Uniform in bandwidth estimation of the gradient lines of a density

Ery Arias-Castro*, David Mason† and Bruno Pelletier‡

January 14, 2018

Dedicated to the memory of Jørgen Hoffmann–Jørgensen

Abstract. Let $X_1, \ldots, X_n$, $n \geq 1$, be independent identically distributed (i.i.d.) $\mathbb{R}^d$ valued random variables with a smooth density function f. We discuss how to use these $X's$ to estimate the gradient flow line of f connecting a point x_0 to a local maxima point (mode) based on an empirical version of the gradient ascent algorithm using a kernel estimator based on a bandwidth h of the gradient ∇f of f. Such gradient flow lines have been proposed to cluster data. We shall establish a uniform in bandwidth h result for our estimator and describe its use in combination with plug in estimators for h.

Index Terms: gradient lines, density estimation, nonparametric clustering, uniform in bandwidth

1 Introduction

Let f be a differentiable density on $\mathbb{R}^d$. Assuming that f is known, consider the following iterative scheme. Fix $a > 0$ and, starting at $x_0 \in \mathbb{R}^d$, define iteratively the gradient ascent method

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

When it exists, define $x_\infty = \lim_{\ell\to\infty} x_\ell$. The rationale behind this iterative gradient ascent scheme is to have the sequence $(x_\ell : \ell \geq 0)$ converge to a local maxima point (mode) of f — representing a cluster center.

In fact, one can use this scheme to cluster a set of data by assigning to each observation the nearest mode along the direction of the gradient at the observation point (Fukunaga and Hostetler [7]), where ∇f is replaced by an estimator $\widehat{\nabla f}$ based on the data. This is close in spirit to Hartigan [9].

In practice, the underlying density f is rarely known and has to be estimated using a kernel density estimator. Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a kernel function — an integrable function satisfying

*Department of Mathematics, University of California, San Diego, USA
†Department of Applied Economics and Statistics, University of Delaware, Newark, DE 19717, USA
‡Département de Mathématiques, IRMAR – UMR CNRS 6625, Université Rennes II, France

$\int_{\mathbb{R}^d} \Phi(x)\mathrm{d}x = 1$ — and for a bandwidth $0 < h \le 1$, let $\Phi_h(u) = h^{-d}\Phi(u/h)$. The corresponding kernel estimator of f based on a random sample $X_1, \dots, X_n$, i.i.d. with density f, is

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

and if Φ is differentiable, then we estimate the gradient of f by the kernel type estimator

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

We shall establish a general uniform in bandwidth h result in a sense to be soon made precise in Section 2 for the sequence of estimators beginning with $\hat{x}_0 = x_0$

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

Before we can do this we must first establish some notation and state two general results.

1.1 Two general results

Let $g : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Starting at $x_0 \in \mathbb{R}^d$, we study the convergence as $a \to 0$ of the sequence

$$x_\ell = x_{\ell-1} + a\nabla g(x_{\ell-1}), \quad \text{for } \ell \ge 1, \tag{2}$$

towards the gradient ascent line of g starting at x_0. In particular, we characterize the limit x_∞, providing a consistency result for the clustering algorithm based on the local maxima point of g. Then, given another differentiable function $\widehat{g}$, meant to approximate g, we compare the sequence (x_ℓ) to $(\hat{x}_\ell)$, where

$$\hat{x}_\ell = \hat{x}_{\ell-1} + a\nabla\widehat{g}(\hat{x}_{\ell-1}), \quad \text{for } \ell \ge 1, \tag{3}$$

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 $X_1, \dots, X_n$, $\widehat{g}$ can be taken to be some kernel estimator $\widehat{f}$ of f.

Recall that a *critical point* of g is a point x^* at which the gradient of g vanishes, that is, such that $\nabla g(x^*) = 0$. A *flow line* or *integral curve* of the positive gradient flow of g is a curve x such that

$$x'(t) = \nabla g(x(t)). \tag{4}$$

Note that, along any flow line, the value of g increases, that is, the function $t \mapsto g(x(t))$ 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 $x(0) = x_0$ (see Section 7.2 of Hirsch et al. [10]); we say that $x(t)$ is the flow line starting at x_0. Let $x^\star$ be a critical point of g. 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 $t \ge 0$ and $\lim_{t\to\infty} x(t) = x^\star$. An accumulation point of a sequence of points through an integral curve $x(t)$, i.e., a sequence of the form $\{x(t_n) : t_1 < t_2 < \dots\}$, $t_n \to \infty$, is called a limit point of $x(t)$. Any limit point of a gradient flow line of g is necessarily a critical point of g.

We start by stating a general result by Arias-Castro et al. [1] (also see [2]) who established the convergence of the gradient ascent scheme (2) towards the flow lines of the underlying function g. Starting from a point x_0 in the attraction basin of an isolated local maxima point $x^\star$, under some conditions stated below, the iteration (2) converges to $x^\star$. By an isolated local maxima point $x^\star$ we mean that for all $\epsilon > 0$ small enough the open ball of radius ϵ around $x^\star$, $B(x^\star, \epsilon)$, contains no local maxima point other than $x^\star$. We will show that in fact, the polygonal line defined by the sequence (x_ℓ) is uniformly close to the flow line starting at x_0 and ending at $x^\star$.

Theorem 1 (Convergence of gradient ascent method) *Let g be a function of class C^3. Let $(x(t) : t \geq 0)$ denote the flow line of g starting at x_0 and ending at an isolated local maxima point $x^\star$ of g. Let (x_ℓ) be the sequence defined in (2) starting at x_0. Then there exists $A = A(x_0, g) > 0$ such that, whenever $a < A$,*

$$\lim_{\ell \to +\infty} x_\ell = x^\star. \tag{5}$$

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 $H_g(x^\star)$ has all eigenvalues in $(-\overline{\nu}, -\underline{\nu})$ for some $0 < \underline{\nu} < \overline{\nu}$. Then, there exists a $C_0 = C(x_0, g, \underline{\nu}, \overline{\nu}) > 0$ such that, for any $0 < a < A$,

$$\sup_{t \geq 0} \|x_a(t) - x(t)\| \leq C_0 a^\delta, \quad \text{with } \delta := \underline{\nu} / (\underline{\nu} + \overline{\nu}). \tag{6}$$

Next, we state a version of a stability result of [1] for flows of smooth functions. Under some conditions, when g and $\widehat{g}$ are close as C^2 functions, then their flow lines are also close. First we need some notation.

For a function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we let $\varphi^{(\ell)}(x)$, $\ell \geq 1$, denote the differential form of φ of order ℓ at a point $x \in \mathbb{R}^d$, and let $H_\varphi(x)$ denote the Hessian matrix of φ evaluated at x when they exist. The differential form $\varphi^{(\ell)}(x)$ of φ at x is the multilinear map from $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ (ℓ times) to $\mathbb{R}$ defined for $\ell \geq 1$ by

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

where, for each $1 \leq i \leq \ell$, u_i has components $u_i = (u_{i,1}, \ldots, u_{i,d})$. We write

$$\varphi^{(0)}(x) = \varphi(x), \ x \in \mathbb{R}^d.$$

Given a multilinear map L of order $\ell \geq 1$ from $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ to $\mathbb{R}$, which we write as

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

we denote by $\|L\|$ its operator norm defined by

$$\|L\| = \sup\left\{|L[u_1, \ldots, u_\ell]| : \|u_1\| = \cdots = \|u_\ell\| = 1\right\}. \tag{7}$$

Note that when $\ell = 1$, $\|L\| = \sqrt{\sum_{i=1}^d L_i^2}$, and when $\ell = 2$

$$\|L\| = \sup_{\|u\|=\|v\|=1} |v' L u| = \sup_{\|u\|=1} |Lu|,$$

where L is the $d \times d$ matrix $\{L_{i,j} : 1 \le i, j \le d\}$, (cf. page 7 of Bhatia [3]), which implies that for any $x \in \mathbb{R}^d$

$$|Lx| \le \|L\| \|x\|. \tag{8}$$

When $\ell = 0$ we set $\|L\| = |L|$.
We denote by $\|L\|_{\max}$ the norm defined by

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

We note for future reference that easy calculations show that

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

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

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

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

$$\mathcal{L}_\varphi(b) = \{x \in \mathbb{R}^d : \varphi(x) \ge b\}. \tag{12}$$

We suppress the dependence on φ whenever no confusion is possible. For any $x \in \mathbb{R}^d$ and $r > 0$ denote the open ball

$$B\,(x, r) = \{y : \|x - y\| < r\}$$

and the closed ball

$$\overline{B}\,(x, r) = \{y : \|x - y\| \le r\}.$$

Here is our stability result. It is a version of Theorem 2 of [1] designed to prove our uniform in bandwidth result stated as Theorem 3 in the next section.

