Hencl, Stanislav; Koski, Aleksis; Onninen, Jani

Sobolev homeomorphic extensions from two to three dimensions

Published in:
Journal of Functional Analysis

DOI:
10.1016/j.jfa.2024.110371

Published: 01/05/2024

Document Version
Publisher's PDF, also known as Version of record

Journal of Functional Analysis 286 (2024) 110371

Contents lists available at ScienceDirect

Journal of Functional Analysis

journal homepage: www.elsevier.com/locate/jfa

Regular Article

Sobolev homeomorphic extensions from two to three dimensions ☆

Stanislav Hencl [a], Aleksis Koski [b,*], Jani Onninen [c,d]

[a] *Charles University, Department of Mathematical Analysis, Sokolovská 83, 186 00, Prague 8, Czech Republic*
[b] *Department of Mathematics and Systems Analysis, P.O. Box 11100, FI-00076, Aalto University, Finland*
[c] *Department of Mathematics, Syracuse University, Syracuse, NY 13244, USA*
[d] *Department of Mathematics and Statistics, P.O. Box 35 (MaD), FI-40014, University of Jyväskylä, Finland*

ARTICLE INFO

Article history:
Received 1 August 2022
Accepted 28 January 2024
Available online 15 February 2024
Communicated by Guido De Philippis

MSC:
primary 46E35, 58E20

Keywords:
Sobolev homeomorphisms
Sobolev extensions
L^1-Beurling-Ahlfors extension

ABSTRACT

We study the basic question of characterizing which boundary homeomorphisms of the unit sphere can be extended to a Sobolev homeomorphism of the interior in 3D space. While the planar variants of this problem are well-understood, completely new and direct ways of constructing an extension are required in 3D. We prove, among other things, that a Sobolev homeomorphism $\varphi\colon \mathbb{R}^2 \xrightarrow{\text{onto}} \mathbb{R}^2$ in $W^{1,p}_{\text{loc}}(\mathbb{R}^2,\mathbb{R}^2)$ for some $p \in [1,\infty)$ admits a homeomorphic extension $h\colon \mathbb{R}^3 \xrightarrow{\text{onto}} \mathbb{R}^3$ in $W^{1,q}_{\text{loc}}(\mathbb{R}^3,\mathbb{R}^3)$ for $1 \leqslant q < \frac{3}{2}p$. Such an extension result is nearly sharp, as the bound $q = \frac{3}{2}p$ cannot be improved due to the Hölder embedding. The case $q = 3$ gains an additional

☆ S. Hencl was supported by the grant GAČR P201/21-01976S A. Koski was supported by the Academy of Finland grant number 307023, the ERC Advanced Grant 834728, received financial support from the Spanish Ministry of Science and Innovation through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000904-S and MTM2017-85934-C3-2-P2), and from the CAM through the line of excellence for University Teaching Staff between CM and UAM. J. Onninen was supported by the NSF grant DMS-2154943.

* Corresponding author.
E-mail addresses: hencl@karlin.mff.cuni.cz (S. Hencl), aleksis.koski@gmail.com (A. Koski), jkonnine@syr.edu (J. Onninen).

https://doi.org/10.1016/j.jfa.2024.110371

interest as it also provides an L^1-variant of the celebrated Beurling-Ahlfors quasiconformal extension result.

1. Introduction

Throughout this paper $\mathbb{B}$ denotes the unit ball in $\mathbb{R}^3$ and $\mathbb{S} = \partial\mathbb{B}$. We study the following *3D–Sobolev homeomorphic extension problem*.

Problem 1. Suppose that a homeomorphism $\varphi\colon \mathbb{S} \xrightarrow{\text{onto}} \mathbb{S}$ admits a continuous extension to $\mathbb{B}$ in the Sobolev space $W^{1,q}(\mathbb{B},\mathbb{R}^3)$ for some $q \in [1,\infty)$. Does the map φ also admit a homeomorphic extension to $\mathbb{B}$ of class $W^{1,q}(\mathbb{B},\mathbb{R}^3)$?

Every boundary homeomorphism $\varphi\colon \mathbb{S} \xrightarrow{\text{onto}} \mathbb{S}$ extends as a homeomorphism to the ball $\mathbb{B}$. On the other hand, according to a famous result of Gagliardo [13], for $1 < q < \infty$, the mapping φ is the Sobolev trace of some (possibly non-homeomorphic) mapping in $W^{1,q}(\mathbb{B},\mathbb{R}^3)$ if and only if it belongs to the fractional Sobolev space $W^{1-\frac{1}{q},q}(\mathbb{S},\mathbb{R}^3)$; that is,

$$\int_{\mathbb{S}}\int_{\mathbb{S}} \frac{|\varphi(x)-\varphi(y)|^q}{|x-y|^{q+1}}\,\mathrm{d}x\,\mathrm{d}y < \infty\,. \tag{1.1}$$

Note that the 2D result [31] that every boundary homeomorphism $\varphi\colon \partial\mathbb{D} \xrightarrow{\text{onto}} \partial\mathbb{D}$ extends as a $W^{1,q}$-homeomorphism, $q < 2$, to the unit disk $\mathbb{D} \subset \mathbb{R}^2$ has no counterpart in higher dimensions. Indeed, there are boundary homeomorphisms from $\mathbb{S}$ onto itself that do not even admit a continuous Sobolev extension in $W^{1,q}(\mathbb{B},\mathbb{R}^3)$ for any $q > 1$, see Example 3.1.

First we give a discrete variant of (1.1); that is, we characterize the boundary homeomorphisms that admit a Sobolev extension in $W^{1,q}(\mathbb{B},\mathbb{R}^3)$ when $q > 2$.

Theorem 1.1. *Let $\varphi\colon \mathbb{S} \xrightarrow{\text{onto}} \mathbb{S}$ be a homeomorphism and $q \in (2,\infty)$. Suppose that $\tilde{\mathcal{D}}_k$ is a dyadic decomposition of $\mathbb{S}$ into closed bi-Lipschitz squares of diameter $c2^{-k}$. Then φ satisfies (1.1) if and only if*

$$\sum_{k=1}^{\infty} 2^{k(q-3)} \sum_{\tilde{Q}_j \in \tilde{\mathcal{D}}_k} \left[\operatorname{diam}\varphi(\tilde{Q}_j)\right]^q < \infty\,. \tag{1.2}$$

For the precise definition of $\tilde{\mathcal{D}}_k$ we refer to Definition 2.1.

