tfrac{\kappa_3}{6} \|\gamma(t) - x^\star\|^3 \\ &\leq (\tfrac{1}{2}\|\mathbf{H}\| + Q_3) Q_2^2 e^{-2\underline{\nu} t}, \end{aligned}$$

where $Q_3 = \frac{\kappa_3}{6} \max_{t \geq 0} \|\gamma(t) - x^\star\|$ and we applied (25) in the second line with $Q_2$ defined above. Therefore, (26) holds with the constant $Q_4 := (\|\mathbf{H}\|/2 + Q_3) Q_2^2$.

We then take $C_6 = \max(Q_2, Q_4)$. ∎

The following, adapted from Hirsch et al. (2004, Sec 17.5), is a stability result for autonomous gradient flows.

**Lemma 7** *Suppose $f$ and $g$ are of class $C^3$. Let $x_0 \in \mathbb{R}^d$, and suppose that*

$$\|\nabla f(x) - \nabla g(x)\| < \eta, \quad \forall x \in \mathcal{S} := \mathcal{L}_f(f(x_0)) \cup \mathcal{L}_g(g(x_0)).$$

*Let $\kappa$ be a Lipschitz constant for $\nabla f$ on $\mathcal{S}$. Let $(x(t) : t \geq 0)$ and $(y(t) : t \geq 0)$ be the flow lines of $f$ and $g$ starting at $x_0$, supposed to be defined on $[0, \infty)$. Then,*

$$\|x(t) - y(t)\| \leq \frac{\eta}{\kappa}\left[e^{\kappa t} - 1\right], \quad \forall t \geq 0.$$

Next is a result on the stability of local maxima.

**Lemma 8** *Suppose $f$ and $g$ are of class $C^3$, and have local maxima at $x$ and $y$, respectively, with $H_f(x)$ having all eigenvalues in $(-\infty, -\nu]$ for some $\nu > 0$. Then for any $C_8 \geq \max\left\{1, \frac{2}{\sqrt{\nu}}, \frac{4\kappa}{3\nu}\right\}$, where $\kappa = \max\left(\kappa_3(f, \bar{B}(x,1)), \kappa_3(g, \bar{B}(y,1))\right)$,*

$$\|x - y\| \leq 1/C_8 \quad \Rightarrow \quad \|x - y\| \leq C_8\big(\|f(x) - g(x)\| + \|f(y) - g(y)\|\big)^{1/2}. \tag{27}$$

**Proof** Let $\mathbf{H}_f$ and $\mathbf{H}_g$ be short for $H_f(x)$ and $H_g(y)$, respectively. We develop $f$ and $g$ around $x$ and $y$, respectively. Assuming $\|x - y\| \leq 1$, we have

$$\begin{aligned} f(y) &= f(x) + \tfrac{1}{2}\mathbf{H}_f[x-y, x-y] + R_f(x,y), && \text{with} \quad |R_f(x,y)| \leq \tfrac{\kappa}{6}\|x-y\|^3; \\ g(x) &= g(y) + \tfrac{1}{2}\mathbf{H}_g[x-y, x-y] + R_g(x,y), && \text{with} \quad |R_g(x,y)| \leq \tfrac{\kappa}{6}\|x-y\|^3. \end{aligned}$$

Summing these two equalities, we obtain

$$\frac{1}{2}(\mathbf{H}_f + \mathbf{H}_g)[x-y, x-y] = f(y) - g(y) + g(x) - f(x) - R_f(x,y) - R_g(x,y).$$

By the triangle inequality and the fact that $\mathbf{H}_g$ is negative semidefinite,

$$\nu\|x-y\|^2 \leq \|(\mathbf{H}_f + \mathbf{H}_g)[x-y, x-y]\| \leq 2\|f(x) - g(x)\| + 2\|f(y) - g(y)\| + \tfrac{2\kappa}{3}\|x-y\|^3.$$

When $\|x - y\| \leq \min\left(\frac{3\nu}{4\kappa}, 1\right)$, we have $\nu\|x-y\|^2 - \frac{2\kappa}{3}\|x-y\|^3 \geq \frac{\nu}{2}\|x-y\|^2$, and therefore

$$\|x-y\|^2 \leq \tfrac{4}{\nu}\left(\|f(x) - g(x)\| + \|f(y) - g(y)\|\right),$$

and from this we conclude that (27) holds with $C_8 = \max(\sqrt{(\frac{4}{\nu})}, \frac{4\kappa}{3\nu}, 1)$. ■

### 4.2 Proof of Theorem 1

Below, $C_m$ refers to the constant defined in Lemma $m$.

We assume that $x_0$ is not a critical point of $f$, for otherwise $x_0 = x^\star$ and there is nothing to prove. Let $t_\ell = a\ell$, which is the time at which the polygonal line $x_a(t)$ passes through $x_\ell$. Let $\mathcal{L}_0$ be short for $\mathcal{L}_f(f(x_0))$. Note that $(x(t) : t \geq 0)$ is bounded since $x$ is a continuous flow line with a beginning and ending points. Let $r_0$ be large enough that $(x(t) : t \geq 0) \subset \bar{B}(x_0, r_0)$.

**Claim.** *Without loss of generality, we may assume that $\mathcal{L}_0$ is bounded.* To see this, suppose the result is true when $\mathcal{L}_0 \subset \bar{B}(x_0, 3r_0)$. We shall prove that it remains true when $\mathcal{L}_0 \nsubseteq \bar{B}(x_0, 3r_0)$. Given such a situation, build another function $\tilde{f}$ in such a way that $\tilde{f}$ is $C^3$ on $\mathbb{R}^d$ with $\tilde{f}(x) = f(x)$ for all $x \in \bar{B}(x_0, 2r_0)$ and $\tilde{f}(x) < f(x_0)$ for $x \notin \bar{B}(x_0, 3r_0)$, so that $\mathcal{L}_{\tilde{f}}(\tilde{f}(x_0)) \subset \bar{B}(x_0, 3r_0)$. To verify that such a function exists, consider the smoothing function $s : \mathbb{R}^d \to \mathbb{R}$ defined by

$$s(x) = \frac{1}{\int_{\bar{B}(0,1)} e^{-1/(1-\|x\|^2)}\mathrm{d}x} e^{-1/(1-\|x\|^2)} \mathbf{1}_{B(0,1)}(x), \quad x \in \mathbb{R}^d,$$

and its dilated versions $s_a$ defined by $s_a(x) = a^{-d} s(x/a)$ for $a > 0$, where $\mathbf{1}_{\bar{B}(0,1)}(x) = 1$ if $x \in \bar{B}(0,1)$ and 0 otherwise. Define the function $g$ by $g(x) = \mathbf{1}_{\bar{B}(0,5r_0/2)} \star s_{r_0/2}(x - x_0)$. Then $g$ is of class $C^\infty$, $g(x) = 1$ for $x \in \bar{B}(x_0, 2r_0)$, $g(x) = 0$ if $x \notin \bar{B}(x_0, 3r_0)$, and $0 < g(x) < 1$ when $2r_0 < \|x - x_0\| < 3r_0$. Then we may take $\tilde{f} = fg$.

Therefore, (13) and (14) hold for $\tilde{f}$, for constants $\tilde{A}$ and $\tilde{C}$, with the same exponent $\delta$ as given in (14). Denote by $\tilde{x}$ and $\tilde{x}_a$ the flow line and polygonal curve constructed from $\tilde{f}$ in the same way $x$ and $x_a$ are from $f$. Then, assuming $\tilde{C} a^\delta \le r_0$, we see by the triangle inequality that $\tilde{x}(t)$ and $\tilde{x}_a(t)$ are determined by $\tilde{f}$ restricted to $\bar{B}(x_0, 2r_0)$, and since $\tilde{f}$ coincides with $f$ there, $x(t) = \tilde{x}(t)$ and $x_a(t) = \tilde{x}_a(t)$, so that (13) and (14) are valid for $f$ if $a \le \min\{\tilde{A}, (r_0/\tilde{C})^{1/\delta}\}$.