Theorem 2 (Stability of smooth flows) *Suppose g and $\widehat{g}$ are of class C^3. Let $(x(t) : t \ge 0)$ be a flow line of g starting at x_0, with $g(x_0) > 0$, and ending at an isolated local maxima point $x^\star$ where $H_g(x^\star)$ has all eigenvalues in $(-\overline{\nu}, -\underline{\nu})$ for some $0 < \underline{\nu} < \overline{\nu}$. Let $\hat{x}(t)$ be the flow line of $\widehat{g}$ starting at x_0. Let $S = \mathcal{L}(g(x_0)/2) \cap \overline{B}(x_0, 3r_0)$, where*

$$r_0 = \max_t \|x(t) - x_0\|, \tag{13}$$

and define

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

Then for all $D > 0$ there exists a constant $C := C(g, x_0, \underline{\nu}, \bar{\nu}, D) \geq 1$ and a function $F(g, x_0, \underline{\nu}, \bar{\nu}, 1/C, D)$ of D such that, whenever $\max\{\eta_0, \eta_1, \eta_2\} \leq 1/C$ and $\eta_3 \leq D$, $\hat{x}(t)$ is defined for all $t \geq 0$ and

$$\sup_{t\geq 0} \|x(t) - \hat{x}(t)\| \leq F(g, x_0, \underline{\nu}, \bar{\nu}, 1/C, D) \max\left\{\sqrt{\eta_0}, \eta_1^\delta\right\}, \tag{14}$$

where $\delta = \underline{\nu}/\left(\underline{\nu} + \bar{\nu}\right)$.

Combining Theorems 1 and 2, we arrive at the following bound for approximating the flow lines of a function g with the polygonal line obtained from the gradient ascent algorithm (3) based on an approximation $\widehat{g}$ to g.

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

$$\hat{x}_a(t) = \hat{x}_{\ell-1} + (t/a - \ell + 1)(\hat{x}_\ell - \hat{x}_{\ell-1}), \quad \forall t \in [(\ell-1)a, \ell a), \tag{15}$$

where $(\hat{x}_\ell)$ is defined in (3). Then for all $D > 0$ there exists a constant $C := C(g, x_0, \underline{\nu}, \bar{\nu}, D) \geq 1$ and a function $F(g, x_0, \underline{\nu}, \bar{\nu}, 1/C, D)$ of D such that, whenever $\max\{\eta_0, \eta_1, \eta_2\} \leq 1/C$ and $\eta_3 \leq D$,

$$\sup_{t\geq 0} \|\hat{x}_a(t) - x(t)\| \leq F(g, x_0, \underline{\nu}, \bar{\nu}, 1/C, D) \left[a^\delta + \max\left\{\sqrt{\eta_0}, \eta_1^\delta\right\}\right], \tag{16}$$

where $\delta = \underline{\nu}/\left(\underline{\nu} + \bar{\nu}\right)$.

In applications, the requirement that $g(x_0) > 0$ can be sidestepped.

2 The estimation of gradient lines of a density

Let $\hat{f}_{n,h}$ be the kernel density estimator of f in (1) with kernel Φ and bandwidth h. Sharp almost-sure convergence rates in the uniform norm of kernel density estimators have been obtained by several authors, for example Einmahl and Mason [5], Giné and Guillou [8], Einmahl and Mason [6], Mason and Swanepoel [12] (also see [13]) and Mason [11].
We first state a bias bound from [1].

Lemma 1 *Assume Φ is nonnegative, C^3 on $\mathbb{R}^d$ with all partial derivatives up to order 3 vanishing at infinity, and satisfies*

$$\int_{\mathbb{R}^d} \Phi(x)\mathrm{d}x = 1, \quad \int_{\mathbb{R}^d} x\Phi(x)\mathrm{d}x = 0 \quad \text{and} \quad \int_{\mathbb{R}^d} \|x\|^2\Phi(x)\mathrm{d}x < \infty. \tag{17}$$

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 for all $0 \leq \ell \leq 3$

$$\sup_{x\in\mathbb{R}^d} \left\|\mathbb{E}\left[\hat{f}_{n,h}^{(\ell)}(x)\right] - f^{(\ell)}(x)\right\| \leq Ch^{(3-\ell)\wedge 2}. \tag{18}$$

Next, by applying the main result of [12] (also see [13] and Theorem 4.1 with Remark 4.2 in [11]), [1] derive the following uniform in bandwidth result for $\hat{f}_{n,h}$ and its derivatives.

Lemma 2 *Suppose that* Φ *is of the form* $\Phi : (x_1, \ldots, x_d) \mapsto \prod_{k=1}^{d} \phi_k(x_k)$*, and that each* ϕ_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 almost surely*

$$\limsup_{n\to\infty} \sup_{\frac{\log n}{n} \leq h^d \leq b_0} \sup_{x\in\mathbb{R}^d} \sqrt{\frac{nh^{d+2\ell}}{\log n}} \left\| \hat{f}^{(\ell)}_{n,h}(x) - \mathbb{E}\left[\hat{f}^{(\ell)}_{n,h}(x)\right] \right\| < \infty, \quad \forall 0 \leq \ell \leq 3. \tag{19}$$

It is straightforward to design a kernel that satisfies the conditions of Lemmas 1 and 2. In fact, the Gaussian kernel $\Phi(x) = (2\pi)^{-d/2} \exp(-\|x\|^2/2)$ is such a kernel.

Theorem 3 *Consider a density* f *satisfying the conditions of Lemma 1. Suppose* $\hat{f}_{n,h}$ *is a kernel estimator of* f *of the form* (1)*, where* Φ *satisfies the conditions of Lemma 1 and 2. Let* $(x(t) : t \geq 0)$ *be the flow line of* f *starting at a point* x_0 *with* $f(x_0) > 0$*, ending at an isolated local maxima point* $x^\star$ *where* $H_f(x^\star)$ *has all eigenvalues in* $(-\overline{\nu}, -\underline{\nu})$ *for some* $0 < \underline{\nu} < \overline{\nu}$*. For* $a > 0$*,* $0 < h \leq 1$ *and* $n \geq 1$ *define* $(\hat{x}_a(t, n, h) : t \geq 0)$ *as in* (15) *with* $\hat{f}$ *taken as* $\hat{f}_{n,h}$ *in* (3)*. i.e. for* $t \in [(\ell-1)a, \ell a)$*,* $\ell \geq 1$*,*

$$\hat{x}_{\ell,n}(h) = \hat{x}_{\ell-1,n}(h) + a\nabla \hat{f}_{n,h}(\hat{x}_{\ell-1}(h)),$$

with $\hat{x}_{0,n}(h) = x_0$*. Suppose that*

$$a_n \to 0,\ \frac{n a_n^{1+6/d}}{\log n} \to \infty \text{ and } a_n < b_n,\ \text{with } b_n \to 0, \tag{20}$$

then there exists a constant $C > 0$ *such that, with probability one, for all* n *large enough, uniformly in* $a_n \leq h^d \leq b_n$*,*

$$\sup_{t\geq 0} \|\hat{x}_a(t, n, h) - x(t)\| \leq C\left(a^\delta + h^{2\delta}\right), \tag{21}$$

where $\delta = \underline{\nu}/\left(\underline{\nu} + \overline{\nu}\right)$*.*