The corresponding 2D–Sobolev homeomorphic extension problem [22] has an easy answer thanks to the available analytic methods of constructing 2D-Sobolev homeomorphisms. Indeed, let $\mathbb{D}$ be the unit disk in $\mathbb{R}^2$ and $q \in [1,\infty)$ then a boundary homeomorphism $\varphi\colon \partial\mathbb{D} \xrightarrow{\text{onto}} \partial\mathbb{D}$ admits a homeomorphic extension to $\overline{\mathbb{D}}$ in $W^{1,q}(\mathbb{D},\mathbb{R}^2)$

if and only if it admits a continuous extension to $\overline{\mathbb{D}}$ in $W^{1,q}(\mathbb{D},\mathbb{R}^2)$. This follows from the Radó-Kneser-Choquet (RKC) theorem [11] for $q \leqslant 2$. The RKC theorem asserts that a homeomorphic boundary value $\varphi\colon \partial\mathbb{D} \overset{\text{onto}}{\longrightarrow} \partial\mathbb{D}$ admits a homeomorphic harmonic extension of $\mathbb{D}$. The harmonic extension belongs to $W^{1,q}(\mathbb{D},\mathbb{R}^2)$ for all $q < 2$ and to $W^{1,2}(\mathbb{D},\mathbb{R}^2)$ exactly when is in the trace space of $W^{1,2}(\mathbb{D},\mathbb{R}^2)$. Similarly the q-harmonic variants of the RKC theorem [2] solve the 2D extension problem for $q > 2$. An analogous approach fails in higher dimensions. Indeed, Laugesen [23] constructed a self-homeomorphism of the sphere $\mathbb{S}$ in $\mathbb{R}^3$ whose harmonic extension to the ball $\mathbb{B}$ is not injective. Thus, the 3D extension problem requires new methods of constructing Sobolev homeomorphisms.

Our main result tells us that the searched homeomorphic extension exists if the boundary homeomorphism satisfies a strengthened version of the condition (1.2).

Theorem 1.2. *Let $q \in (1,\infty)$. Suppose that $\tilde{\mathcal{D}}_k$ is a dyadic decomposition of $\mathbb{S}$ into closed bi-Lipschitz squares of diameter $c2^{-k}$. If a homeomorphism $\varphi\colon \mathbb{S} \overset{\text{onto}}{\longrightarrow} \mathbb{S}$ satisfies*

$$\sum_{k=1}^{\infty} 2^{k(q-3)} \sum_{\tilde{Q}_j \in \tilde{\mathcal{D}}_k} \left[\mathcal{H}^1\big(\varphi(\partial \tilde{Q}_j)\big)\right]^q < \infty\,, \tag{1.3}$$

then it admits a homeomorphic extension $h\colon \overline{\mathbb{B}} \overset{\text{onto}}{\longrightarrow} \overline{\mathbb{B}}$ in $W^{1,q}(\mathbb{B},\mathbb{R}^3)$.

Here $\mathcal{H}^1$ stands for 1-dimensional Hausdorff measure and so $\mathcal{H}^1\big(\varphi(\partial \tilde{Q}_j)\big)$ measures the length of the curve $\varphi(\partial \tilde{Q}_j)$.

For a Sobolev homeomorphism $\varphi\colon \mathbb{S} \overset{\text{onto}}{\longrightarrow} \mathbb{S}$ the trivial radial extension $h(x) = |x|\varphi(x)$ produces a self homeomorphism of $\overline{\mathbb{B}}$ which has the same Sobolev regularity as the given boundary map φ. Clearly, such an extension is far from being optimal. Our next result, however, nearly characterizers the first order Sobolev spaces that admit a Sobolev homeomorphic extension to $\mathbb{B}$.

Theorem 1.3. *Let $\varphi\colon \mathbb{S} \overset{\text{onto}}{\longrightarrow} \mathbb{S}$ be a homeomorphism in $W^{1,p}(\mathbb{S},\mathbb{R}^3)$ for some $p \in [1,\infty)$. Then φ admits a homeomorphic extension $h\colon \overline{\mathbb{B}} \overset{\text{onto}}{\longrightarrow} \overline{\mathbb{B}}$ in $W^{1,q}(\mathbb{B},\mathbb{R}^3)$ for $1 \leqslant q < \frac{3}{2}p$.*

For the sharpness of this result we refer to the general embedding result by Sickel and Triebel [28, Theorem 3.2.1]. Namely for $p \in (1,\infty)$ we have $W^{1,p}(\mathbb{S},\mathbb{R}^3) \subset W^{1-\frac{1}{q},q}(\mathbb{S},\mathbb{R}^3)$ if and only if $q \leqslant \frac{3}{2}p$. Even assuming that the mappings are homeomorphisms does not improve the inclusion at least when $p \geqslant 2$, see Example 3.2. We do not know if one can take $q = \frac{3}{2}p$ in Theorem 1.3.

Theorem 1.3 follows from Theorem 1.2. On the contrary there are self homeomorphisms of $\mathbb{S}$ which satisfy (1.3) and do not belong to any Sobolev class $W^{1,p}(\mathbb{S},\mathbb{R}^3)$, $p \geqslant 1$, see Example 3.3.

In topology and analysis, a number of extension problems have been studied. A demand for Sobolev homeomorphic extension problems comes from the variational

approach to Geometric Function Theory (GFT) [4,15,21,26] and mathematical models of Nonlinear Elasticity (NE) [3,6,9]. Both theories enquire into homeomorphisms $h\colon \mathbb{X} \xrightarrow{\text{onto}} \mathbb{Y}$ of smallest *stored energy*

$$\mathsf{E}_{\mathbb{X}}[h] = \int_{\mathbb{X}} \mathbf{E}(x, h, Dh)\,\mathrm{d}x\,, \qquad \mathbf{E}\colon \mathbb{X} \times \mathbb{Y} \times \mathbb{R}^{n\times n}$$

where the so-called *stored energy function* $\mathbf{E}$ characterizes the mechanical and elastic properties of the material occupying the domains. In a pure displacement setting, typically an orientation-preserving boundary homeomorphism $\varphi\colon \partial\mathbb{X} \xrightarrow{\text{onto}} \partial\mathbb{Y}$ is given. The class of admissible deformations consists of Sobolev homeomorphisms or just Sobolev mappings $h\colon \overline{\mathbb{X}} \xrightarrow{\text{onto}} \overline{\mathbb{Y}}$ with non-negative Jacobian determinant $J_h(x) = \det Dh(x) \geqslant 0$ (an axiomatic assumption in NE) which coincides with φ on the boundary and having a finite stored energy. In such variational problems, a first issue to address is the non-emptiness of the class of admissible deformations; that is, to solve the corresponding Sobolev homeomorphic extension problem.