From now on, we assume that $\mathcal{L}_0$ is bounded. Note that $\mathcal{L}_0$ is also closed since $f$ is continuous, so in fact $\mathcal{L}_0$ is compact. Let

$$S = \mathcal{L}_0 \oplus \bar{B}(0, \kappa_1(f, \mathcal{L}_0)) =: \{x \in \mathbb{R}^d \,:\, \mathrm{dist}(x, \mathcal{L}_0) \le \kappa_1(f, \mathcal{L}_0)\}, \tag{28}$$

where $\mathrm{dist}(x, \mathcal{L}_0) = \inf\{\|x - y\| \,:\, y \in \mathcal{L}_0\}$. For any $0 \le \ell \le 3$, let $\kappa_\ell = \kappa_\ell(f, S)$, where $S$ is defined in (28). For any $x \in \mathbb{R}^d$, let

$$\kappa_2(x) = \kappa_2(f, \bar{B}(x, \|\nabla f(x)\|)) = \sup\{\|f^{(2)}(y)\| \,:\, y \in \bar{B}(x, \|\nabla f(x)\|)\}. \tag{29}$$

Notice that $\mathcal{L}_0 \subset S$ and that, by construction, $\bar{B}(x, \|\nabla f(x)\|) \subset S$ for any $x \in \mathcal{L}_0$. Hence, $\kappa_2(x) \le \kappa_2$ for all $x$ in $\mathcal{L}_0$.

**Claim**. *For any $x \in \mathbb{R}^d$ with $\nabla f(x) \neq 0$ and any $0 \le b \le 1 \wedge (2\sqrt{d}\kappa_2(x))^{-1}$, we have $f(x + b\nabla f(x)) > f(x)$ and $f$ is increasing along the line segment $[x, x + b\nabla f(x)]$.* Using a Taylor expansion of $f$ at $x$, we have

$$f(x + b\nabla f(x)) = f(x) + b\|\nabla f(x)\|^2 + R(x, b),$$

where $|R(x,b)| \le \frac{1}{2} b^2 \kappa_2(x) \|\nabla f(x)\|^2 \le \frac{b}{2}\|\nabla f(x)\|^2$, since $b \le (2\sqrt{d}\kappa_2(x))^{-1} \le \kappa_2^{-1}(x)$. Then

$$\zeta(b) := f(x + b\nabla f(x)) \ge f(x) + \frac{b}{2}\|\nabla f(x)\|^2 > f(x). \tag{30}$$

Now for any $0 < \beta < b$,

$$\zeta'(\beta) = \nabla f\,(x + \beta \nabla f(x)) \cdot \nabla f(x),$$

and by a Taylor expansion of the components of $\nabla f$

$$\nabla f\,(x + \beta\nabla f(x)) = \nabla f(x) + R_2(x, \beta),$$

where $\|R_2(x, \beta)\| \le \beta\sqrt{d}\kappa_2(x)\|\nabla f(x)\|$. Hence, for any $0 < \beta < b$

$$\zeta'(\beta) = \|\nabla f(x)\|^2 + R(x, \beta) \cdot \nabla f(x) \ge \frac{1}{2}\|\nabla f(x)\| > 0$$

since $\beta < b \le (2\sqrt{d}\kappa_2(x))^{-1}$ and so $f$ is increasing along the line segment $[x, x + b\nabla f(x)]$.

**Claim.** *For $a$ sufficiently small, $x_a(t) \in \mathcal{L}_0$ for all $t \ge 0$.* Indeed, since $\kappa_2(x) \le \kappa_2$ for all $x$ in $\mathcal{L}_0$, we have $1 \wedge (2\sqrt{d}\kappa_2(x))^{-1} \ge 1 \wedge (2\sqrt{d}\kappa_2)^{-1}$ for all $x$ in $\mathcal{L}_0$. Consequently, by the previous claim, for any $x$ in $\mathcal{L}_0$ and $a \le 1 \wedge (2\sqrt{d}\kappa_2)^{-1}$, we have $f(x + a\nabla f(x)) > f(x)$ and the values of $f$ are increasing along the line segment $[x, x + a\nabla f(x)]$. In particular, since $x_a$ starts at $x_0 \in \mathcal{L}_0$, we have $f(x_1) = f(x_0 + \nabla f(x_0)) > f(x_0)$, and the segment $[x_0, x_1]$ belongs to $\mathcal{L}_0$. By recursion, we deduce that $x_a(t)$ belongs to $\mathcal{L}_0$ for all $t \geq 0$.

From now on, we assume that

$$a \leq A_1 := 1 \wedge (2\sqrt{d}\kappa_2)^{-1}. \tag{31}$$

**Claim.** *$f$ is increasing along the polygonal curve $x_a$.* By the previous arguments, the values of $f$ are increasing along the line segment $[x_\ell, x_{\ell+1}]$, for all $\ell \geq 0$.

**Claim.** *$(x_\ell)$ converges to a critical point of $f$.* We just showed that the sequence $(f(x_\ell) : \ell \geq 0)$ is increasing, and since it is bounded by $\kappa_0$, it converges. By the first inequality in (30) and the fact that $\|x_{\ell+1} - x_\ell\| = a\|\nabla f(x_\ell)\|$ by construction, we have

$$f(x_{\ell+1}) - f(x_\ell) \geq \frac{1}{2}a\|\nabla f(x_\ell)\|^2 = \frac{1}{2a}\|x_{\ell+1} - x_\ell\|^2, \tag{32}$$

for all $\ell \geq 1$. Hence, for all $\ell \geq 1$, and all $k \geq 1$, we have

$$f(x_{\ell+k}) - f(x_\ell) \geq \frac{1}{2a}\sum_{i=1}^{k}\|x_{\ell+i} - x_\ell\|^2 \geq \frac{1}{2a}\|x_{\ell+k} - x_\ell\|^2,$$

by the triangle inequality. Since $(f(x_\ell))$ is convergent, it is a Cauchy sequence, and consequently, so is $(x_\ell)$, so that $\tilde{x} := \lim_{\ell\to\infty} x_\ell$ exists. And by (32) and the fact that $f$ is $C^1$, we have

$$\nabla f(\tilde{x}) = \lim_{\ell\to\infty} \nabla f(x_\ell) = 0,$$

so that $\tilde{x}$ is a critical point of $f$.

**Claim.** *We have*

$$\|x(t_\ell) - x_\ell\| \leq \left[e^{\ell a \kappa_2 \sqrt{d}} - 1\right]\kappa_1 a, \quad \forall \ell \geq 0. \tag{33}$$

Indeed, let $e_\ell = x(t_\ell) - x_\ell$. Using (6), we have

$$\begin{aligned} e_{\ell+1} &= x(t_{\ell+1}) - x_{\ell+1} \\ &= e_\ell + [x(t_{\ell+1}) - x(t_\ell) - a\nabla f(x(t_\ell))] + a\left[\nabla f(x(t_\ell)) - \nabla f(x_\ell)\right]. \end{aligned} \tag{34}$$

By the definition of $\kappa_2$, and a Taylor expansion,

$$\|\nabla f(x(t_\ell)) - \nabla f(x_\ell)\| \leq \sqrt{d}\kappa_2\|x(t_\ell) - x_\ell\| = \sqrt{d}\kappa_2\|e_\ell\|. \tag{35}$$

We also have

$$\begin{aligned} x(t_{\ell+1}) - x(t_\ell) - a\nabla f(x(t_\ell)) &= \int_{t_\ell}^{t_{\ell+1}} x'(s)\mathrm{d}s - \frac{a}{t_{\ell+1} - t_\ell}\int_{t_\ell}^{t_{\ell+1}} x'(t_\ell)\mathrm{d}s \\ &= \int_{t_\ell}^{t_{\ell+1}} (x'(s) - x'(t_\ell))\mathrm{d}s, \end{aligned}$$

by the definitions of $x(t)$ and $t_\ell$. Consequently,

$$\|x(t_{\ell+1}) - x(t_\ell) - a\nabla f(x(t_\ell))\| \le \int_{t_\ell}^{t_{\ell+1}} \|x'(s) - x'(t_\ell)\| \mathrm{d}s.$$

For $s \in [t_\ell, t_{\ell+1}]$, we have

$$\|x'(s) - x'(t_\ell)\| = \|\nabla f(x(s)) - \nabla f(x(t_\ell))\| \le \kappa_2 \sqrt{d} \|x(s) - x(t_\ell)\|,$$

and

$$\|x(s) - x(t_\ell)\| = \left\| \int_{t_\ell}^{s} x'(t) \mathrm{d}t \right\| \le \int_{t_\ell}^{s} \|x'(t)\| \mathrm{d}t = \int_{t_\ell}^{s} \|\nabla f(x(t))\| \mathrm{d}t \le \kappa_1 (s - t_\ell).$$

Hence

$$\|x'(s) - x'(t_\ell)\| \le \sqrt{d} \kappa_2 \kappa_1 (s - t_\ell),$$

and, recalling that $t_\ell = a\ell$,

$$\|x(t_{\ell+1}) - x(t_\ell) - a\nabla f(x(t_\ell))\| \le \sqrt{d} \kappa_2 \kappa_1 (t_{\ell+1} - t_\ell)^2 = \sqrt{d} \kappa_2 \kappa_1 a^2. \tag{36}$$

Plugging (36) and (35) into (34), we deduce that

$$\|e_{\ell+1}\| \le \sqrt{d} \kappa_2 \kappa_1 a^2 + (1 + \sqrt{d} \kappa_2 a) \|e_\ell\|.$$

The inequality (33) is now a direct consequence of Lemma 4. (Recall that $x(t_0) = x_0$.)