Remark Let

$$\hat{h}_n = H_n(X_1, \ldots, X_n)$$

be a bandwidth estimator so that with probability 1

$$\hat{h}_n \to 0 \text{ and } \liminf_n \frac{\hat{h}_n^d}{a_n} > 0,$$

where a_n satisfies the conditions in (20). Notice that under the assumptions and notation of Theorem 3 we have, with probability 1, for the *plug in* estimator $\hat{x}_a(t, n, \hat{h}_n)$, for all large enough n,

$$\sup_{t\geq 0} \|\hat{x}_a(t, n, \hat{h}_n) - x(t)\| \leq C\left(a^\delta + \hat{h}_n^{2\delta}\right). \tag{22}$$

For a general treatment of bandwidth selection and data-driven bandwidths consult Sections 2.3 and 2.4 of Deheuvels and Mason [4], as well as the references therein.

3 Proofs of Theorem 2 and Theorem 3

To show the reader how all of these results fit together, we shall prove Theorem 3 first.

3.1 Proof of Theorem 3

As in the proof of Theorem 2 in the next subsection, we may assume without loss of generality that $\mathcal{L}_g(f(x_0/2) \subset \overline{B}(x_0, 3r_0)$, with $r_0 = \sup_{t\geq 0} \|x(t) - x_0\|$, which implies that $\mathcal{L}(f(x_0/2)$ is compact.

For any integer $0 \leq \ell \leq 3$, $n \geq 1$ and $0 < h \leq 1$, let

$$\eta_{\ell,n}(h) = \sup_{x\in S} \|\hat{f}_{n,h}^{(\ell)}(x) - f^{\ell}(x)\|,$$

where the norm used is defined in (7). From (18) and (19), we see from the triangle inequality that for some constant $A_\ell > 0$, uniformly in $a_n \leq h^d \leq b_n$, for all large n

$$\eta_{\ell,n}(h) \leq A_\ell \left(h^{(3-\ell)\wedge 2} + \sqrt{\frac{\log n}{nh^{d+2\ell}}} \right)$$

$$\leq A_\ell \left(b_n^{(3-\ell)\wedge 2} + \sqrt{\frac{\log n}{na_n^{1+2\ell/d}}} \right).$$

It is easily checked using (20) that for any $0 \leq \ell \leq 2$

$$\sup_{a_n\leq h^d\leq b_n} \eta_{\ell,n}(h) \to 0, \text{ a.s.},$$

while

$$\limsup_{n\to\infty} \sup_{a_n\leq h^d\leq b_n} \eta_{3,n}(h) \leq A_3, \text{ a.s.}$$

Also one finds that uniformly in $a_n \leq h^d \leq b_n$ for all large n for some constant $B > 0$

$$h^{(3-\ell)\wedge 2} + \sqrt{\frac{\log n}{nh^{d+2\ell}}} \leq Bh^2, \text{ for } \ell = 0, 1.$$

Thus since $\delta < 1/2$, uniformly in $a_n \leq h^d \leq b_n$ for all n large enough,

$$\max\{\sqrt{\eta_{0,n}(h)}, \eta_{1,n}^{\delta}(h)\} \leq Ah^{2\delta},$$

with $A = \max\{\sqrt{A_0 B}, (A_1 B)^{\delta}\}$. We finish the proof by applying Corollary 1. $\square$

3.2 Proof of Theorem 2

Our proof will follow that of Theorem 2 of [1], however with some major modifications and clarifications needed to obtain the present result. We shall require the following two lemmas, which we state here without proof. They are respectively Lemma 5 and 6 of Theorem 2 of [1].

Lemma 3 *Suppose that g is of class C^3. Let $x^\star$ be an isolated local maxima point of g where $H_g(x^\star)$ 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 $\mathcal{L}_g(g(x^\star) - \epsilon)$ that contains $x^\star$. Then there is a constant $C_3 = C_3(g, x^\star)$ such that*

$$\overline{B}(x^\star, \sqrt{(2\epsilon/\overline{\nu})}) \subset \mathcal{C}(\epsilon) \subset \overline{B}(x^\star, \sqrt{2\epsilon/\underline{\nu}}), \quad \text{for all } \epsilon \leq C_3, \tag{23}$$

and

$$g(x^\star) - g(x) \leq \frac{\overline{\nu}}{2}\|x - x^\star\|^2, \quad \text{for all } x \text{ such that } \|x - x^\star\| \leq \sqrt{C_3/\overline{\nu}}. \tag{24}$$

Lemma 4 *Suppose that g is of class C^3. Let $(x(t) : t \geq 0)$ be the flow line of g starting at x_0 and ending at $x^\star$ where $H_g(x^\star)$ has all its eigenvalues in $(-\infty, -\underline{\nu})$, with $\underline{\nu} > 0$. Then, there is $C_4 = C_4(g, x_0)$ such that, for all $t \geq 0$,*

$$\|x(t) - x^\star\| \leq C_4 e^{-\underline{\nu}t}, \tag{25}$$

and

$$g(x^\star) - g(x(t)) \leq C_4 e^{-2\underline{\nu}t}. \tag{26}$$

The following, adapted from Hirsch et al. [10, Section 17.5], is a stability result for autonomous gradient flows.

Lemma 5 *Suppose φ and ψ are of class C^1 and for a measurable subset $\mathcal{S} \subset \mathbb{R}^d$*

$$\|\nabla\varphi(x) - \nabla\psi(x)\| < \varepsilon, \quad \forall x \in \mathcal{S}.$$

Let K be a Lipschitz constant for $\nabla\varphi$ on $\mathcal{S}$. Let $(x(t) : t \geq t_0)$ and $(y(t) : t \geq t_0)$ with $t_0 \geq 0$, be the flow lines of φ and ψ starting at x_1 and y_1, respectively, i.e. $x(t_0) = x_1$ and $y(t_0) = y_1$, and

$$x'(t) = \nabla\varphi(x(t)) \text{ and } y'(t) = \nabla\psi(y(t)), \text{ for } t \geq t_0.$$

Assume that the flow lines $x(t)$ and $y(t)$ are in $\mathcal{S}$. Then,

$$\|x(t) - y(t) - (x_1 - y_1)\| \leq \frac{\varepsilon}{K}\left[e^{Kt} - 1\right], \quad \forall t \geq t_0.$$

For the convenience of the reader we state here the Weyl Perturbation Theorem (see Corollary III.2.6 of Bhatia [3].)

Weyl Perturbation Theorem Let M and H be n by n Hermitian matrices, where M has eigenvalues $\mu_1 \geq \cdots \geq \mu_n$ and H has eigenvalues $\nu_1 \geq \cdots \geq \nu_n$. If $\|M - H\| \leq \varepsilon$, then $|\mu_i - \nu_i| \leq \varepsilon$ for $i = 1, \ldots, n$.

Next is a result on the stability of local maxima points.