Note that an arbitrary orientation-preserving Sobolev homeomorphism h need not be *strictly orientation-preserving* in the sense that $J_h(x) = \det Dh(x) > 0$ almost everywhere. For every $q < 3$, there even exists a homeomorphism $h\colon \mathbb{B} \xrightarrow{\text{onto}} \mathbb{B}$ in $W^{1,q}(\mathbb{B}, \mathbb{R}^3)$ with $J_h(x) = 0$ for almost every $x \in \mathbb{B}$, see [14]. However, the homeomorphic extensions $h\colon \mathbb{B} \xrightarrow{\text{onto}} \mathbb{B}$ constructed in Theorem 1.3 and Theorem 1.2 are piecewise linear. Thus, they are strictly orientation-preserving provided that the given boundary homeomorphism itself preserves the orientation. In particular, these homeomorphisms have finite distortion. The theory of mappings of finite distortion arose out of a need to extend the ideas and applications of the classical theory of quasiconformal mappings to the degenerate elliptic setting [15,21]. We recall that a homeomorphism $h\colon \mathbb{X} \xrightarrow{\text{onto}} \mathbb{Y}$ of Sobolev class $W^{1,1}_{\text{loc}}(\mathbb{X}, \mathbb{R}^n)$ defined on a domain $\mathbb{X} \subset \mathbb{R}^n$ has *finite distortion* if

$$|Dh(x)|^n \leqslant K(x) J_h(x) \tag{1.4}$$

for some measurable function $1 \leqslant K(x) < \infty$. Here, $|Dh(x)|$ is the operator norm of the weak differential $Dh(x)\colon \mathbb{X} \to \mathbb{R}^n$ of h at a point $x \in \mathbb{X}$. We obtain *quasiconformal* mappings if $K \in L^\infty(\mathbb{X})$. There are several other distortion functions of great interest in GFT. Each of them is designed to measure the deviation from conformality of a given mapping $h\colon \mathbb{X} \to \mathbb{R}^n$ in terms of the tangent linear map $Dh(x)\colon \mathbb{R}^n \to \mathbb{R}^n$. The most interesting, from the applied point of view, is the inner distortion function. In NE one is typically provided information not only on the differential matrix, but also on its $(n-1) \times (n-1)$–minors; that is, the *cofactor matrix* $D^\sharp h$ called *co-differential* of h. Now, for a homeomorphism $h \in W^{1,1}_{\text{loc}}(\mathbb{X}, \mathbb{R}^n)$ of finite distortion we introduce its *inner distortion* function, to be the smallest $K_{\scriptscriptstyle I}(x) = K_{\scriptscriptstyle I}(x, f) \geqslant 1$ satisfying

$$|D^\sharp f(x)|^n = K_{\scriptscriptstyle I}(x) \cdot J_f(x)^{n-1}$$

The most pronounced extension result in GFT is the Beurling-Ahlfors quasiconformal extension theorem [7]. It states that a self-homeomorphism of the unit disk $\mathbb{D}$ is quasiconformal if and only if the boundary correspondence homeomorphism $\varphi\colon \partial\mathbb{D} \overset{\text{onto}}{\longrightarrow} \partial\mathbb{D}$ is quasisymmetric. The Beurling-Ahlfors result has found a number of applications in Teichmüller theory, Kleinian groups, conformal welding and dynamics, see e.g. [4,19]. It has generalized to the n-dimensional quasiconformal maps as well, first for $n = 3$ by Ahlfors [1] and then for $n = 4$ by Carleson [8]. A full n-dimensional version of the Beurling-Ahlfors extension is due to Tukia and Väisälä [30]. Their extension uses, among other things, Sullivan's theory [29] of deformations of Lipschitz embeddings. Moreover, Astala, Iwaniec, Martin and Onninen [5], as a part of their studies of deformations with smallest mean distortion, characterizes self homeomorphisms of the unit circle that admit a homeomorphic extension to the unit disk $\mathbb{D}$ with integrable distortion. This L^1–Beurling-Ahlfors extension theorem enjoys the following 3D-variant.

Theorem 1.4. *Let $\psi\colon \mathbb{S} \overset{\text{onto}}{\longrightarrow} \mathbb{S}$ be an orientation-preserving homeomorphism. Suppose that the inverse $\psi^{-1} = \varphi$ satisfies (1.3) with $q = 3$. Then ψ admits a homeomorphic extension $f\colon \overline{\mathbb{B}} \overset{\text{onto}}{\longrightarrow} \overline{\mathbb{B}}$ with integrable inner distortion.*

Theorem 1.4 is actually a relatively straightforward consequence of Theorem 1.2, thanks to an important connection between the conformal energy of a homeomorphism and the inner distortion function of the inverse mapping. Indeed it is easy to see, at least formally, that the pullback of the 3-form $K_I(y, f)\,\mathrm{d}y \in \wedge^3\mathbb{B}$ by the inverse mapping $f^{-1}\colon \mathbb{B} \overset{\text{onto}}{\longrightarrow} \mathbb{B}$ is equal to $|Df^{-1}(x)|^3\,\mathrm{d}x \in \wedge^3\mathbb{B}$. This observation is the key to the identity,

$$\int_{\mathbb{B}} |Dh(x)|^3\,\mathrm{d}x = \int_{\mathbb{B}} K_I(y, f)\,\mathrm{d}y\,, \qquad \text{where } h = f^{-1}\colon \mathbb{B} \overset{\text{onto}}{\longrightarrow} \mathbb{B}\,. \tag{1.5}$$

The optimal Sobolev regularity of deformations to guarantee the identity is well-understood today, [10,16,17,24]. In particular, if a homeomorphism $h\colon \mathbb{B} \overset{\text{onto}}{\longrightarrow} \mathbb{B}$ of finite distortion belongs to the Sobolev class $W^{1,3}(\mathbb{B}, \mathbb{R}^3)$, then the inverse $f = h^{-1}$ has integrable inner distortion. Thus, Theorem 1.4 simply follows from Theorem 1.2. It is worth noting that the borderline case in Theorem 1.3 (p = 3 and q = 2), if true, would have an interesting corollary. Namely, a homeomorphism $\psi\colon \mathbb{R}^2 \overset{\text{onto}}{\longrightarrow} \mathbb{R}^2$ of locally integrable distortion would then admit a homeomorphic extension $f\colon \mathbb{R}^3 \overset{\text{onto}}{\longrightarrow} \mathbb{R}^3$ with locally integrable inner distortion.

Acknowledgements. We would like to thank the referee for their many insightful comments and suggestions which particularly helped in improving the presentation of the paper considerably.

2. A discrete characterization, proof of Theorem 1.1

Let $\mathbb{I} = [a,b]^2$ be an initial square in $\mathbb{R}^2$. The standard *dyadic decomposition* of $\mathbb{I}$ consists of closed squares $\tilde{Q} \subset \mathbb{I}$ with sides parallel to the sides of $\mathbb{I}$ and of side length $l(\tilde{Q}) = 2^{-k}(b-a)$, $k = 1, 2, 3, \dots$; refers to the *k-th generation* in the construction. That is, the squares in the k-th generation have the form

$$\tilde{Q}_j = 2^{-k}(\mathbb{I} + v_j) \subset \mathbb{I}\,, \qquad \text{for some } v_j \in \mathbb{R}^2\,.$$

They cover $\mathbb{I}$ and have side length $2^{-k}(b-a)$. The collection of the k-th generation squares are denoted by $\tilde{\mathcal{D}}_k$. There are 2^{2k} squares in $\tilde{\mathcal{D}}_k$. The interiors of the squares in the same generation $\tilde{\mathcal{D}}_k$ are pairwise disjoint.

Let $\mathbb{Q}^3 = [0,1]^3$ be the unit cube in $\mathbb{R}^3$. We define the k-th generation dyadic decomposition of $\partial\mathbb{Q}^3$ as follows: first we divide each of the six faces of $\partial\mathbb{Q}$ into the k-th generation squares and then the k-th generation dyadic decomposition of $\partial\mathbb{Q}^3$ simply consists of the union of these closed squares.