**Claim.** *$(x_\ell)$ converges to $x^\star$.* By this we mean that $\tilde{x}$ coincides with $x^\star$. Indeed, for any $\eta > 0$, denote by $\mathcal{C}(\eta)$ the connected component of $\mathcal{L}_f(f(x^\star) - \eta)$ that contains $x^\star$. Let $\mathbf{H}$ be a shorthand for $H_f(x^\star)$. Suppose all the eigenvalues of $\mathbf{H}$ are in $(-\overline{\nu}, -\underline{\nu})$ for some $\overline{\nu} > \underline{\nu} > 0$. Because $\mathbf{H}$ is negative definite, when $\epsilon > 0$ is small enough $\bar{B}(x^\star, \epsilon)$ contains no critical point of $f$ other than $x^\star$. Let $\ell_\epsilon$ be such that $\|x_\ell - \tilde{x}\| \le \epsilon/3$ when $\ell \ge \ell_\epsilon$, which is well-defined since $(x_\ell)$ converges to $\tilde{x}$. Using the triangle inequality, and then Lemma 6 and (33), for $\ell = \ell_{\epsilon,a} := \max\left\{\ell_\epsilon, \left\lceil \frac{1}{a\underline{\nu}} \log(3/(C_6\epsilon)) \right\rceil\right\}$, we have

$$\begin{aligned}
\|x^\star - \tilde{x}\| &\le \|x^\star - x(t_\ell)\| + \|x(t_\ell) - x_\ell\| + \|x_\ell - \tilde{x}\| \\
&\le \epsilon/3 + \left[e^{\sqrt{d}\kappa_2 a \ell_{\epsilon,a}} - 1\right] \kappa_1 a + \epsilon/3 \\
&\le \epsilon,
\end{aligned}$$

when $a \le A_2$ for some $A_2 > 0$ (depending on $\epsilon > 0$) sufficiently small. Hence, $\tilde{x} \in \bar{B}(x^\star, \epsilon)$. Since $\tilde{x}$ is a critical point, and the only critical point in $\bar{B}(x^\star, \epsilon)$ is $x^\star$, necessarily $\tilde{x} = x^\star$. This proves (13) for $a \le A := \min(1, A_1, A_2)$, where $A_1$ is defined in (31).

Henceforth, we assume that $a \le A$, so that $x_\ell \to x^\star$ as $\ell \to \infty$, and focus on proving (14).

**Bound for large $t$.** A Taylor expansion gives

$$\nabla f(x) = \mathbf{H}(x - x^\star) + R(x, x^\star), \quad \text{where } \|R(x, x^\star)\| \le \tfrac{\sqrt{d}}{2} \kappa_3 \|x - x^\star\|^2.$$

We then have

$$\begin{aligned}
x_{\ell+1} - x^\star &= x_\ell - x^\star + a\nabla f(x_\ell) \\
&= (\mathbf{I} + a\mathbf{H})(x_\ell - x^\star) + aR(x_\ell, x^\star),
\end{aligned}$$

so that

$$\|x_{\ell+1} - x^\star\| \le (1 - a\nu)\|x_\ell - x^\star\| + a\tfrac{\sqrt{d}}{2}\kappa_3\|x_\ell - x^\star\|^2,$$

for some $\nu > \underline{\nu}$. As $x_\ell \to x^\star$, there is $\ell_0$ such that, for $\ell \ge \ell_0$, $\nu - \frac{\sqrt{d}}{2}\kappa_3\|x_\ell - x^\star\| > \underline{\nu}$, implying

$$\|x_{\ell+1} - x^\star\| \le (1 - a\underline{\nu})\|x_\ell - x^\star\|, \quad \forall \ell \ge \ell_0.$$

By recursion, we deduce that there is a constant $Q_1 > 0$ such that

$$\|x_\ell - x^\star\| \le Q_1(1 - a\underline{\nu})^\ell \le Q_1 e^{-\underline{\nu}\ell a}, \quad \forall \ell \ge 0. \tag{37}$$

Fix $t \in [t_\ell, t_{\ell+1}]$. Starting with the triangle inequality, we have

$$\begin{aligned}\|x(t) - x_a(t)\| &\le \|x(t) - x_\star\| + \|x_\star - x_\ell\| + \|x_\ell - x_a(t)\| \\ &\le C_6 e^{-\underline{\nu} t} + Q_1 e^{-\underline{\nu}\ell a} + (t - t_\ell)\|\nabla f(x_\ell)\| \\ &\le Q_2 e^{-\underline{\nu} t} + \kappa_1 a. \end{aligned} \tag{38}$$

In the first line, we applied (25), (37), and used the definition of $x_a$. In the second line, we let $Q_2 = C_6 + Q_1 e^{\underline{\nu} A}$ and used the definition of $\kappa_1$ in (11).

**Bound for small $t$.** On the other hand, we also have

$$\begin{aligned}\|x(t) - x_a(t)\| &\le \|x(t) - x(t_\ell)\| + \|x(t_\ell) - x_\ell\| + \|x_\ell - x_a(t)\| \\ &\le \kappa_1(t_{\ell+1} - t_\ell) + \|x(t_\ell) - x_\ell\| + \|x_\ell - x_{\ell+1}\| \\ &= \kappa_1 a + \|x(t_\ell) - x_\ell\| + a\|\nabla f(x_\ell)\| \\ &\le 2\kappa_1 a + \|x(t_\ell) - x_\ell\|. \end{aligned}$$

Because $f$ is $C^3$, there is $\epsilon > 0$ such that all the eigenvalues of $H_f(x)$ exceed $-\overline{\nu}$ when $x \in \bar{B}(x^\star, \epsilon)$. Let $\ell_\epsilon$ be such that $x(t), x_\ell \in \bar{B}(x^\star, \epsilon)$ for all $t \ge a\ell_\epsilon$ and $\ell \ge \ell_\epsilon$, which implies

$$\|\nabla f(x(t)) - \nabla f(x_\ell)\| \le \overline{\nu}\|x(t) - x_\ell\|.$$

Using this inequality instead of (35), we can refine (33) into

$$\|x(t_\ell) - x_\ell\| \le \left[e^{\ell a \overline{\nu}} - 1\right]\kappa_1 a, \quad \forall \ell \ge \ell_\epsilon,$$

and since $\epsilon$ is fixed, we can combine this with (33) to get

$$\|x(t_\ell) - x_\ell\| \le \left[e^{\ell a \overline{\nu}} - 1\right]\kappa_1 a + Q_3 a, \quad \forall \ell \ge 0, \tag{39}$$

for some constant $Q_3$. We thus have

$$\|x(t) - x_a(t)\| \le \left[2\kappa_1 + (e^{\overline{\nu} t} - 1)\kappa_1 + Q_3\right] a, \tag{40}$$

using the fact that $t \ge t_\ell = a\ell$.

Combining (38) and (40), we have

$$\|x(t) - x_a(t)\| \le (\kappa_1 + Q_3)a + \min\left\{\kappa_1 a e^{\overline{\nu} t}, Q_2 e^{-\underline{\nu} t}\right\}.$$

From this, we deduce (14) from elementary calculations.

### 4.3 Proof of Theorem 2

Below, $C_m$ refers to the constant defined in Lemma $m$.