Lemma 6 *Suppose f and g are of class C^3, and have local maxima points at x and y, respectively, with $H_f(x)$ having all eigenvalues in $(-\infty, -\nu]$ for some $\nu > 0$. Then for any $0 < b \le 1$ and $\kappa \ge \max\left(\kappa_3(f, \overline{B}\,(x,b)), \kappa_3(g, \overline{B}\,(x,b))\right)$,*

$$\|x-y\| \le \min\left\{\frac{3\nu}{4\kappa}, b\right\} \quad \Rightarrow \quad \|x-y\| \le \frac{2}{\sqrt{\nu}}\big(\,|f(x)-g(x)| + |f(y)-g(y)|\,\big)^{1/2}. \tag{27}$$

Proof Let $\mathbf{H}_f$ and $\mathbf{H}_g$ be short for the Hessian matrices $H_f(x)$ and $H_g(y)$, respectively. We develop f and g around x and y, respectively. Assuming $\|x-y\| \le \min\left\{\frac{3\nu}{4\kappa}, b\right\}$, which implies that $y \in \overline{B}\,(x,b)$, we have

$$f(y) = f(x) + \frac{1}{2}\mathbf{H}_f[x-y, x-y] + R_f(x,y), \qquad \text{with} \quad |R_f(x,y)| \le \frac{\kappa}{6}\|x-y\|^3;$$
$$g(x) = g(y) + \frac{1}{2}\mathbf{H}_g[x-y, x-y] + R_g(x,y), \qquad \text{with} \quad |R_g(x,y)| \le \frac{\kappa}{6}\|x-y\|^3.$$

Summing these two equalities, we obtain

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

Let $\nu > 0$ be such that the largest eigenvalue of $\mathbf{H}_f$ is bounded by $-\nu$. By the triangle inequality and the fact that $\mathbf{H}_g$ is negative semidefinite,

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

Thus, when $\|x-y\| \le \min\left\{\frac{3\nu}{4\kappa}, b\right\}$, we have $\nu\|x-y\|^2 - \frac{2\kappa}{3}\|x-y\|^3 \ge \frac{\nu}{2}\|x-y\|^2$, so that

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

and from this we conclude (27). □

It would help the reader to make his or her way through the intricate arguments that follow to always keep in mind that η_0, η_1, η_2 and $\epsilon > 0$ are assumed to be sufficiently small and $t_\epsilon > 0$ sufficiently large as needed, and $\eta_3 \le D$, where $D > 0$ is a pre-chosen constant.

Bound on $\|\hat{x}^\star - x^\star\|$.

Our first goal is to derive a bound on $\|\hat{x}^\star - x^\star\|$. Arguing as in the proof of Theorem 1 of [1], we may assume, without loss of generality [WLOG], that $\mathcal{L}_g(g(x_0)/2) \subset \overline{B}(x_0, 3r_0)$, where r_0 is as in (13). So from now on, we assume that $\mathcal{L}_g(g(x_0)/2)$ is compact and we set

$$S = \mathcal{L}_g(g(x_0)/2). \tag{28}$$

Note that since $g\,(x\,(t))$ increases along $t \ge 0$, $x\,(t) \in S$ for all $t \ge 0$.

We also let κ_ℓ be short for $\kappa_\ell(g, S)$, as defined in (11).

Claim 1. *For η_0 sufficiently small, $\hat{x}(t) \in S$, for all $t \geq 0$, with S as in (28).* Indeed, suppose there is $t > 0$ such that $\hat{x}(t) \notin S$. Fix $\varrho = g(x_0)/2$. Then, by continuity, there is $0 \leq t' < t$ such that $g(\hat{x}(t')) = g(x_0) - \varrho$. Since both $\hat{x}(t')$ and $x_0 \in S$, we have

$$\begin{aligned} \widehat{g}(\hat{x}(t')) &= \widehat{g}(\hat{x}(t')) - g(\hat{x}(t')) + g(\hat{x}(t')) \\ &\leq \eta_0 + g(x_0) - \varrho \\ &= \eta_0 + \widehat{g}(x_0) + g(x_0) - \widehat{g}(x_0) - \varrho \\ &\leq \widehat{g}(x_0) + 2\eta_0 - \varrho, \end{aligned}$$

by the triangle inequality, applied twice. Since $\widehat{g}(\hat{x}(t')) \geq \widehat{g}(x_0)$, we see that this situation does not arise when $\eta_0 < \varrho/2$. This establishes Claim 1.

From now on we shall assume that η_0 is sufficiently small, so that

$$\hat{x}(t) \in S, \text{ for all } t \geq 0. \tag{29}$$

Claim 2. *For all η_0, η_1 and η_2 sufficiently small, $\hat{x}^\star = \lim_{t\to\infty} \hat{x}(t)$ is well defined and is close to $x^\star$.* Since $\widehat{g}$ is of class C^3 by assumption, the map $x \mapsto \nabla\widehat{g}(x)$ is C^1, and since by Claim 1 for all η_0 sufficiently small $\hat{x}(t)$ stays in S and S is compact, $\hat{x}(t)$ is defined for all $t \geq 0$ by the first corollary to the first theorem in [10, Section 17.5].

Applying Lemma 5 with $t_0 = 0$ and $x_1 = y_1 = x_0$ we get

$$\|\hat{x}(t) - x(t)\| \leq \frac{\eta_1}{\sqrt{d}\kappa_2} e^{\sqrt{d}\kappa_2 t}, \quad \forall t \geq 0, \tag{30}$$

For $\epsilon \in (0, C_3)$, where C_3 is as in Lemma 3, let t_ϵ be such that $x(t) \in B(x^\star, \sqrt{(2\epsilon/\overline{\nu})})$ for all $t \geq t_\epsilon$, which is well-defined since $x(t) \to x^\star$ as $t \to \infty$. Hence

$$\begin{aligned} \|\hat{x}(t_\epsilon) - x^\star\| &\leq \|\hat{x}(t_\epsilon) - x(t_\epsilon)\| + \|x(t_\epsilon) - x^\star\| \\ &\leq \frac{\eta_1}{\sqrt{d}\kappa_2} e^{\sqrt{d}\kappa_2 t_\epsilon} + \sqrt{\frac{2\epsilon}{\overline{\nu}}} =: \delta_1. \end{aligned} \tag{31}$$

Assume that η_1 and ϵ are small enough so that $\delta_1 < \sqrt{C_3/\overline{\nu}}$. Letting $\mathcal{C}(\epsilon)$ be as in Lemma 3, by (23) we have

$$\overline{B}(x^\star, \delta_1) \subset \mathcal{C}(\epsilon_1), \text{ with } \epsilon_1 = \frac{\overline{\nu}}{2}\delta_1^2,$$

noting that $\sqrt{\epsilon_1 2/\overline{\nu}} = \delta_1$ and $\epsilon_1 < C_3/2$. Thus $\hat{x}(t_\epsilon)$ belongs to $\mathcal{C}(\epsilon_1)$ and in particular $g(\hat{x}(t_\epsilon)) \geq g(x^\star) - \epsilon_1$. Using this last inequality, we deduce from the triangle inequality and the fact that $t \mapsto \widehat{g}(\hat{x}(t))$ is increasing that for $t \geq t_\epsilon$,

$$\begin{aligned} g(\hat{x}(t)) &\geq \widehat{g}(\hat{x}(t)) - \eta_0 \geq \widehat{g}(\hat{x}(t_\epsilon)) - \eta_0 \\ &\geq g(\hat{x}(t_\epsilon)) - 2\eta_0 \geq g(x^\star) - \epsilon_2, \end{aligned}$$

where

$$\epsilon_2 := \epsilon_1 + 2\eta_0. \tag{32}$$

Since $\hat{x}(t_\epsilon) \in \mathcal{C}(\epsilon_1) \subset \mathcal{C}(\epsilon_2)$ and $(\hat{x}(t) : t \geq t_\epsilon)$ is connected and in $\mathcal{L}_g(g(x^\star) - \epsilon_2)$, we necessarily have $(\hat{x}(t) : t \geq t_\epsilon) \subset \mathcal{C}(\epsilon_2)$. Assume that ϵ, η_0 and η_1 are small enough so that $\epsilon_2 \leq C_3$. Then, by Lemma 3, $\mathcal{C}(\epsilon_2) \subset \overline{B}\left(x^\star, \sqrt{2\epsilon_2/\underline{\nu}}\right)$, and so

$$\|\hat{x}(t) - x^\star\| \leq \epsilon_3 := \sqrt{2\epsilon_2/\underline{\nu}}, \text{ for all } t \geq t_\epsilon. \tag{33}$$