Now, since $\overline{\mathbb{B}}$ is a bi-Lipschitz equivalent with $\mathbb{Q}^3$, defining a k-th generation dyadic decomposition of $\partial\mathbb{B} = \mathbb{S}$ can be easily induced from the above case.

Definition 2.1. Let $\Phi\colon \mathbb{R}^3 \to \mathbb{R}^3$ be a bi-Lipschitz map which takes $\mathbb{Q}^3$ onto $\overline{\mathbb{B}}$. Then the k-th generation dyadic decomposition of $\mathbb{S}$, denoted by $\tilde{\mathcal{D}}_k$, consists of $\Phi(\tilde{Q}_j)$, where $\tilde{Q}_j$ is a k-th generation dyadic square of $\partial\mathbb{Q}^3$.

Theorem 2.2. *Let $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a homeomorphism, $\mathbb{I}_R = [-R,R]^2 \subset \mathbb{R}^2$ for $R > 0$ and let $N \in \mathbb{N}$. Denote the collection of k-th generation dyadic squares of $\mathbb{I}_N$ by $\tilde{\mathcal{D}}_k^N$. Then, for $2 < q < \infty$ we have*

$$\int_{\mathbb{I}_R}\int_{\mathbb{I}_R} \frac{|\varphi(x)-\varphi(y)|^q}{|x-y|^{q+1}}\,\mathrm{d}x\,\mathrm{d}y < \infty \quad \text{for every } R > 0 \tag{2.1}$$

if and only if

$$\sum_{k=1}^{\infty} 2^{k(q-3)} \sum_{\tilde{Q}_j \in \tilde{\mathcal{D}}_k^N} \left[\operatorname{diam}\varphi(\tilde{Q}_j)\right]^q < \infty \quad \text{for every } N \in \mathbb{N}\,. \tag{2.2}$$

Proof. First we assume the condition (2.1) with $R = 2^{12}$. Now, the mapping $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2$ admits an extension $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ in $W^{1,p}(\mathbb{I}_R \times [-R,R], \mathbb{R}^3)$ which is continuous and agrees with φ on $\mathbb{R}^2 \times \{0\}$ (see (1.1) and the paragraph before). It suffices to prove (2.2) with $N = 1$.

Fix $\tilde{Q}_{k,j} \in \tilde{\mathcal{D}}_k^1$ for some $k \in \mathbb{N}$ and $j \in \{1, \dots, 2^{2k}\}$. We denote the centre of $\tilde{Q}_{k,j} \subset \mathbb{R}^2$ by $x_\circ$. Let $\mathbb{B}_R^3$ be the 3-dimensional ball in $\mathbb{R}^3$ centred at $x_\circ$ with radius $R > 0$ and

$$\mathbb{B}_R^2 = \mathbb{B}_R^3 \cap (\mathbb{R}^2 \times \{0\})\,. \tag{2.3}$$

Choose $\eta \in (2, q)$. According to the Sobolev imbedding theorem on spheres [15, Lemma 2.19] there is a constant $C > 0$ such that for a.e. $s \in (0, R)$ we have

$$\operatorname{diam} f(\partial \mathbb{B}^3_s) \leqslant C\, s^{1-\frac{2}{\eta}} \left(\int\limits_{\partial \mathbb{B}^3_s} |Df|^\eta \right)^{\frac{1}{\eta}} .$$

This is the moment where we used the assumption $q > 2$. By (2.3) we always have

$$\operatorname{diam} f(\partial \mathbb{B}^2_s) \leqslant \operatorname{diam} f(\partial \mathbb{B}^3_s) .$$

Since $\varphi \colon \mathbb{R}^2 \overset{\text{onto}}{\longrightarrow} \mathbb{R}^2$ is a homeomorphism we get

$$\operatorname{diam} \varphi(\mathbb{B}^2_s) = \operatorname{diam} \varphi(\partial \mathbb{B}^2_s) .$$

For fixed $r \in (0, R/2)$, the above estimates give

$$\operatorname{diam} \varphi(\mathbb{B}^2_r) \leqslant C\, s^{1-\frac{2}{\eta}} \left(\int\limits_{\partial \mathbb{B}^3_s} |Df|^\eta \right)^{\frac{1}{\eta}} \qquad \text{for a.e. } s \in (r, R)$$

and

$$\left[\operatorname{diam} \varphi(\mathbb{B}^2_r)\right]^\eta \int\limits_r^{2r} \frac{\mathrm{d}s}{s^{\eta-2}} \leqslant C \int\limits_{\mathbb{B}^3_{2r} \setminus \mathbb{B}^3_r} |Df|^\eta . \tag{2.4}$$

Thus

$$\operatorname{diam} \varphi(\mathbb{B}^2_r) \leqslant C r^{1-\frac{3}{\eta}} \left(\int\limits_{\mathbb{B}^3_{2r}} |Df|^\eta \right)^{\frac{1}{\eta}}$$

and

$$\operatorname{diam} \varphi(\tilde{Q}_{k,j}) \leqslant C 2^{-k(1-3/\eta)} \left(\int\limits_{\mathbb{B}^3_{2^{3-k}}} |Df|^\eta \right)^{\frac{1}{\eta}} . \tag{2.5}$$

The k-th dyadic decomposition $\tilde{\mathcal{D}}_k = \{\tilde{Q}_{k,j} \colon k \in \mathbb{N}\,,\ j = 1, \dots, 2^{2k}\}$ of $\mathbb{I}_1 \subset \mathbb{R}^2$ defines a corresponding Whitney decomposition of $\mathbb{I}_1 \times [0, 2] \subset \mathbb{R}^3$,

$$\mathcal{W}_k = \{\tilde{Q}^3_{k,j} \colon k \in \mathbb{N}\,,\ j = 1, \dots, 2^{2k}\}$$

where

$$\tilde{Q}^3_{k,j} = \tilde{Q}_{k,j} \times [2^{-k+1}, 2^{-k+2}] .$$

Let $x \in \tilde{Q}^3_{k,j}$ and $c = 2^{11}$. Then $B^3_{c2^{-k}}(x) = B^3(x, c2^{-k}) \supset \mathbb{B}^3_{2^{3-k}}$ and so

$$\operatorname{diam} \varphi(\tilde{Q}_{k,j}) \leqslant C 2^{-k(1-3/\eta)} \left(\fint_{B^3_{c2^{-k}}(x)} |Df|^\eta \right)^{\frac{1}{\eta}}$$

by (2.5). In particular, we have

$$\operatorname{diam} \varphi(\tilde{Q}_{k,j}) \leqslant C 2^{-k} \left[\mathbf{M}_c |Df|^\eta(x) \right]^{\frac{1}{\eta}} \qquad \text{for all } x \in \tilde{Q}^3_{k,j} . \tag{2.6}$$

Here $\mathbf{M}_c$ denotes the Hardy-Littlewood maximal operator,

$$\mathbf{M}_c |Df|^\eta(x) = \sup_{r<c} \frac{1}{|B^3_r(x)|} \int_{B^3_r(x)} |Df|^\eta .$$