Arguing as in the proof of Theorem 1, we may assume, without any loss of generality, that $\mathcal{L}_f(f(x_0)/2) \subset \bar{B}(x_0, 3r_0)$. So from now on, we assume that $\mathcal{L}_f(f(x_0)/2)$ is compact and we set $S = \mathcal{L}_f(f(x_0)/2)$. For any $0 \le \ell \le 3$, we also let $\kappa_\ell$ be short for $\kappa_\ell(f, S)$.

**Claim.** *For $\eta_0$ sufficiently small, $\hat{x}(t) \in S$.* Indeed, suppose there is $t \ge 0$ such that $\hat{x}(t) \notin S$. Fix $\epsilon = f(x_0)/2$. Then, by continuity, there is $0 \le t' < t$ such that $f(\hat{x}(t')) = f(x_0) - \epsilon$. Since $\hat{x}(t'), x_0 \in S$, we have

$$\begin{aligned} \hat{f}(\hat{x}(t')) &= \hat{f}(\hat{x}(t')) - f(\hat{x}(t')) + f(\hat{x}(t')) \\ &\le \eta_0 + f(x_0) - \epsilon \\ &= \eta_0 + \hat{f}(x_0) + f(x_0) - \hat{f}(x_0) - \epsilon \\ &\le \hat{f}(x_0) + 2\eta_0 - \epsilon, \end{aligned}$$

by the triangle inequality, applied twice. Since $\hat{f}(\hat{x}(t')) \ge \hat{f}(x_0)$, we see that this situation does not arise when $\eta_0 < \epsilon/2$.

**Claim.** *$\hat{x}^\star = \lim_{t\to\infty} \hat{x}(t)$ is well defined and is close to $x^\star$.* Since $\hat{f}$ is of class $C^3$ by assumption, the map $x \mapsto \nabla \hat{f}(x)$ is $C^1$, and since $\hat{x}(t)$ stays in $S$ and $S$ is compact, $\hat{x}(t)$ is defined for all $t \ge 0$ by the first corollary to the first theorem in (Hirsch et al., 2004, Sec. 17.4). For any $\epsilon \in (0, C_5)$, with $\epsilon < f(x^\star) - f(x_0)/2$, let $t_\epsilon$ be such that $x(t) \in \bar{B}(x^\star, \sqrt{(2\epsilon/\overline{\nu})})$ for all $t \ge t_\epsilon$, which is well-defined since $x(t) \to x^\star$ as $t \to \infty$. By Lemma 7, we have

$$\|\hat{x}(t) - x(t)\| \le \frac{\eta_1}{\sqrt{d}\kappa_2} e^{\sqrt{d}\kappa_2 t}, \quad \forall t \ge 0. \tag{41}$$

Hence

$$\|\hat{x}(t_\epsilon) - x^\star\| \le \|\hat{x}(t_\epsilon) - x(t_\epsilon)\| + \|x(t_\epsilon) - x^\star\| \le \frac{\eta_1}{\sqrt{d}\kappa_2} e^{\sqrt{d}\kappa_2 t_\epsilon} + \sqrt{\frac{2\epsilon}{\overline{\nu}}} =: \delta_1. \tag{42}$$

Assume that $\eta_1$ and $\epsilon$ are small enough that $\delta_1 < \sqrt{(2C_5/\overline{\nu})}$. Letting $\mathcal{C}(\epsilon)$ be as in Lemma 5, by (22) we have

$$\bar{B}(x^\star, \delta_1) \subset \mathcal{C}(\epsilon_1),$$

with $\epsilon_1 := \frac{\overline{\nu}}{2}\delta_1^2$. Thus $\hat{x}(t_\epsilon)$ belongs to $\mathcal{C}(\epsilon_1)$ and in particular $f(\hat{x}(t_\epsilon)) \ge f(x^\star) - \epsilon_1$. Using this last inequality, we deduce by the triangle inequality and the fact that $t \mapsto \hat{f}(\hat{x}(t))$ is increasing that for all $t \ge t_\epsilon$,

$$f(\hat{x}(t)) \ge \hat{f}(\hat{x}(t)) - \eta_0 \ge \hat{f}(\hat{x}(t_\epsilon)) - \eta_0 \ge f(\hat{x}(t_\epsilon)) - 2\eta_0 \ge f(x^\star) - \epsilon_2,$$

where $\epsilon_2 := \epsilon_1 + 2\eta_0$. Since $\hat{x}(t_\epsilon) \in \mathcal{C}(\epsilon_1) \subset \mathcal{C}(\epsilon_2)$ and $\{\hat{x}(t) : t \ge t_\epsilon\}$ is connected and in $\mathcal{L}_f(f(x^\star) - \epsilon_2)$, we necessarily have $\{\hat{x}(t) : t \ge t_\epsilon\} \subset \mathcal{C}(\epsilon_2)$. Assume $\epsilon, \eta_0, \eta_1$ are small enough that $\epsilon_2 \le C_5$. Then, by Lemma 5, $\mathcal{C}(\epsilon_2) \subset \bar{B}(x^\star, \sqrt{2\epsilon_2/\underline{\nu}})$, and so

$$\|\hat{x}(t) - x^\star\| \le \epsilon_3 := \sqrt{2\epsilon_2/\underline{\nu}}, \quad \text{for all } t \ge t_\epsilon. \tag{43}$$

Assume $\epsilon, \eta_0, \eta_1$ are small enough that $\bar{B}(x^\star, \epsilon_3) \subset S$. For any $x$ and $y$ in $\bar{B}(x^\star, \epsilon_3)$, we then have

$$\|H_f(x) - H_f(y)\| \le d\|H_f(x) - H_f(y)\|_{\max} \le d^{\frac{3}{2}} \kappa_3 \|x - y\|. \tag{44}$$

Using (44), for any $x$ in $\bar{B}(x^\star, \epsilon_3)$

$$\begin{aligned}\|H_{\hat{f}}(x) - H_f(x^\star)\| &\le \|H_{\hat{f}}(x) - H_f(x)\| + \|H_f(x) - H_f(x^\star)\| \\ &\le \eta_2 + d^{\frac{3}{2}}\kappa_3\|x - x^\star\| \\ &\le \eta_2 + d^{\frac{3}{2}}\kappa_3\epsilon_3. \end{aligned} \tag{45}$$

We then apply Weyl's inequality (Stewart and Sun, 1990, Cor. IV.4.9) to conclude that, when $\eta_2$ and $\epsilon_3$ are small enough, for all $x$ in $\bar{B}(x^\star, \epsilon_3)$, the eigenvalues of $H_{\hat{f}}(x)$ are all in $(-\infty, -\underline{\nu})$. We assume that $\epsilon, \eta_0, \eta_1, \eta_2$ are small enough that this is the case. This implies that any critical point of $\hat{f}$ in $\bar{B}(x^\star, \epsilon_3)$ is isolated and a local maximum of $\hat{f}$. Using (43) and the compactness of $\bar{B}(x^\star, \epsilon_3)$, by Cantor's intersection theorem $K := \cap_{t\ge t_\epsilon}\overline{\{\hat{x}(u) : u \ge t\}}$ is nonempty. In addition, $K$ is composed of critical points of $\hat{f}$; see Hirsch et al. (2004, Section 9.3, Proposition, p. 206 and Theorem, p. 205) or Absil and Kurdyka (2006, Lemma 5). Therefore we conclude that $K$ is a singleton, which we denote by $\hat{x}^\star$. This is a critical point of $\hat{f}$ in $\bar{B}(x^\star, \epsilon_3)$ and is the limit of $\hat{x}(t)$ as $t \to \infty$. Moreover, $\hat{x}^\star$ is a local maximum of $\hat{f}$.

Since our assumptions imply that $x^\star$ is also a local maximum, we can apply Lemma 8 to bound $\|\hat{x}^\star - x^\star\|$. In our setting, applying the triangle inequality, we may take $C_8 = \max\left\{1, \frac{2}{\sqrt{\underline{\nu}}}, \frac{4\kappa}{3\underline{\nu}}\right\}$, where $\kappa = \kappa_3 + \eta_3$. Assume $\epsilon, \eta_0, \eta_1$ are small enough that $\epsilon_3 \le 1/C_8$. Then, by (43) and Lemma 8, we conclude that $\|\hat{x}^\star - x^\star\| \le C_8\sqrt{2\eta_0}$. Hence we have shown that there exists a constant $Q_0 := Q_0(f, \underline{\nu}) \ge 1$ such that, whenever $\max\{\eta_0, \eta_1, \eta_2\} \le 1/Q_0$ and $\eta_3 \le Q_0$,

$$\|\hat{x}^\star - x^\star\| \le Q_0\sqrt{\eta_0}. \tag{46}$$

Let $\mathbf{H}$ and $\hat{\mathbf{H}}$ be short for $H_f(x^\star)$ and $H_{\hat{f}}(\hat{x}^\star)$, respectively. We now bound $\|\hat{x}(t) - x(t)\|$ in two ways.