Assume ϵ, η_0, η_1 are small enough so that $\overline{B}(x^\star, \epsilon_3) \subset S$. For any x and y in $\overline{B}(x^\star, \epsilon_3)$ we get by (10) that

$$\|H_g(x) - H_g(y)\| \leq d\|H_g(x) - H_g(y)\|_{\max} \leq d^{3/2}\kappa_3\|x - y\|. \tag{34}$$

Using (34) and (33), for any $x \in \overline{B}(x^\star, \epsilon_3)$

$$\|H_{\widehat{g}}(x) - H_g(x^\star)\| \leq \|H_{\widehat{g}}(x) - H_g(x)\| + \|H_g(x) - H_g(x^\star)\| \tag{35}$$

$$\leq \eta_2 + d^{3/2}\kappa_3\|x - x^\star\| \leq \eta_2 + d^{3/2}\kappa_3\epsilon_3. \tag{36}$$

Let $\nu > \underline{\nu}$, but close enough such that all the eigenvalues of $\mathbf{H}$ are still in $(-\infty, -\nu)$. We then apply the Weyl Perturbation Theorem, cited above, to conclude that for all η_2 and ϵ_3 small enough and $x \in \overline{B}(x^\star, \epsilon_3)$ so that

$$\eta_2 + d^{3/2}\kappa_3\epsilon_3 \leq \nu - \underline{\nu} \tag{37}$$

the eigenvalues of $H_{\widehat{g}}(x)$ are all in $(-\infty, -\underline{\nu})$. We shall assume that $\epsilon, \eta_0, \eta_1, \eta_2$ are small enough so that this is the case. Using (33) and compactness of $\overline{B}(x^\star, \epsilon_3)$, we get by Cantor's intersection theorem that

$$K := \cap_{t \geq t_\epsilon} \overline{\{\widehat{x}(u) : u \geq t\}}$$

is nonempty. In addition K is composed of critical points of $\widehat{g}$. (See [10], Section 9.3, Proposition, p. 206 and Theorem p. 205). Therefore we conclude that K is a singleton, which we denote $\hat{x}^\star$. This is a critical point of $\widehat{g}$ in $\overline{B}(x^\star, \epsilon_3)$ and is the limit of $\widehat{x}(t)$ as $t \to \infty$. Moreover, $\hat{x}^\star$ is a local maxima point of $\widehat{g}$. This proves Claim 2.

We have just shown that for $\epsilon > 0, \eta_0, \eta_1$and η_2 sufficiently small

$$\|\hat{x}^\star - x^\star\| \leq \epsilon_3.$$

To summarize, the analysis from equations (30) through (37) shows that for all $\epsilon > 0$, η_0, η_1 and η_2 small enough, $\overline{B}(x^*, \epsilon_3) \subset S$, $\hat{x}^* \in \overline{B}(x^*, \epsilon_3)$, $\eta_2 + d^{3/2}\kappa_3\epsilon_3 \leq \nu - \underline{\nu}$ and (33) holds, where

$$\delta_1 = \frac{\eta_1}{\sqrt{d}\kappa_2} e^{\sqrt{d}\kappa_2 t_\epsilon} + \sqrt{\frac{2\epsilon}{\overline{\nu}}},\ \epsilon_1 = \frac{\overline{\nu}}{2}\delta_1^2,\ \epsilon_2 = \epsilon_1 + 2\eta_0, \tag{38}$$

and

$$\epsilon_3 = \sqrt{2\epsilon_2/\overline{\nu}}. \tag{39}$$

Notice that ϵ_3 is a function of $(\epsilon, \eta_0, \eta_1, \eta_2)$ and

$$\frac{\nu - \underline{\nu} - \eta_2}{d^{3/2}\kappa_3} \geq \epsilon_3 = \sqrt{\frac{2(\epsilon_1 + 2\eta_0)}{\overline{\nu}}} = \sqrt{\frac{2\left(\frac{\overline{\nu}}{2}\delta_1^2 + 2\eta_0\right)}{\overline{\nu}}}.$$

Letting $\kappa = \kappa_3 + \eta_3$ and $b = \epsilon_3$ in Lemma 6 we see by (27) that whenever

$$\|\hat{x}^\star - x^\star\| \leq \min\left\{\epsilon_3, \frac{3\underline{\nu}}{4(\kappa_3 + \eta_3)}\right\},$$

then

$$\|\hat{x}^\star - x^\star\| \leq \frac{2\sqrt{2\eta_0}}{\sqrt{\underline{\nu}}}. \tag{40}$$

Clearly when $\eta_3 \leq D$ for some $D > 0$ and $\epsilon_3 \leq \frac{3}{4}\underline{\nu}/(\kappa_3 + D)$ then

$$\min\left\{\epsilon_3, \frac{3\underline{\nu}}{4(\kappa_3 + \eta_3)}\right\} \geq \min\left\{\epsilon_3, \frac{3\underline{\nu}}{4(\kappa_3 + D)}\right\} = \epsilon_3.$$

Putting everything together, we can conclude for every $D > 0$ there exists a constant

$$q_0 := q_0(g, x_0, \underline{\nu}, \bar{\nu}, D) \geq 1$$

such that whenever $\max\{\epsilon, \eta_0, \eta_1, \eta_2\} \leq 1/q_0$ and $\eta_3 \leq D$

$$\|\hat{x}^\star - x^\star\| \leq \frac{2\sqrt{2\eta_0}}{\sqrt{\underline{\nu}}} =: Q_0\sqrt{\eta_0}. \tag{41}$$

*Throughout the remainder of the proof, we shall assume $\max\{\epsilon, \eta_0, \eta_1, \eta_2\} \leq 1/q_0$ and $\eta_3 \leq D$ so that (41) holds.

Bound on $\|x(t) - \hat{x}(t)\|$ for large t.

Next we obtain a bound on $\|x(t) - \hat{x}(t)\|$ for large $t > 0$. Let $\mathbf{H}$ and $\hat{\mathbf{H}}$ be short for $H_g(x^\star)$ and $H_{\hat{g}}(\hat{x}^\star)$, respectively. We proceed with a linearization of the flows near the critical points. Let $\nu > \underline{\nu}$, but close enough such that all the eigenvalues of $\mathbf{H}$ are still in $(-\infty, -\nu)$. By combining (36) and (41)

$$\|\hat{\mathbf{H}} - \mathbf{H}\| \leq \eta_2 + d^{\frac{3}{2}}\kappa_3 Q_0\sqrt{\eta_0}. \tag{42}$$