Raising the estimate (2.6) to the power q and then integrating it over the cube $\tilde{Q}^3_{k,j}$ we have

$$2^{-3k} \left[\operatorname{diam} \varphi(\tilde{Q}_{k,j}) \right]^q \leqslant C 2^{-qk} \fint_{\tilde{Q}^3_{k,j}} \left[\mathbf{M}_c |Df|^\eta(x) \right]^{\frac{q}{\eta}} .$$

Thus,

$$\begin{aligned} \sum_{k=1}^{\infty} \sum_{j=1}^{2^{2k}} 2^{k(q-3)} \left[\operatorname{diam} \varphi(\tilde{Q}_{k,j}) \right]^q &\leqslant C \sum_{k=1}^{\infty} \sum_{j=1}^{2^{2+2k}} \fint_{Q^3_{k,j}} \left[\mathbf{M}_c |Df|^\eta(x) \right]^{\frac{q}{\eta}} \\ &= C \fint_{\mathbb{I}_1 \times [0,2]} \left[\mathbf{M}_c |Df|^\eta(x) \right]^{\frac{q}{\eta}} . \end{aligned}$$

Since $q/\eta > 1$ we can use the boundedness of the Hardy-Littlewood maximal function in $L^{\frac{q}{\eta}}$ for the function $|Df|^\eta$ to obtain

$$\sum_{k=1}^{\infty} \sum_{j=1}^{2^{2k}} 2^{k(q-3)} \left[\operatorname{diam} \varphi(\tilde{Q}_{k,j}) \right]^q \leqslant C \fint_{\mathbb{I}_c \times [-2c, 2c]} |Df|^q$$

as claimed.

Secondly we assume (2.2) for $N = 1$ and some $q \in (1, \infty)$. Our goal is show that

$$\int_{\mathbb{I}_1}\int_{\mathbb{I}_1} \frac{|\varphi(x) - \varphi(y)|^q}{|x-y|^{q+1}}\,\mathrm{d}x\,\mathrm{d}y < \infty\,.$$

We say that two dyadic squares on the same level k are *neighbours* if their boundaries have at least one intersection point. We also define the *dyadic distance* $d^*(S, S')$ of two squares $S, S' \in \tilde{\mathcal{D}}^1_k$ as the number of neighbours one has to travel through to reach S' from S, so that two dyadic neighbours themselves have a distance of 0. If $S, S' \in \tilde{\mathcal{D}}^1_k$ are such squares then we denote $S|S'$ if the dyadic distance between S and S' is either 1 or 2. We first note that

$$\int_{\mathbb{I}_1}\int_{\mathbb{I}_1} \frac{|\varphi(x) - \varphi(y)|^q}{|x-y|^{q+1}}\,\mathrm{d}x\,\mathrm{d}y \leqslant \sum_{k=1}^{\infty}\sum_{S|S'}\int_S\int_{S'} \frac{|\varphi(x) - \varphi(y)|^q}{|x-y|^{q+1}}\,\mathrm{d}x\,\mathrm{d}y \tag{2.7}$$

where the inner sum is taken over all pairs $S, S' \in \tilde{\mathcal{D}}^1_k$ for which $S|S'$ holds. This is due to the geometric fact that for every pair of points $x, y \in \mathbb{I}_1$ there are dyadic squares with $S|S'$ so that $x \in S$ and $y \in S'$.

Let now $S|S'$ with $x \in S \in \tilde{\mathcal{D}}^1_k$ and $y \in S' \in \tilde{\mathcal{D}}^1_k$. Denote by $S_1 \in \tilde{\mathcal{D}}^1_k$ and $S_2 \in \tilde{\mathcal{D}}^1_k$ two different dyadic squares so that (S, S_1, S_2, S') form a sequence of dyadic squares for which each successive pair is a neighbour. Then we simply estimate that

$$\begin{aligned}|\varphi(x) - \varphi(y)| &\leqslant \operatorname{diam}\varphi(S) + \operatorname{diam}\varphi(S_1) + \operatorname{diam}\varphi(S_2) + \operatorname{diam}\varphi(S')\\ &\leqslant \sum_{d^*(S,\tilde{Q})\leqslant 2} \operatorname{diam}\varphi(\tilde{Q}).\end{aligned}$$

Note that the sum in the last expression has at most 49 terms. Hence if we sum this expression over all dyadic squares S, every dyadic square will be repeated at most 49 times. Plugging this into (2.7) and using (2.2) gives

$$\begin{aligned}\int_{\mathbb{I}_1}\int_{\mathbb{I}_1} \frac{|\varphi(x) - \varphi(y)|^q}{|x-y|^{q+1}}\,\mathrm{d}x\,\mathrm{d}y &\leqslant \sum_{k=1}^{\infty}\sum_{S\in\tilde{\mathcal{D}}^1_k}\int_S\int_{S'} \frac{49^p\,[\operatorname{diam}\varphi(S)]^q}{2^{-(q+1)k}}\,\mathrm{d}x\,\mathrm{d}y\\ &\leqslant 49^q \sum_{k=1}^{\infty} \frac{2^{-2k}2^{-2k}}{2^{-(q+1)k}} \sum_{S\in\tilde{\mathcal{D}}^1_k} \big[\operatorname{diam}\varphi(S)\big]^q\\ &< \infty. \quad \square\end{aligned}$$

Clearly, Theorem 1.1 is an immediate consequence of Theorem 2.2.

3. Examples

An arbitrary homeomorphism $\varphi\colon \partial\mathbb{D} \overset{\text{onto}}{\longrightarrow} \partial\mathbb{D}$ admits a homeomorphic extension to the unit disk $\mathbb{D} \subset \mathbb{R}^2$ in the Sobolev class $W^{1,q}(\mathbb{D}, \mathbb{R}^2)$ for all $q < 2$. Our next example shows that such a result has no 3D counterpart.

Example 3.1. There is a Sobolev homeomorphism $\varphi\colon \mathbb{S} \overset{\text{onto}}{\longrightarrow} \mathbb{S}$ such that $\varphi \notin W^{1-\frac{1}{q},q}(\mathbb{S}, \mathbb{R}^3)$ for any $q > 1$ and hence it does not admit a continuous extension $f\colon \overline{\mathbb{B}} \to \mathbb{R}^3$ in $W^{1,q}(\mathbb{B}, \mathbb{R}^3)$.

Proof. We simplify our writing here and construct a Sobolev homeomorphism $\varphi\colon [0,1] \times [0,1] \overset{\text{onto}}{\longrightarrow} [0,1] \times [0,2]$ with $\varphi(0,0) = \varphi(1,1)$. Note that this causes no loss of generality due to a suitable bilipschitz change of variables in both domain and target side, and the fact that the 2D sphere may be appropriately covered by such atlases.