**Bound for large $t$.** We proceed with a linearization of the flows near the critical points. Let $\nu > \underline{\nu}$, but close enough that all the eigenvalues of $\mathbf{H}$ are still in $(-\infty, -\nu)$. Note first that $x^\star$ is an interior point of $S$. Suppose that $\max\{\eta_0, \eta_1, \eta_2\} \le 1/Q_0$ and $\eta_3 \le Q_0$ so that (46) holds. By combining (45) and (46)

$$\|\hat{\mathbf{H}} - \mathbf{H}\| \le \eta_2 + d^{\frac{3}{2}}\kappa_3 Q_0\sqrt{\eta_0}. \tag{47}$$

Suppose in addition that $\eta_0$ is small enough that $\hat{x}^\star$ is also an interior point to $S$, which is possible by (46), and that $\|\hat{\mathbf{H}} - \mathbf{H}\|$ is small enough that $\hat{\mathbf{H}}$ also has all its eigenvalues in $(-\infty, -\nu)$, which is possible by (47) and Weyl's inequality for $\eta_0$ and $\eta_2$ small enough. Then there exists $r_\ddagger > 0$ such that

$$\bar{B}(x^\star, r_\ddagger) \subset S \quad \text{and} \quad \bar{B}(\hat{x}^\star, r_\ddagger) \subset S,$$

and since $x(t) \to x^\star$ and $\hat{x}(t) \to \hat{x}^\star$ as $t \to \infty$, there exists a time $t_\ddagger > 0$ such that

$$x(t) \in \bar{B}(x^\star, r_\ddagger) \subset S \quad \text{and} \quad \hat{x}(t) \in \bar{B}(\hat{x}^\star, r_\ddagger), \quad \text{for any } t \ge t_\ddagger.$$

Letting $x_\ddagger(t) = x(t) - x^\star$ and $\hat{x}_\ddagger(t) = \hat{x}(t) - \hat{x}^\star$, by a Taylor expansion, for all $t \geq t_\ddagger$ we have

$$x_\ddagger'(t) = \nabla f(x(t)) = \mathbf{H}\, x_\ddagger(t) + R(t), \qquad \text{with} \quad \|R(t)\| \leq \tfrac{\sqrt{d}\kappa_3}{2} \|x_\ddagger(t)\|^2 \ ; \tag{48}$$

$$\hat{x}_\ddagger'(t) = \nabla \hat{f}(\hat{x}(t)) = \hat{\mathbf{H}}\, \hat{x}_\ddagger(t) + \hat{R}(t), \qquad \text{with} \quad \|\hat{R}(t)\| \leq \tfrac{\sqrt{d}(\kappa_3+\eta_3)}{2} \|\hat{x}_\ddagger(t)\|^2 \ . \tag{49}$$

The difference gives

$$\begin{aligned} x_\ddagger'(t) - \hat{x}_\ddagger'(t) &= \mathbf{H}x_\ddagger(t) - \hat{\mathbf{H}}\hat{x}_\ddagger(t) + R(t) - \hat{R}(t) \\ &= \mathbf{H}(x_\ddagger(t) - \hat{x}_\ddagger(t)) + (\mathbf{H} - \hat{\mathbf{H}})\hat{x}_\ddagger(t) + R(t) - \hat{R}(t), \end{aligned} \tag{50}$$

and after integration between 0 and $t > 0$, we get

$$x_\ddagger(t) - \hat{x}_\ddagger(t) = -e^{t\mathbf{H}}(x^\star - \hat{x}^\star) + \int_0^t e^{(t-s)\mathbf{H}} \big[(\mathbf{H} - \hat{\mathbf{H}})\hat{x}_\ddagger(s) + R(s) - \hat{R}(s)\big] \mathrm{d}s. \tag{51}$$

To check that, note that $x_\ddagger(0) - \hat{x}_\ddagger(0) = x^\star - \hat{x}^\star$, and by differentiating (51), we get

$$\begin{aligned} x_\ddagger'(t) - \hat{x}_\ddagger'(t) = -\mathbf{H}e^{t\mathbf{H}}(x^\star - \hat{x}^\star) + \mathbf{H}e^{t\mathbf{H}} \int_0^t e^{-s\mathbf{H}} \big[(\mathbf{H} - \hat{\mathbf{H}})\hat{x}_\ddagger(s) + R(s) - \hat{R}(s)\big] \mathrm{d}s \\ + (\mathbf{H} - \hat{\mathbf{H}})\hat{x}_\ddagger(t) + R(t) - \hat{R}(t). \end{aligned} \tag{52}$$

From (51), $e^{t\mathbf{H}}(x^\star - \hat{x}^\star)$ may be expressed as

$$e^{t\mathbf{H}}(x^\star - \hat{x}^\star) = -(x_\ddagger(t) - \hat{x}_\ddagger(t)) + \int_0^t e^{(t-s)\mathbf{H}} \big[(\mathbf{H} - \hat{\mathbf{H}})\hat{x}_\ddagger(s) + R(s) - \hat{R}(s)\big] \mathrm{d}s. \tag{53}$$

By reporting (53) in (52) we indeed obtain (50).

Using the triangle inequality in (51), and the fact that all the eigenvalues of $\mathbf{H}$ and $\hat{\mathbf{H}}$ are in $(-\infty, -\nu)$ we then get by (48) and (49) that

$$\begin{aligned} \|x_\ddagger(t) - \hat{x}_\ddagger(t)\| &\leq e^{-\nu t}\|x^\star - \hat{x}^\star\| \\ &\quad + \sqrt{d} \int_0^t e^{-\nu(t-s)} \left[\eta_2 \|\hat{x}_\ddagger(s)\| + \tfrac{\kappa_3}{2}\|x_\ddagger(s)\|^2 + \tfrac{\kappa_3+\eta_3}{2}\|\hat{x}_\ddagger(s)\|^2\right] \mathrm{d}s. \end{aligned}$$

By Lemma 6, $\max(\|x_\ddagger(t)\|, \|\hat{x}_\ddagger(t)\|) \leq C_6 e^{-\nu t}$ for all $t \geq 0$. We use this to bound the integral above. We have

$$\begin{aligned} &\int_0^t e^{-\nu(t-s)} \left[\eta_2 \|\hat{x}_\ddagger(s)\| + \tfrac{\kappa_3}{2}\|x_\ddagger(s)\|^2 + \tfrac{\kappa_3+\eta_3}{2}\|\hat{x}_\ddagger(s)\|^2\right] \mathrm{d}s \\ &\quad \leq \int_0^t e^{-\nu(t-s)} \left[\eta_2 C_6 e^{-\nu s} + \tfrac{\kappa_3}{2} C_6^2 e^{-2\nu s} + \tfrac{\kappa_3+\eta_3}{2} C_6^2 e^{-2\nu s}\right] \mathrm{d}s \\ &\quad \leq C_6 e^{-\nu t} \left[\eta_2 t + (\kappa_3 + \eta_3) C_6 \frac{1 - e^{-\nu t}}{\nu}\right]. \end{aligned}$$

Hence

$$\|x_\ddagger(t) - \hat{x}_\ddagger(t)\| \leq e^{-\nu t}\|x^\star - \hat{x}^\star\| + \sqrt{d} C_6 e^{-\nu t} \left[\eta_2 t + (\kappa_3 + \eta_3) C_6 \frac{1 - e^{-\nu t}}{\nu}\right]. \tag{54}$$

By the triangle inequality, $\|x(t) - \hat{x}(t)\| \le \|x^\star - \hat{x}^\star\| + \|x_\ddagger(t) - \hat{x}_\ddagger(t)\|$, and using (46) and (54), we deduce that

$$\|x(t) - \hat{x}(t)\| \le (1 + e^{-\nu t}) Q_0 \sqrt{\eta_0} + \sqrt{d} C_6 e^{-\nu t} \left[\eta_2 t + (\kappa_3 + \eta_3) C_6 \frac{1 - e^{-\nu t}}{\nu}\right], \quad \text{for all } t \ge t_\ddagger.$$

By increasing the constant factors as needed, we arrive at

$$\|x(t) - \hat{x}(t)\| \le Q_1 \big(\sqrt{\eta_0} + e^{-\nu t} \left[\eta_2 t + \kappa_3 + \eta_3\right] \big), \quad \text{for all } t \ge 0, \tag{55}$$

for some constant $Q_1 > 0$.