Choose $\nu > \nu_2 > \nu_1 > \underline{\nu}$. Clearly the eigenvalues of $\mathbf{H}$ are also in $(-\infty, -\nu_2)$. Suppose that η_0 and η_2 are small enough that

$$\eta_2 + d^{\frac{3}{2}}\kappa_3 Q_0\sqrt{\eta_0} < \nu - \nu_2.$$

Thus $\|\hat{\mathbf{H}} - \mathbf{H}\| \leq \nu - \nu_2$ and by Weyl's inequality the eigenvalues of $\hat{\mathbf{H}}$ are in

$$(-\infty, -\nu + (\nu - \nu_2)) = (-\infty, -\nu_2). \tag{43}$$

Recall that WLOG we assume that $S = \mathcal{L}_g(g(x_0)/2)$. By the definition of S, clearly there is an $r_+ > 0$ such that $\bar{B}(x^\star, r_+) \subset S$. Note that for any $D > 0$ fixed the constant $q_0 \geq 1$ can be taken large enough so that (29), (31), (33), (34), (36) and (41) hold simultaneously. Fix an $\epsilon > 0$ small enough so that this is the case, and also such that $\sqrt{\epsilon} < (\sqrt{\underline{\nu}/2})r_+/2$. Recall the constants (38) and note that $\epsilon_2 \geq \epsilon$. Then recall by (33) there is a t_ϵ (depending on ϵ and the trajectory $x(t)$) such that

$$\|\hat{x}(t) - x^\star\| \leq \sqrt{2\epsilon_2/\underline{\nu}}, \qquad \text{for all } t \geq t_\epsilon,$$

which in combination with (41) gives

$$\|\hat{x}(t) - \hat{x}^\star\| \le \sqrt{2\epsilon_2/\underline{\nu}} + Q_0\sqrt{\eta_0}, \qquad \text{for all } t \ge t_\epsilon. \tag{44}$$

Also by (25) for all $t \ge t_\epsilon$, where $t_\epsilon > 0$ is large enough,

$$\|x(t) - x^\star\| \le r_+/2. \tag{45}$$

We see by (41) that when η_0 and η_1 are small enough we get $\bar{B}(\hat{x}^\star, r_+/2) \subset \bar{B}(x^\star, r_+)$ and we see by (44) that when η_0 and η_1 are small enough, $\|\hat{x}(t) - \hat{x}^\star\| \le r_+/2$ (note that this is possible since we have fixed $\sqrt{\epsilon} < (\sqrt{\underline{\nu}/2})r_+/2$). Setting $r_\ddagger = r_+/2$ and

$$t_\ddagger = t_\epsilon, \tag{46}$$

we get that

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

and

$$x(t) \in \bar{B}(x^\star, r_\ddagger) \quad \text{and} \quad \hat{x}(t) \in \bar{B}(\hat{x}^\star, r_\ddagger), \qquad \text{for any } t \ge t_\ddagger, \tag{47}$$

when η_0, η_1, and η_2 are small enough and $\eta_3 \le D$, and also keeping (45) in mind. (Note that $t_\ddagger$ depends only on g and the trajectory $x(t)$).

Letting

$$x_\ddagger(t) = x(t) - x^\star \text{ and } \hat{x}_\ddagger(t) = \hat{x}(t) - \hat{x}^\star,$$

by a Taylor expansion, for all $t \ge t_\ddagger$ we have

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

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

The difference gives

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

Claim 3 *We get after integrating* (50),

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

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

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

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

$$e^{t\mathbf{H}}(x^\star - \hat{x}^\star) = -\left(x'_\ddagger(t) - \hat{x}'_\ddagger(t)\right) + \int_0^t e^{(t-s)\mathbf{H}}\big[(\mathbf{H} - \hat{\mathbf{H}})\hat{x}_\ddagger(s) + R(s) - \hat{R}(s)\big]\mathrm{d}s. \tag{53}$$

Putting (53) in (52) we get (50). This verifies Claim 3.

Now since all of the eigenvalues of $\mathbf{H}$ are in $(-\infty, -\nu)$ we have

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

Using this fact with the triangle inequality along with (8), (42) and the inequalities in (48) and (49) we get

$$\|x_\ddagger(t) - \hat{x}_\ddagger(t)\|$$
$$\le e^{-\nu t}\|x^\star - \hat{x}^\star\| + \int_0^t e^{-\nu(t-s)}\left[\Delta\|\hat{x}_\ddagger(s)\| + \sqrt{d}\left(\frac{\kappa_3}{2}\|x_\ddagger(s)\|^2 + \frac{\kappa_3+\eta_3}{2}\|\hat{x}_\ddagger(s)\|^2\right)\right]\mathrm{d}s, \tag{54}$$

where

$$\Delta = \eta_2 + d^{\frac{3}{2}}\kappa_3 Q_0\sqrt{\eta_0}.$$

Recall that by Lemma 4, for some $C_4 = C_4(g, x_0)$,

$$\|x_\ddagger(t)\| \le C_4 e^{-\nu_1 t} \text{ for all } t \ge 0. \tag{55}$$

Claim 4. *For $\epsilon > 0$, η_0, η_1, and η_2 small enough and that $\eta_3 \le D$ so that (41), (43) and (47) hold, there is a constant $C'_4 := C'_4(g, x_0, \underline{\nu}, \bar{\nu}, \epsilon, D)$ such that*

$$\|\hat{x}_\ddagger(t)\| \le \max C'_4 e^{-\nu_1 t}, \quad \text{for all } t \ge 0. \tag{56}$$

Proof. We assume WLOG that $S = \mathcal{L}_g\left(g\left(x_0\right)/2\right)$ and is compact. Thus

$$\sup_{x,y\in S}\|x - y\| = L < \infty. \tag{57}$$

Let $\hat{\kappa}_3$ be short for $\kappa_3(\widehat{g}, S)$. We have that,

$$\hat{\kappa}_3 \le \kappa_3 + \eta_3 \le \kappa_3 + D.$$

We assume that $\epsilon > 0$, η_0, η_1, and η_2 are small enough and that $\eta_3 \le D$ so that (41) and (47) hold.

A Taylor expansion of $\nabla\widehat{g}$ at $x \in \bar{B}(\widehat{x}^\star, r_0)$ gives

$$\widehat{\nabla g}(x) = \widehat{\mathbf{H}}(x - \widehat{x}^\star) + \widehat{R}(x, \widehat{x}^\star), \tag{58}$$

with

$$\|\widehat{R}(x, \widehat{x}^\star)\| \le \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}\|x - \widehat{x}^\star\|^2.$$

Therefore by (58) and $\widehat{x}'\left(t\right) = \widehat{\nabla g}(\widehat{x}\left(t\right))$, we have,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\widehat{x}(t) - \widehat{x}^\star\right) - \widehat{\mathbf{H}}\left(\widehat{x}(t) - \widehat{x}^\star\right) = \widehat{R}\left(\widehat{x}(t), \widehat{x}^\star\right), \tag{59}$$

and since $\widehat{x}(0) = x_0$ and $\widehat{x}\,(t)$ satisfies the differential equation (59) it is readily checked that

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

Since all the eigenvalues of $\widehat{\mathbf{H}}$ are in $(-\infty, -\nu_2)$ we have

$$\left\| e^{\alpha\widehat{\mathbf{H}}} \right\| \le e^{-\nu_2\alpha}, \quad \text{for all } \alpha > 0.$$

Then,

$$\|\widehat{x}(t) - \widehat{x}^\star\| \le e^{-\nu_2 t}\|\widehat{x}_0 - \widehat{x}^\star\| + \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}\int_0^t e^{-\nu_2(t-s)}\|\widehat{x}(s) - \widehat{x}^\star\|^2\mathrm{d}s. \tag{60}$$

Set

$$\widehat{u}(t) = e^{\nu_2 t}\|\widehat{x}(t) - \widehat{x}^\star\|$$

and

$$\widehat{U}(t) = \|x_0 - \widehat{x}^\star\| + \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}\int_0^t e^{\nu_2 s}\|\widehat{x}(s) - \widehat{x}^\star\|^2\mathrm{d}s. \tag{61}$$