Let $s\colon \mathbb{R} \to \mathbb{R}$ be a 1-periodic piecewise linear "saw" function defined by

$$s(x) = \begin{cases} 2x & \text{for } x \in [0, \frac{1}{2}], \\ 2 - 2x & \text{for } x \in [\frac{1}{2}, 1]. \end{cases}$$

We set $s_k(x) = s(x10^k)$ and obtain a 10^{-k}-periodic saw function. By induction we choose an increasing sequence of integers n_k such that

$$\begin{gathered} 10^{-kq} 10^{(q-1)\frac{1}{2}n_k} \geqslant 2^k \text{ and} \\ \Big(\sum_{j=1}^{k-1} 10^{-j} \cdot 2 \cdot 10^{n_j}\Big) 10^{-\frac{1}{2}n_k} \leqslant \frac{1}{8} 10^{-k}. \end{gathered} \tag{3.1}$$

We set

$$r_k = 10^{-\frac{1}{2}n_k} \text{ and } \phi(x) = \sum_{j=1}^{\infty} 10^{-j} s_{n_j}(x).$$

Note that ϕ, being a uniform limit of continuous functions, is also continuous. It is not difficult to check that the mapping $\varphi\colon [0,1]^2 \overset{\text{onto}}{\longrightarrow} [0,1] \times [0,2]$, defined by

$$\varphi(x_1, x_2) = [x_1, x_2 + \phi(x_1)] \text{ is a homeomorphism.}$$

We estimate

$$\begin{aligned} &\int\limits_{(0,1)^2 \times (0,1)^2} \frac{|\varphi(x) - \varphi(y)|^q}{|x - y|^{q+1}} \, dx \, dy \\ &\geqslant C \int\limits_{(0,1)^2 \times (0,1)^2} \frac{(|\phi(x_1) - \phi(y_1)| - |x_2 - y_2|)^q}{|x - y|^{q+1}} \, dx \, dy \end{aligned} \tag{3.2}$$

and note that the term $\frac{|x_2-y_2|^q}{|x-y|^{q+1}} \leqslant \frac{1}{|x-y|}$ in the last integral is integrable. Therefore, it suffices to show that the integral

$$\int\limits_{(0,1)^2\times(0,1)^2} \frac{|\phi(x_1)-\phi(y_1)|^q}{|x-y|^{q+1}}\,dx\,dy \tag{3.3}$$

diverges.

For that, let us fix $k \in \mathbb{N}$ and denote

$$A_1 := \big\{x_1 \in [0,1] :\ x_1 \in [-\tfrac{1}{8}10^{-n_k} + j10^{-n_k}, \tfrac{1}{8}10^{-n_k} + j10^{-n_k}] \text{ for } j \in \mathbb{N}\cup\{0\}\big\},$$

i.e. $s_{n_k}(x_1) \in [0, \frac{1}{4}]$ for every $x_1 \in A_1$ and

$$A_2 = \{y_1 \in [0,1] :\ y_1 \in [\tfrac{3}{8}10^{-n_k} + j10^{-n_k}, \tfrac{5}{8}10^{-n_k} + j10^{-n_k}] \text{ for } j \in \mathbb{N}\cup\{0\}\},$$

i.e. $s_{n_k}(y_1) \in [\frac{3}{4}, 1]$ for every $y_1 \in A_2$. Given $x_1 \in A_1$ we set

$$A_2(x_1) = A_2 \cap (x_1 - r_k, x_1 + r_k).$$

It is easy to see that for every $x_1 \in A_1$ and $y_1 \in A_2$ we have

$$10^{-k}|s_{n_k}(x_1) - s_{n_k}(y_1)| \geqslant \frac{1}{2}10^{-k}.$$

Further for every x_1 and y_1 we have

$$\Big|\sum_{j=k+1}^{\infty} 10^{-j}s_{n_j}(x_1) - \sum_{j=k+1}^{\infty} 10^{-j}s_{n_j}(y_1)\Big| \leqslant \sum_{j=k+1}^{\infty} 10^{-j} \leqslant \frac{1}{8}10^{-k}.$$

The function $10^{-j}s_{n_j}$ is Lipschitz with Lipschitz constant $10^{-j}\frac{1}{10^{-n_j}/2}$. Hence in view of (3.1), for every x_1 and y_1 with $|x_1 - y_1| < r_k$ we have

$$\Big|\sum_{j=1}^{k-1} 10^{-j}s_{n_j}(x_1) - \sum_{j=1}^{k-1} 10^{-j}s_{n_j}(y_1)\Big| \leqslant \sum_{j=1}^{k-1} 10^{-j}\cdot 2\cdot 10^{n_j}\cdot|x_1 - y_1| \leqslant \frac{1}{8}10^{-k}.$$

It follows that for every $x_1 \in A_1$ and $y_1 \in A_2$ with $|x_1 - y_1| < r_k$ we have

$$\begin{aligned}|\phi(x_1)-\phi(y_1)| \geqslant & 10^{-k}|s_{n_k}(x_1)-s_{n_k}(y_1)| \\ & - \Big|\sum_{j=k+1}^{\infty} 10^{-j} s_{n_j}(x_1) - \sum_{j=k+1}^{\infty} 10^{-j} s_{n_j}(y_1)\Big| \\ & - \Big|\sum_{j=1}^{k-1} 10^{-j} s_{n_j}(x_1) - \sum_{j=1}^{k-1} 10^{-j} s_{n_j}(y_1)\Big| \\ \geqslant & \frac{1}{4}10^{-k}.\end{aligned}$$

To show (3.3) we estimate the integral

$$C \int_{A_1} \int_{A_2(x_1)} \int_0^1 \int_0^1 \frac{10^{-kq}}{\big(|x_1-y_1|+|x_2-y_2|\big)^{q+1}}\, dx_2\, dy_2\, dy_1\, dx_1.$$

Since applying a change of variables $s = x_2 - y_2$ and $t = x_2 + y_2$ we obtain

$$\begin{aligned}\int_0^1 \int_0^1 \frac{1}{\big(|a|+|x_2-y_2|\big)^{q+1}}\, dx_2\, dy_2 &\geqslant C \int_{\frac{1}{2}}^{\frac{3}{2}} 1\, dt \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{\big(|a|+|s|\big)^{q+1}}\, ds \\ &\geqslant C\frac{1}{|a|^q}\end{aligned}$$

we may estimate (3.3) from below by the integral

$$C \int_{A_1} \int_{A_2(x_1)} \frac{10^{-kq}}{|x_1-y_1|^q}\, dy_1\, dx_1. \tag{3.4}$$

We use again a change of variables $s = x_1 - y_1$ and $t = x_1 + y_1$. Since $|A_1| \geqslant \frac{1}{4}$ and $|A_2| \geqslant \frac{1}{4}$ it is not difficult to see that the sets $A_1 + A_2$ and $A_1 - A_2$ are large enough, i.e. they occupy a large percentage of each interval of size much bigger than 10^{-n_k}. Together with the fact that $r_k = 10^{-\frac{1}{2}n_k}$ is much bigger than the period of s_{n_k} which is 10^{-n_k} we may estimate the integral (3.4) from below as

$$C \int_{r_k/2}^{r_k} \frac{10^{-kq}}{|s|^q}\, ds \geqslant C\frac{10^{-kq}}{r_k^{q-1}}.$$

By (3.1) we finally conclude that the integral (3.3) diverges as we wanted. □

The following example shows the sharpness of Theorem 1.3.