**Bound for small** $t$**.** We also have the following refinement of (41). Since $f$ is $C^3$, there exists $\epsilon > 0$ such that all the eigenvalues of $H_f(x)$ exceed $-\overline{\nu}$ when $x \in \bar{B}(x^\star, \epsilon)$. Note that this implies that $\nabla f$ is Lipschitz on $\bar{B}(x^\star, \epsilon)$ with constant $\overline{\nu}$.

Keeping $\epsilon > 0$ fixed, let $t_\epsilon$ be such that $x(t) \in \bar{B}(x^\star, \epsilon)$ and $\hat{x}(t) \in \bar{B}(\hat{x}^\star, \epsilon/2)$, for all $t \ge t_\epsilon$. Assume that $\eta_0$ is small enough that $\|\hat{x}^\star - x^\star\| \le \epsilon/2$, which is possible by (46). Then we also have $\hat{x}(t) \in \bar{B}(x^\star, \epsilon)$

We may now apply Lemma 7 to get

$$\|x(t) - \hat{x}(t)\| \le \frac{\eta_1}{\overline{\nu}} e^{\overline{\nu} t}, \quad \forall t \ge t_\epsilon. \tag{56}$$

Since $\epsilon$ is fixed, by (41), for any $0 \le t \le t_\epsilon$, we have

$$\|x(t) - \hat{x}(t)\| \le \frac{\eta_1}{\sqrt{d}\kappa_2} e^{\sqrt{d}\kappa_2 t} \le \frac{e^{|\sqrt{d}\kappa_2 - \overline{\nu}| t_\epsilon}}{\sqrt{d}\kappa_2} \eta_1 e^{\overline{\nu} t}. \tag{57}$$

Combining (56) and (57) we deduce that

$$\|x(t) - \hat{x}(t)\| \le Q_2 \eta_1 e^{\overline{\nu} t}, \quad \forall t \ge 0, \tag{58}$$

for some constant $Q_2$.

We now combine (55) and (58), and use the fact that $t e^{-\nu t} \le \frac{1}{\nu - \underline{\nu}} e^{-\underline{\nu} t}$ for all $t \ge 0$, to arrive at

$$\|x(t) - \hat{x}(t)\| \le Q_3 \min\left[\sqrt{\eta_0} + e^{-\underline{\nu} t}, \eta_1 e^{\overline{\nu} t}\right], \quad \forall t \ge 0, \tag{59}$$

for some constant $Q_3$. We shall show that the bound (15) follows from (59). To verify this, we start with

$$\min\left[\sqrt{\eta_0} + e^{-\underline{\nu} t}, \eta_1 e^{\overline{\nu} t}\right] \le 2B(t), \quad B(t) := \min\left[\max\{\sqrt{\eta_0}, e^{-\underline{\nu} t}\}, \eta_1 e^{\overline{\nu} t}\right].$$

Set $t_0 = \frac{1}{2\underline{\nu}} \log(1/\eta_0)$ and note that

$$\max\{\sqrt{\eta_0}, e^{-\underline{\nu} t}\} = \begin{cases} e^{-\underline{\nu} t} & \text{when } t \le t_0 \\ \sqrt{\eta_0} & \text{when } t \ge t_0. \end{cases}$$

- When $t \ge t_0$, then we simply observe that $B(t) \le \eta_0^{1/2}$.

- When $t \le t_0$, we have $B(t) = \min\left\{e^{-\underline{\nu}t}, \eta_1 e^{\overline{\nu}t}\right\}$. Let $t_1 = \frac{1}{\underline{\nu}+\overline{\nu}}\log(1/\eta_1)$. Note that the map defined on $[0,\infty)$ by $t \mapsto \min\left\{e^{-\underline{\nu}t}, \eta_1 e^{\overline{\nu}t}\right\}$ is increasing over $[0,t_1]$, decreasing over $[t_1,\infty)$, and that

$$\min\left\{e^{-\underline{\nu}t}, \eta_1 e^{\overline{\nu}t}\right\} = \begin{cases} \eta_1 e^{\overline{\nu}t} & \text{when } t \le t_1 \\ e^{-\underline{\nu}t} & \text{when } t \ge t_1. \end{cases}$$

  - When $t_1 \ge t_0$, $B(t) \le B(t_0) = \eta_1 e^{\overline{\nu}t_0} = \eta_1 \eta_0^{-\frac{\overline{\nu}}{2\underline{\nu}}}$.
  - When $t_1 < t_0$, then $B(t) \le B(t_1) = e^{-\underline{\nu}t_1} = \eta_1^{\frac{\underline{\nu}}{\underline{\nu}+\overline{\nu}}}$

  Since $t_0 \le t_1$ if, and only if, $\eta_1\eta_0^{-\frac{\overline{\nu}}{2\underline{\nu}}} \le \eta_1^{\frac{\underline{\nu}}{\underline{\nu}+\overline{\nu}}}$, we conclude that $B(t) \le \min\left\{\eta_1^{\frac{\underline{\nu}}{\underline{\nu}+\overline{\nu}}}, \eta_1\eta_0^{-\frac{\overline{\nu}}{2\underline{\nu}}}\right\}$ for all $t \le t_0$.

Hence, we worked (59) into

$$\sup_{t\ge 0}\|x(t) - \hat{x}(t)\| \le 2Q_3 \max\left\{\sqrt{\eta_0}, \min\left[\eta_1^{\delta}, \eta_0^{\frac{\delta-1}{2\delta}}\eta_1\right]\right\},$$

where $\delta = \frac{\underline{\nu}}{\underline{\nu}+\overline{\nu}}$. We note that

$$\sqrt{\eta_0} \le \eta_1^{\delta} \iff \eta_0^{\frac{1}{2\delta}} \le \eta_1 \iff \sqrt{\eta_0} \le \eta_1\eta_0^{\frac{1}{2}-\frac{1}{2\delta}} \iff \sqrt{\eta_0} \le \eta_0^{\frac{\delta-1}{2\delta}}\eta_1$$

and that

$$\eta_1^{\delta} \le \eta_0^{\frac{\delta-1}{2\delta}}\eta_1 \iff \eta_0^{\frac{1-\delta}{2\delta}} \le \eta_1^{1-\delta} \iff \sqrt{\eta_0} \le \eta_1^{\delta}.$$

Using these equivalences we deduce that

$$\max\left\{\sqrt{\eta_0}, \min\left[\eta_1^{\delta}, \eta_0^{\frac{\delta-1}{2\delta}}\eta_1\right]\right\} = \max\left\{\sqrt{\eta_0}, \eta_1^{\delta}\right\}.$$

### 4.4 Proof of Lemma 2

For any $d$-tuple $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}^d$, let $|\beta| = \beta_1 + \dots + \beta_d$, and let

$$\partial^{\beta} g(x) = \frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} \dots \partial x_d^{\beta_d}} g(x) \tag{60}$$

denote the $\beta$-th partial derivative of a function $g : \mathbb{R}^d \to \mathbb{R}$. Let $C$ be such that $|\partial^{\beta} f(x)| \le C$ for all $x \in \mathbb{R}^d$ and all $\beta$ such that $|\beta| \le 3$.

Fix $\beta \in \mathbb{N}^d$ with $|\beta| = \ell \le 3$. Since the partial derivatives of $\Phi$ up to the order 3 vanish at infinity, and those of $f$ are bounded, we obtain by integrating by parts

$$\begin{aligned}
\mathbb{E}\left[\partial^{\beta}\hat{f}(x)\right] &= \frac{1}{h^{d+\ell}}\mathbb{E}\left[\partial^{\beta}\Phi\left(\frac{x-X}{h}\right)\right] \\
&= \frac{1}{h^d}\int_{\mathbb{R}^d}\Phi\left(\frac{x-u}{h}\right)\partial^{\beta}f(u)\mathrm{d}u \\
&= \int_{\mathbb{R}^d}\Phi(u)\,\partial^{\beta}f(x-hu)\mathrm{d}u.
\end{aligned}$$

When $\ell = 3$, we simply deduce that

$$\left|\mathbb{E}\big[\partial^\beta \hat{f}(x)\big] - \partial^\beta f(x)\right| \leq \left|\mathbb{E}\big[\partial^\beta \hat{f}(x)\big]\right| + C \leq 2C,$$

using Jensen's inequality.