Thus by (60), $\widehat{u}(t) \le \widehat{U}(t)$ and $\widehat{U}'(t) = \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}e^{-\nu_2 t}\widehat{u}^2(t)$, so

$$\begin{aligned}\frac{\widehat{U}'(t)}{\widehat{U}(t)} &= \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}e^{-\nu_2 t}\widehat{u}(t)\frac{\widehat{u}(t)}{\widehat{U}(t)} \\ &\le \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}e^{-\nu_2 t}\widehat{u}(t) = \widehat{\kappa}_3\tfrac{\sqrt{d}}{2}\|\widehat{x}(t) - \widehat{x}^\star\| \\ &\le \frac{\sqrt{d}}{2}(\kappa_3 + D)\|\hat{x}(t) - \hat{x}^\star\|. \end{aligned} \tag{62}$$

Recall that $\nu_2 > \nu_1 > \underline{\nu}$. We can choose WLOG $r_\ddagger$ in (47) small enough so that

$$r_\ddagger \le \left[\frac{\sqrt{d}}{2}(\kappa_3 + D)\right]^{-1}(\nu_2 - \nu_1).$$

Assuming that this is the case, we get from (62)

$$\frac{\hat{U}'(t)}{\hat{U}(t)} \le \nu_2 - \nu_1, \quad \text{for all } t \ge t_\ddagger.$$

By integrating between $t_\ddagger$ and t, we deduce that

$$\log\widehat{U}(t) \le \log\widehat{U}(t_\ddagger) + (\nu_2 - \nu_1)(t - t_\ddagger),$$

and so

$$\|\hat{x}(t) - \hat{x}^\star\| = e^{-\nu_2 t}\hat{u}(t) \le e^{-\nu_2 t}\hat{U}(t) \le c_1 e^{-\nu_1 t}, \quad \text{for all } t \ge t_\ddagger,$$

with

$$c_1 := \widehat{U}(t_\ddagger)e^{-(\nu_2-\nu_1)t_\ddagger}.$$

For $t < t_\ddagger$, we simply have

$$\|\widehat{x}(t) - \widehat{x}^\star\| \le c_2 e^{-\nu_1 t},$$

where

$$c_2 = \max_{0 \le t \le t_\ddagger} \|\widehat{x}(t) - \widehat{x}^\star\| e^{\nu_1 t}.$$

Notice that by (57) and (61), keeping in mind that we always assume by Claim 1 that η_0 is sufficiently small so that $\hat{x}(t) \in S$, for all $t \ge 0$,

$$\widehat{U}(t_\ddagger) = \|x_0 - \widehat{x}^\star\| + \widehat{\kappa}_3 \tfrac{\sqrt{d}}{2} \int_0^{t_\ddagger} e^{\nu_2 s} \|\widehat{x}(s) - \widehat{x}^\star\|^2 \mathrm{d}s$$

$$\le L + (\kappa_3 + D)\, \tfrac{\sqrt{d}L^2}{2\nu} e^{\nu_2 t_\ddagger}$$

and thus

$$c_1 \le \left(L + (\kappa_3 + D)\, \tfrac{\sqrt{d}L^2}{2\nu} e^{\nu t_\ddagger}\right) e^{-(\nu_2 - \nu_1) t_\ddagger} =: \overline{c}_1$$

and

$$c_2 \le L e^{\nu_1 t_\ddagger} =: \overline{c}_2.$$

Hence (56) holds with the constant $C_4' = \max(\overline{c}_1, \overline{c}_2)$, which proves Claim 4.
This, in combination with (55), shows that for all $t \ge 0$

$$\max(\|x_\ddagger(t)\|, \|\hat{x}_\ddagger(t)\|) \le C_M e^{-\nu_1 t}, \tag{63}$$

where $C_M = \max(C_4, C_4')$.
We shall use (63) to bound the integral in (54). We have by (63) and $\nu > \nu_1 > \underline{\nu}$

$$\int_0^t e^{-\nu(t-s)} \left[\Delta \|\hat{x}_\ddagger(s)\| + \sqrt{d}\left(\frac{\kappa_3}{2}\|x_\ddagger(s)\|^2 + \frac{\kappa_3 + \eta_3}{2}\|\hat{x}_\ddagger(s)\|^2\right)\right] \mathrm{d}s,$$

$$\le \int_0^t e^{-\underline{\nu}(t-s)} \left[\Delta C_M e^{-\nu_1 s} + \sqrt{d}\left(\frac{\kappa_3}{2} C_M^2 e^{-2\nu_1 s} + \frac{\kappa_3 + \eta_3}{2} C_M^2 e^{-2\nu_1 s}\right)\right] \mathrm{d}s$$

$$\le \int_0^t e^{-\underline{\nu}(t-s)} \left[\Delta C_M e^{-\nu_1 s} + \sqrt{d}\,(\kappa_3 + \eta_3)\, C_M^2 e^{-2\underline{\nu} s}\right] \mathrm{d}s$$

$$\le C_M e^{-\underline{\nu} t} \left[\Delta \frac{1 - e^{-(\nu_1 - \underline{\nu})t}}{\nu_1 - \underline{\nu}} + \sqrt{d}\,(\kappa_3 + \eta_3)\, C_M \frac{1 - e^{-\underline{\nu} t}}{\underline{\nu}}\right].$$

Applying this bound in (54) we get

$$\|x_\ddagger(t) - \hat{x}_\ddagger(t)\|$$

$$\le e^{-\underline{\nu} t}\|x^* - \hat{x}^*\| + C_M e^{-\underline{\nu} t} \left[\Delta \frac{1 - e^{-(\nu_1 - \underline{\nu})t}}{\nu_1 - \underline{\nu}} + \sqrt{d}\,(\kappa_3 + \eta_3)\, C_M \frac{1 - e^{-\underline{\nu} t}}{\underline{\nu}}\right]. \tag{64}$$

By the triangle inequality

$$\|x(t) - \hat{x}(t)\| \le \|x^* - \hat{x}^*\| + \|x_\ddagger(t) - \hat{x}_\ddagger(t)\|$$

and using (41) and (64) we deduce that for all $t \geq t_{\ddagger}$,

$$\|x(t) - \hat{x}(t)\|$$
$$\leq (1 + e^{-\underline{\nu}t})Q_0\sqrt{\eta_0} + C_M e^{-\underline{\nu}t}\left[\Delta\frac{1 - e^{-(\nu_1 - \underline{\nu})t}}{\nu_1 - \underline{\nu}} + \sqrt{d}(\kappa_3 + \eta_3)C_M\frac{1 - e^{-\underline{\nu}t}}{\underline{\nu}}\right].$$

Keeping in mind that we assume that $\eta_3 \leq D$, η_0, η_1 and $\eta_2 \leq 1/q_0 \leq 1$, which makes $\Delta \leq 1 + d^{3/2}\kappa_3 Q_0$. Therefore for $t_{\ddagger} = t_\epsilon > 0$ suitably large we get that for some constant $Q_1 = Q_1(g, x_0, \underline{\nu}, \overline{\nu}, \epsilon, D) > 0$,

$$\|x(t) - \hat{x}(t)\| \leq Q_1\big(\sqrt{\eta_0} + e^{-\underline{\nu}t}\big), \quad \text{for all } t \geq t_\epsilon. \tag{65}$$

(Recall that in (46) we defined $t_{\ddagger} := t_\epsilon$.)