Example 3.2. Let $p \geqslant 2$ and $q > \frac{3}{2}p$. There is a Sobolev homeomorphism $\varphi\colon \mathbb{S} \xrightarrow{\text{onto}} \mathbb{S}$ such that $\varphi \in W^{1,p}(\mathbb{S},\mathbb{R}^3)$ but $\varphi \notin W^{1-\frac{1}{q},q}(\mathbb{S},\mathbb{R}^3)$. Hence such a φ does not admit a continuous extension $h\colon \overline{\mathbb{B}} \to \mathbb{R}^3$ in the Sobolev class $W^{1,q}(\mathbb{B},\mathbb{R}^3)$.

Proof. For simplicity we give a formula for φ from $\mathbb{D}$ onto itself and not from $\mathbb{S}$ onto $\mathbb{S}$. It is clear that this causes no loss of generality due to a suitable bilipschitz change of variables. Given our $p \geqslant 2$ and $q > \frac{3}{2}p$ we choose $\alpha > 0$ such that

$$1 - \frac{2}{p} < \alpha < 1 - \frac{3}{q}.$$

We set

$$\varphi(x) = \frac{x}{|x|}|x|^{\alpha}.$$

A simple computation gives that $\varphi \in W^{1,p}(\mathbb{D},\mathbb{R}^2)$. Either by a direct computation we also obtain that $\varphi \notin W^{1-\frac{1}{q},q}(\mathbb{D},\mathbb{R}^2)$ (see e.g. [27, Lemma 1, page 44]) or assuming by contradiction that $\varphi \in W^{1-\frac{1}{q},q}(\mathbb{D},\mathbb{R}^2)$. In the latter case φ admits a continuous extension $h\colon \mathbb{D}\times(-1,1) \to \mathbb{R}^3$ in the Sobolev class $W^{1,q}(\mathbb{D}\times(-1,1),\mathbb{R}^3)$. In particular, h is locally $(1-\frac{3}{q})$-Hölder continuous but this is impossible because $h = \varphi$ on $\mathbb{D}\times\{0\}$ is just $(1-\frac{2}{\alpha})$-Hölder continuous. □

Theorem 1.3 follows from Theorem 1.2. In the following example we show that on the contrary there is a homeomorphism $\varphi\colon \mathbb{S} \xrightarrow{\text{onto}} \mathbb{S}$ which satisfy the condition (1.3) in Theorem 1.3 and does not belong to any Sobolev class $W^{1,p}(\mathbb{S},\mathbb{R}^3)$, $p \geqslant 1$. Again, we define φ only on $[0,1]^2$, and a bilipschitz change of variables easily generalizes this homeomorphism from $\mathbb{S}$ onto $\mathbb{S}$.

Example 3.3. Consider

$$\varphi(x,y) = [g(x),y] \text{ where } g(x) = x + C(x) \tag{3.5}$$

and C is Cantor function. Not the standard 1/3 Cantor function, but $1/K$ Cantor function (for $K \geqslant 2$), i.e. in each step we remove the middle $1/K$-part of the interval. It is not difficult to show that this Cantor function is Hölder continuous with exponent $\alpha = \frac{\log\frac{1}{2}}{\log(\frac{1}{2}(1-\frac{1}{K}))}$. Let us note that

$$\lim_{k\to\infty} \alpha = \lim_{K\to\infty} \frac{\log\frac{1}{2}}{\log(\frac{1}{2}(1-\frac{1}{K}))} = 1.$$

Let $\tilde{\mathcal{D}}_k$, $k \in \mathbb{N}$, be the collection of k-th generation dyadic square of $[0,1]^2$ into $(2^k)^2$ squares of sidelength 2^{-k}. It is easy to see that $\mathscr{H}^1(\varphi(\partial\tilde{Q}_{k,j})) < \infty$ for all k and j by (3.5). Using Hölder continuity of h we get

$$\sum_{k=0}^{\infty}\sum_{j=1}^{2^{2k}} 2^{-(3-q)k}\mathscr{H}^1(\varphi(\partial\tilde{Q}_{k,j}))^q \leqslant C\sum_{k=0}^{\infty} 2^{2k}2^{-(3-q)k}[2^{-\alpha k}]^q.$$

This sum is finite whenever $q(1-\alpha)<1$, which we can guarantee by choosing K large enough at the start, in which case also (1.3) holds. By Theorem 1.2 we obtain that we can extend this boundary homeomorphism as a $W^{1,q}$ homeomorphism inside. However, the mapping φ does not belong to $W^{1,1}_{\mathrm{loc}}([0,1]^2,\mathbb{R}^2)$ as it fails the ACL condition on all vertical segments (it just has bounded variation).

4. Structure of the proof of Theorem 1.2

In this section we give a brief overview of the arguments we need to prove our main extension result, Theorem 1.2.

Before we address the case of extending a boundary map φ from the unit sphere to itself, we aim to first describe an extension method which extends a homeomorphic boundary map $\varphi:\mathbb{R}^2\to\mathbb{R}^2$ as a homeomorphism of the upper half space to itself. This will comprise the majority of the proof (Sections 5 to 8), while the topological arguments used to extend this method to the spherical case will be explained in Section 9.

Recalling that $S_0=[0,1]^2$ is the unit square in the plane, our aim is to define a continuous injective extension $h:[0,1]^3\to\mathbb{R}^3_+$ which agrees with a given homeomorphism φ on $[0,1]^2\times\{0\}$ (this is identified with S_0). The construction of h is split into two parts: First we construct a monotone extension of φ in Sections 5 to 7 and then describe how this monotone extension may be modified to be injective in Section 8. Here monotonicity is in the sense of Morrey, meaning that the preimage of each point is connected.

The basic idea is to decompose the domain space $[0,1]^3$ dyadically into cubes $U_{k,j}$. Recall the original standard dyadic decomposition of S_0 into dyadic squares $\tilde{Q}_{k,j}$. We define $U_{k,j}=\tilde{Q}_{k,j}\times[2^{-k},2^{-(k-1)}]$. Thus $U_{k,j}$ is a cube of side length 2^{-k} and the union of all such cubes decompose the domain space $[0,1]^3$. The idea is to map each cube to a 'cylindrical' region $V_{k,j}$ in the target.

To define the region $V_{k,j}$, we consider the dyadic squares $\tilde{Q}_{k,j}\subset\mathbb{R}^2$ on the domain side. For each such square, we will define a curve $\Gamma_{k,j}$ on the target side as a piecewise linear approximation of the image curve $\varphi(\partial\tilde{Q}_{k,j})$. Section 5 will explain the precise details, but in particular we get that the curves $\Gamma_{k,j}$ form a tiling of the plane on each dyadic level k, and satisfy the total estimate

$$\sum_{k=1}^{\infty}\sum_{m=1}^{4^k} 2^{k(q-3)}\mathcal{H}^1(\Gamma_{k,j})^q<\infty. \tag{4.1}$$

The top face of $V_{k,j}$ will be the horizontal region bounded by the curve $\Gamma_{k,j}\times\{2^{-(k-1)}\}$, and the bottom face will consist of the union of the regions bounded by $\hat{\Gamma}^{(m)}_{k,j}\times\{2^{-k}\}$, where $\hat{\Gamma}^{(m)}_{k,j}$ for $m=1,\dots,4$ denote the four dyadic children of $\Gamma_{k,j}$. See Fig. 1.