When $\ell = 2$, we use a Taylor expansion of order 1, to get

$$\big|\partial^\beta f(x - hu) - \partial^\beta f(x)\big| \leq \sqrt{d} C h \|u\|, \quad \forall x, u \in \mathbb{R}^d,$$

and deduce that

$$\left|\mathbb{E}\big[\partial^\beta \hat{f}(x)\big] - \partial^\beta f(x)\right| \leq h \sqrt{d} C \int_{\mathbb{R}^d} \|u\| \Phi(u) \mathrm{d}u,$$

using the fact that $\Phi$ integrates to 1.

When $\ell \leq 1$, we use a Taylor expansion of order 2, to get

$$\big|\partial^\beta f(x - hu) - \partial^\beta f(x) + h\big(\partial^\beta f\big)^{(1)}(x)[u]\big| \leq dCh^2 \|u\|^2, \quad \forall x, u \in \mathbb{R}^d,$$

and deduce that

$$\left|\mathbb{E}\big[\partial^\beta \hat{f}(x)\big] - \partial^\beta f(x)\right| \leq h^2 \, dC \int_{\mathbb{R}^d} \|u\|^2 \Phi(u) \mathrm{d}u,$$

using the fact that $\Phi$ integrates to 1 and kills moments of order 1 by assumption (18).

### 4.5 Proof of Lemma 3

From Theorem 4.1 in Mason (2012), we immediately deduce the following. (Note that in the statement of condition (G.iii) of Theorem 4.1 in Mason (2012), $\mathcal{G}$ should be corrected to be $\mathcal{G}_0$).

**Lemma 9** *Let $f$ be a density on $\mathbb{R}^d$ and let $X \sim f$. Let $\mathcal{G}$ be a class of uniformly bounded measurable functions $\mathbb{R}^d \times (0, 1] \to \mathbb{R}$, such that*

$$\sup_{g \in \mathcal{G}} \sup_{h \in (0,1]} \frac{1}{h^d} \mathbb{E}\left[g(X, h)^2\right] < \infty, \tag{61}$$

*and such that the class*

$$\mathcal{G}_0 = \{x \mapsto g(x, h) : g \in \mathcal{G}, h \in (0, 1)\} \tag{62}$$

*is pointwise measurable and of VC-type. Then there exists a $0 < b_0 < 1$ such that if $X_1, X_2, \ldots$ is an iid sequence from $f$,*

$$\limsup_{n\to\infty} \sup_{g \in \mathcal{G}} \sup_{\frac{\log n}{n} \leq h^d \leq b_0} \sqrt{\frac{n}{h^d \log n}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, h) - \mathbb{E}[g(X, h)] \right| < \infty, \quad \textit{almost surely.} \tag{63}$$

For the definitions of VC-type and pointwise measurable, we refer to Mason (2012, Sec. 4.2) or van der Vaart and Wellner (1996).

**Remark 1** *The assumption that the class $\mathcal{G}_0$ be pointwise measurable insures that the supremum of functionals defined on $\mathcal{G}_0$ be measurable. Another condition that is often imposed on a class of functions is image-Suslin measurable. For details see page 138 of de la Peña and Giné (1999).*

Let $\Phi$ be a kernel and $f$ be a density as in Lemma 3, and let $X \sim f$. Fixing $\beta \in \mathbb{N}^d$ such that $|\beta| \leq 3$, we apply this lemma to

$$\mathcal{G} = \big\{(x,h) \mapsto \partial^\beta \Phi(\tfrac{u-x}{h}) : u \in \mathbb{R}^d\big\}.$$

For any $x, u \in \mathbb{R}^d$ and $h \in (0,1]$,

$$\big|\partial^\beta \Phi(\tfrac{u-x}{h})\big| \leq \|\partial^\beta \Phi\|_\infty,$$

so that $\mathcal{G}$ is uniformly bounded, and

$$\mathbb{E}\left[\partial^\beta \Phi\left(\frac{u-X}{h}\right)^2\right] = \int_{\mathbb{R}^d} \partial^\beta \Phi\left(\frac{u-x}{h}\right)^2 f(x)\mathrm{d}x,$$

which by the change of variables $v = \frac{u-x}{h}$ equals

$$h^d \int_{\mathbb{R}^d} \partial^\beta \Phi(v)^2 f(u-hv)\mathrm{d}v \leq h^d \|f\|_\infty \|\partial^\beta \Phi\|_\infty \int_{\mathbb{R}^d} \big|\partial^\beta \Phi(v)\big| \,\mathrm{d}v, \tag{64}$$

where $\|f\|_\infty$, $\|\partial^\beta \Phi\|_\infty$, and $\int_{\mathbb{R}^d} \big|\partial^\beta \Phi(v)\big| \,\mathrm{d}v$ are finite by assumption. Hence $\mathcal{G}$ satisfies (61). In addition, $\mathcal{G}_0$ is seen to be pointwise measurable by consideration of the subclass

$$\big\{x \mapsto \partial^\beta \Phi(\tfrac{u-x}{h}) : u \in \mathbb{Q}^d, h \in (0,1] \cap \mathbb{Q}\big\}.$$

To see that $\mathcal{G}_0$ is of VC-type, notice that for any $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, $\partial^\beta \Phi(x) = \prod_{k=1}^d \phi_k^{(\beta_k)}(x_k)$. By assumption, $\phi_k^{(\beta_k)}$ is of bounded variation on $\mathbb{R}$, so that by Nolan and Pollard (1987, Lem 22) the class of functions given by

$$\mathrm{g}_{0,k} := \left\{s \in \mathbb{R} \mapsto \phi_k^{(\beta_k)}(\tfrac{s-t}{h}) : s \in \mathbb{R},\ 0 < h \leq 1\right\}$$

is of VC-type. Then an application of Einmahl and Mason (2000, Lem A1) shows that the class of functions $\mathcal{G}_0$, which is equivalently expressed as

$$\mathcal{G}_0 := \left\{(u_1, \dots, u_d) \mapsto g_1(u_1) \dots g_d(u_d) : g_k \in \mathrm{g}_{0,k},\ k = 1, \dots, d\right\},$$

is of VC-type.

Therefore, the conditions of Lemma 9 are met, so that we can assert that (63) holds. Noticing that

$$\frac{1}{n}\sum_{i=1}^n \partial^\beta \Phi(\tfrac{u-X_i}{h}) = h^{\ell+d} \partial^\beta f_{n,h}(u),$$

and consequently

$$\mathbb{E}\left[\partial^\beta \Phi(\tfrac{u-X}{h})\right] = h^{\ell+d}\, \mathbb{E}\left[\partial^\beta f_{n,h}(u)\right],$$

we see that (63) yields

$$\limsup_{n\to\infty} \sup_{u\in\mathbb{R}^d} \sup_{\frac{\log n}{n} \leq h^d \leq b_0} \sqrt{\frac{n}{h^d \log n}} h^{\ell+d} \left|\partial^\beta f_{n,h}(u) - \mathbb{E}\left[\partial^\beta f_{n,h}(u)\right]\right| < \infty, \quad \text{almost surely,}$$

which is exactly (20).

### 4.6 Proof of Theorem 3

As in the proofs of Theorems 1 and 2, we may assume without loss of generality that $\mathcal{L}_f(f(x_0/2)) \subset \bar{B}(x_0, 3r_0)$, with $r_0 = \sup_{t\geq 0} \|x(t) - x_0\|$, which implies that $\mathcal{L}_f(f(x_0/2)$ is compact. In this subsection,

$$S = \mathcal{L}_f(f(x_0)/2), \quad \kappa_\ell = \kappa_\ell(f, S), \quad \hat{f} = \hat{f}_{n,h}, \tag{65}$$

for short.