Notice that since g is in C^3, there is an $\epsilon > 0$ such that all the eigenvalues of $H_g(x)$ exceed $-\overline{\nu}$ when $x \in \overline{B}(x^\star, \epsilon)$, $\epsilon > 0$, being fixed. Note that this implies that ∇g is Lipschitz on $\overline{B}(x^\star, \epsilon)$ with constant $\overline{\nu}$. Let t_ϵ be large enough such that for all $t \geq t_\epsilon$, $x(t) \in B(x^\star, \epsilon/2)$. Assume that η_0 is small enough so that $\|\widehat{x}^\star - x^\star\| \leq \epsilon/2$, which is possible by (41). Moreover by (65) for a suitably large $t_\epsilon > 0$ and small $\eta_0 > 0$ with $\eta_2 \leq 1/q_0 \leq 1$ and $\eta_3 \leq D$

$$\|x(t) - \hat{x}(t)\| \leq Q_1\left(\sqrt{\eta_0} + e^{-\underline{\nu}t_\epsilon}\right) \leq \epsilon/2, \quad \text{for all } t \geq t_\epsilon, \tag{66}$$

Then we also have $\widehat{x}(t) \in \overline{B}(x^\star, \epsilon)$ for all $t \geq t_\epsilon$. We may now apply Lemma 5 with $\mathcal{S} = \overline{B}(x^\star, \epsilon)$, $t_0 = t_\epsilon$, $x_1 = x(t_\epsilon)$, $y_1 = \widehat{x}(t_\epsilon)$, keeping in mind that ∇g is Lipschitz on $\overline{B}(x^\star, \epsilon)$ with constant $\overline{\nu}$, to get

$$\|x(t) - \hat{x}(t) - (x(t_\epsilon) - \hat{x}(t_\epsilon))\| \leq \frac{\eta_1}{\overline{\nu}}e^{\overline{\nu}t}, \quad \forall t \geq t_\epsilon. \tag{67}$$

Bound on $\|x(t) - \hat{x}(t)\|$ for small t.

Since ϵ is fixed, by (30) we also get by Lemma 5 the following bound on $\|x(t) - \hat{x}(t)\|$ for small $t \geq 0$

$$\|x(t) - \hat{x}(t)\| \leq \frac{\eta_1}{\sqrt{d}\kappa_2}e^{\sqrt{d}\kappa_2 t} \leq \frac{\eta_1 e^{\left|\sqrt{d}\kappa_2 - \overline{\nu}\right|t_\epsilon}}{\sqrt{d}\kappa_2}e^{\overline{\nu}t}, \quad 0 \leq t \leq t_\epsilon. \tag{68}$$

Completion of the Proof of Theorem 2

Combining (67) and (68) we get

$$\|x(t) - \hat{x}(t)\| \leq Q_2\eta_1 e^{\overline{\nu}t}, \quad \forall t \geq 0, \tag{69}$$

for some constant $Q_2 = Q_2(g, x_0, \underline{\nu}, \overline{\nu}, \epsilon, D)$. Then from (65) and (69) we arrive at

$$\|x(t) - \hat{x}(t)\| \leq Q_3 \min\left[\sqrt{\eta_0} + e^{-\underline{\nu}t}, \eta_1 e^{\overline{\nu}t}\right], \quad \forall t \geq 0, \tag{70}$$

for some constant $Q_3 = Q_3(g, x_0, \underline{\nu}, \overline{\nu}, \epsilon, D)$. Indeed, the curves $t \mapsto Q_1\left(\sqrt{\eta_0} + e^{-\underline{\nu}t}\right)$ and $t \mapsto Q_2\eta_1 e^{\overline{\nu}t}$ intersect at some point t larger than t_ϵ if

$$Q_1\left(\sqrt{\eta_0} + e^{-\underline{\nu}t_\epsilon}\right) \geq Q_2\eta_1 e^{\overline{\nu}t_\epsilon} \Longleftrightarrow Q_1 \geq Q_2\frac{\eta_1 e^{\overline{\nu}t_\epsilon}}{\sqrt{\eta_0} + e^{-\underline{\nu}t_\epsilon}},$$

and this is guaranteed if we choose Q_1 large enough that $Q_1 \geq Q_2 \frac{1}{q_0} e^{(\underline{\nu}+\overline{\nu})t_\epsilon}$. (Recall the bounds in (41) and note that Q_2 does not depend on Q_1).

We are now ready to finish the proof of Theorem 2. We shall show that the bound (14) follows from (70). To verify this, we start with

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

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

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

Suppose that η_0 is small enough so that $t_0 \geq t_\ddagger$.

- When $t \geq t_0$, then we simply observe that $B(t) \leq \eta_0^{1/2}$.
- When $t \leq t_0$, 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 $[t_1, \infty)$, and that
$$\min\{\sqrt{\eta_0}, e^{-\underline{\nu}t}\} = \begin{cases} \eta_1 e^{\overline{\nu}t} \text{ when } t \leq t_1 \\ e^{-\underline{\nu}t}, \text{ when } t \geq t_1. \end{cases}$$
- When $t_1 \geq t_0$ and $t \leq t_0$, we see that $B(t) = \eta_1 e^{\overline{\nu}t_0} \leq \eta_1 \eta_0^{-\frac{\overline{\nu}}{2\underline{\nu}}}$.
- When $t_1 < t_0$ and $t \leq t_0$, then $B(t) \leq B(t_1) = e^{-\underline{\nu}t_1} \leq \eta_1^{\frac{\underline{\nu}}{\underline{\nu}+\overline{\nu}}}$.

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

Hence, we worked (70) into

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

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

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

and

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

Using these equivalences we deduce that

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

Putting together our bounds on $\|x(t) - \hat{x}(t)\|$ for $t > 0$ large and $t \geq 0$ small, we can now conclude from (70) that for all $\epsilon > 0$ small enough and all $D > 0$ there exists a constant

$C := C(g, x_0, \underline{\nu}, \bar{\nu}, D) \geq 1$ and a function $F(g, x_0, \underline{\nu}, \bar{\nu}, \epsilon, D)$ of ϵ and D such that, whenever $\max\{\epsilon, \eta_0, \eta_1, \eta_2\} \leq 1/C$ and $\eta_3 \leq D$, $\hat{x}(t)$ is defined for all $t \geq 0$ and

$$\sup_{t \geq 0} \|x(t) - \hat{x}(t)\| \leq F(g, x_0, \underline{\nu}, \bar{\nu}, \epsilon, D) \max\left\{\sqrt{\eta_0}, \eta_1^{\delta}\right\}, \tag{71}$$

holds, where $\delta := \underline{\nu}/\left(\underline{\nu} + \overline{\nu}\right)$. We now take $\epsilon = 1/C$ in (71). This completes the proof of Theorem 2. □

Acknowledgements
The second named author thanks the editors for inviting him to contribute to this proceedings. He had intended give a talk based on the results in this paper at the 2017 High Dimensional Probability VIII conference. Unfortunately he had to cancel his participation due heath reasons. Part of this paper was written while he was a guest of Australian National University.

References

[1] E. Arias-Castro, D.M. Mason and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. J. Mach. Learn. Res. **17**, Paper No. 43 (2016)

[2] E. Arias-Castro, D.M. Mason and B. Pelletier. Errata: On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. J. Mach. Learn. Res. **17**, Paper No. 206 (2016)

[3] R. Bhatia. Matrix Analysis. Graduate Texts in Mathematics, **169**. Springer-Verlag, New York, 1997.

[4] P. Deheuvels and D.M. Mason. General asymptotic confidence bands based on kernel-type function estimators. Stat. Inference Stoch. Process. **7**, 225–277 (2004)

[5] 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)

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

[7] 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**, 32–40 (1975)

[8] 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)

[9] J.A. Hartigan. Clustering Algorithms. Wiley, New York, 1975

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

[11] D.M. Mason. Proving consistency of non-standard kernel estimators. Stochastic Inference for Stochastic Processes. **15**, 151–176 (2012)

[12] 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)

[13] D.M. Mason and J. Swanepoel. Errata: A general result on the uniform in bandwidth consistency of kernel-type function estimators. Test **24**, 205–206 (2015)