For any integer $0 \leq \ell \leq 2$, we let

$$\eta_\ell^\star = \sup_{x\in S} \left\| \hat{f}^{(\ell)}(x) - f^\ell(x) \right\| , \quad \eta_\ell = \sup_{x\in S} \left\| \big(\log \hat{f}\big)^{(\ell)}(x) - \big(\log f\big)^{(\ell)}(x) \right\| ,$$

where the norm used is defined in (8). (Keep in mind that we are suppressing in the notation $\hat{f}^{(\ell)}$ and $\eta_\ell$ the dependence on $n$ and $h$.) From (19) and (20), we see that, since $\frac{nh^{d+6}}{\log n} \to \infty$, for any $0 \leq \ell \leq 2$, $\eta_\ell^\star \to 0$ almost surely as $n \to \infty$ while $\eta_3^\star = O(1)$ almost surely. Since $f(x) \geq f(x_0)/2 > 0$ for all $x$ in $S$, and since $\eta_0^\star \to 0$ almost surely, then almost surely, for all $n$ large enough, $\log \hat{f}(x)$ is well-defined for all $x$ in $S$. We have

$$\frac{\partial}{\partial x_i} \log f(x) = \frac{1}{f} \frac{\partial}{\partial x_i} f(x), \qquad \frac{\partial}{\partial x_i} f^{-k}(x) = -k f^{-(k+1)}(x) \frac{\partial}{\partial x_i} f(x), \tag{66}$$

and similarly for $\hat{f}$ almost surely for all $n$ large enough, using the fact that $f(x) \geq f(x_0)/2$ for all $x$ in $S$ once again. We see using (66) that for each $0 \leq \ell \leq 3$ and $\beta \in \mathbb{N}^d$ with $|\beta| = \ell$ there is a continuously differentiable function $F_{\ell,\beta}$ defined on $(0, \infty) \times \mathbb{R}^d \times \cdots \times \mathbb{R}^{d\ell}$, where $\mathbb{R}^{d\ell}$ is suppressed if $\ell = 0$, such that for all $x \in S$

$$\partial^\beta \log f(x) = F_{\ell,\beta} \Big( f(x), \partial^\alpha f(x),\, \alpha \in \mathbb{N}^d \text{ with } |\alpha| = k,\, k = 1, \ldots, \ell \Big),$$

where with some abuse of notation, $\partial^\alpha f(x)$, $\alpha \in \mathbb{N}^d$ with $|\alpha| = k$, $k \geq 1$, represents a $dk$-vector in $\mathbb{R}^{dk}$, and, similarly, almost surely, for all large enough $n$

$$\partial^\beta \log \hat{f}(x) = F_{\ell,\beta} \Big( \hat{f}(x), \partial^\alpha \hat{f}(x),\, \alpha \in \mathbb{N}^d \text{ with } |\alpha| = k,\, k = 1, \ldots, \ell \Big).$$

Observe that the set of points

$$\Big\{ \Big( f(x), \partial^\alpha f(x),\, \alpha \in \mathbb{N}^d \text{ with } |\alpha| = k,\, k = 1, \ldots, \ell \Big) \,:\, x \in S \Big\} \tag{67}$$

lies in a compact subset of $(0, \infty) \times \mathbb{R}^d \times \cdots \times \mathbb{R}^{d\ell}$, and almost surely for all large enough $n$ the same is true for the set of points formed as in (67) with $f$ replaced by $\hat{f}$. Since a compact subset of $(0, \infty) \times \mathbb{R}^d \times \cdots \times \mathbb{R}^{d\ell}$ can be chosen to include both of these sets, using the mean value theorem we see that for some constant $C(\ell, \beta) > 0$

$$\begin{aligned} &\sup_{x\in S} \left| \partial^\beta \log f(x) - \partial^\beta \log \hat{f}(x) \right| \\ &\quad \leq C(\ell, \beta) \max \left\{ \sup_{x\in S} \left| \partial^\alpha f(x) - \partial^\alpha \hat{f}(x) \right| \,:\, \alpha \in \mathbb{N}^d \text{ with } |\alpha| = k,\, k = 0, \ldots, \ell \right\}. \end{aligned}$$

Using (10) this proves that there exists a constant $C > 0$ such that almost surely for all $n$ large enough

$$\eta_\ell \le C(\eta_0^\star + \cdots + \eta_\ell^\star), \quad 0 \le \ell \le 3. \tag{68}$$

Hence, almost surely, $\eta_\ell \to 0$ for all $\ell = 0, 1, 2$ and $\limsup \eta_3 < \infty$. We are then in a position to apply Corollary 1. Noting that $\sqrt{\left(\frac{\log n}{nh^{d+2}}\right)} = o(h^2)$ under the condition $\frac{nh^{d+6}}{\log n} \to \infty$, and using the inequalities in (68), almost surely for all $n$ large enough, $\eta_0 \le Ch^2$ and $\eta_1 \le Ch^2$ for some constant $C > 1$, and since $\delta < 1/2$, almost surely, for all $n$ large enough, $\max\{\sqrt{\eta_0}, \eta_1^\delta\} \le Ch^{2\delta}$. We conclude by applying Corollary 1.

## Acknowledgments

The authors are grateful to Jacob Sterbenz for pointers and stimulating discussions, and to three anonymous referees for comments and for bringing to their attention some important references that were missing, in particular, (Merlet and Pierre, 2010; Stetter, 1973; Comaniciu and Meer, 2002). EAC was supported by a grant from the US National Science Foundation (DMS-0915160). BP was supported by a grant from the French National Research Agency (ANR 09-BLAN-0051-01).

## References

P.-A. Absil and K. Kurdyka. On the stable equilibrium points of gradient systems. *Systems & Control Letters*, 55:573–577, 2006.

W.-J. Beyn. On the numerical approximation of phase portraits near stationary points. *SIAM J. Numer. Anal.*, 24(5):1095–1113, 1987. ISSN 0036-1429. doi: 10.1137/0724072. URL `http://dx.doi.org/10.1137/0724072`.

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(6):3319–3363, 2010.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

M.A. Carreira-Perpinan. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.

M.A. Carreira-Perpinan and C.K.I. Williams. On the number of modes of a gaussian mixture. In *Scale Space Methods in Computer Vision*, volume 2695, pages 625–640. Lecture Notes in Computer Science, 2003.

M.-Y. Cheng, P. Hall, and J.A. Hartigan. Estimating gradient trees. In *A festschrift for Herman Rubin*, volume 45 of *IMS Lecture Notes Monogr. Ser.*, pages 237–249. Inst. Math. Statist., Beachwood, OH, 2004.

Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):1–18, 2002.

V.H. de la Peña and E. Giné. *Decoupling. From dependence to independence. Randomly stopped processes. U-statistics and processes. Martingales and beyond.* Probability and its Applications (New York). Springer-Verlag, New York, 1999.

L. Devroye and L. Gyorfi. *Nonparametric Density Estimation: The $L_1$ View.* John Wiley & Sons, New-York, 1985.

U. Einmahl and D.M. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, 13:1–37, 2000.

U. Einmahl and D.M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33:1380–1403, 2005.

K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21 (1):32–40, 1975.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annals of the Institute Henri Poincaré: Probability and Statistics*, 38:907–921, 2002.

J.A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

M.W. Hirsch and S. Smale. *Differential equations, dynamical systems, and linear algebra.* Academic Press, 1974.

M.W. Hirsch, S. Smale, and R.L. Devaney. *Differential Equations, Dynamical Systems & An Introduction to Chaos.* Academic Press, second edition, 2004.

M. C. Irwin. *Smooth Dynamical Systems.* Academic Press, New York - London, 1980.

J. Li, S. Ray, and B.G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8:1687–1723, 2007.

D.M. Mason. Proving consistency of non-standard kernel estimators. *Stochastic Inference for Stochastic Processes*, 15:151–176, 2012.

D.M. Mason and J. Swanepoel. A general result on the uniform in bandwidth consistency of kernel-type function estimators. *Test*, 20:72–94, 2011.

B. Merlet and M. Pierre. Convergence to equilibrium for the backward Euler scheme and applications. *Commun. Pure Appl. Anal.*, 9(3):685–702, 2010.

D. Nolan and D. Pollard. U-processes: rates of convergence. *Annals of Statistics*, 15: 780–799, 1987.

H. J. Stetter. *Analysis of discretization methods for ordinary differential equations.* Springer-Verlag, New York-Heidelberg, 1973. Springer Tracts in Natural Philosophy, Vol. 23.

G.W. Stewart and J.G. Sun. *Matrix perturbation theory*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1990.

G. Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New-York, 1996.