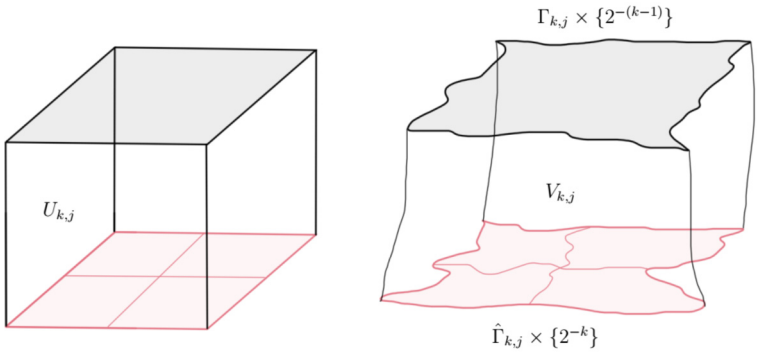


Fig. 1. The cube $U_{k,j}$ and its image set $V_{k,j}$ defined as a region spanned by the curve $\Gamma_{k,j} \times \{2^{-(k-1)}\}$ and its corresponding curve $\hat{\Gamma}_{k,j} \times \{2^{-k}\}$ on the next level.

We aim to define the extension h so that it maps each horizontal section of $U_{k,j}$ to the horizontal section of $V_{k,j}$ of the same height. The horizontal sections of $V_{k,j}$ will still need to be defined, however, and to do this we will need to construct an appropriate homotopy between the curve $\Gamma_{k,j}$ to the curve $\hat{\Gamma}_{k,j}$ which we define as the outer boundary of $\bigcup_{m=1}^{4} \hat{\Gamma}_{k,j}^{(m)}$, i.e. the curve corresponding to $\Gamma_{k,j}$ on the next dyadic level. In terms of estimating the Sobolev norm of h, our main goal is to show the following.

Goal: The map $h : U_{k,j} \to V_{k,j}$ will be a Lipschitz mapping. The Lipschitz constant of the map should be estimated from above by a uniform constant times the quantity $(\mathcal{H}^1(\Gamma_{k,j}) + \sum_{m=1}^{4} \mathcal{H}^1(\hat{\Gamma}_{k,j}^{(m)}))2^k$, or possibly this quantity added together with the same quantity over all of the neighbours of $\Gamma_{k,j}$.

After Sections 5 to 7 we will have defined the monotone extension h on each dyadic cube $U_{k,j}$ so that the goal estimate above holds, and this extension is further modified into an injective extension h in Section 8 with the same estimates still holding. The $W^{1,q}$-norm of h can then be estimated by estimating the differential $|Dh|$ above by the Lipschitz constant of h. Combined with the goal estimate this gives

$$\int_{U_{k,j}} |Dh(z)|^q \, dz \leqslant 2^{k(q-3)} \left(\mathcal{H}^1 \left(\Gamma_{k,j} \right) + \sum_{m=1}^{4} \mathcal{H}^1 \left(\hat{\Gamma}_{k,j}^{(m)} \right) \right)^q .$$

Combined with (4.1) this will yield that h belongs to the Sobolev space $W^{1,q}([0,1]^3)$ as desired. The proof of Theorem 1.2 is then finished in Section 9 where we explain the slight changes in the arguments needed for the spherical case.

5. Decomposition of the domain and target side

In this section we start with the standard dyadic decomposition $\tilde{\mathcal{D}}_k$ of the boundary and define a modification of it in order to control the lengths of the image curves of the image grid under the given boundary map φ. Furthermore, we will define piecewise

linear replacements of these image curves. These divisions on the domain and target side will be used in later sections to assist in defining the extension map we use to prove our main result, Theorem 1.2. We also show in this section that Theorem 1.3 follows from Theorem 1.2.

Lemma 5.1. *Let $\tilde{\mathcal{D}}_k = \{\tilde{Q}_{k,j} : k \in \mathbb{N}, j = 1 \dots 2^{2k}\}$ be the dyadic decomposition of the unit square $Q_0 = [0,1]^2$ into closed squares of side length 2^{-k} for each fixed k. Let $p > 1$ and $\varphi : \overline{Q_0} \to \overline{Q_0}$ be a homeomorphism in the space $\varphi \in W^{1,p}(3Q_0, \mathbb{R}^2)$. Then there exists a set of closed quadrilaterals $\mathcal{D}_k = \{Q_{k,j} : k \in \mathbb{N}, j = 1 \dots 2^{2k}\}$ such that*

(1) For each point $\tilde{v} \in Q_0$ which is a vertex of a dyadic square of side length 2^{-k} in $\tilde{\mathcal{D}}_k$, there exists exactly one corresponding point $v \in Q_0$ which is a vertex of a quadrilateral from $\mathcal{D}_k$. The vertices v of a quadrilateral $Q_{k,j}$ in $\mathcal{D}_k$ are exactly the points which correspond to the vertices $\tilde{v}$ of the dyadic square $\tilde{Q}_{k,j}$. Moreover, for the coordinates of these points $v = [v_1, v_2]$ and $\tilde{v} = [\tilde{v}_1, \tilde{v}_2]$ we have (see Fig. 2)

$$v_1 - \tilde{v}_1 \in \left[\frac{2^{-k}}{10} - \frac{2^{-k}}{40}, \frac{2^{-k}}{10}\right] \text{ and } v_2 - \tilde{v}_2 \in \left[\frac{2^{-k}}{10} - \frac{2^{-k}}{40}, \frac{2^{-k}}{10}\right] \tag{5.1}$$

for all pairs of corresponding vertices.

(2) The quadrilaterals $Q_{k,j}$ for each fixed level k are thus mutually disjoint apart from their boundaries.

(3) If we inherit the parent-child relation between dyadic squares from $\tilde{\mathcal{D}}$ to $\mathcal{D}$, then the following holds. The children $Q_1, \dots Q_4 \in \mathcal{D}_{k+1}$ of a given square $Q \in \mathcal{D}_k$ (i.e. $Q = Q_1 \cup Q_2 \cup Q_3 \cup Q_4$) need not be contained in Q nor does their union need to cover Q. However, for $\hat{Q} = \cup_{i=1}^4 Q_i$ the boundaries ∂Q and $\partial \hat{Q}$ always intersect exactly at two points.

(4) For each k, j we have the inequality

$$2^{-k} \int_{\partial Q_{k,j}} |D\varphi(t)|^p dt \leqslant C \int_{2Q_{k,j}} |D\varphi(z)|^p dz. \tag{5.2}$$

Proof. (1) and (4): Let us first explain that it is possible to choose the grid so that (1) is satisfied and we have the key inequality (5.2).

This follows essentially from [18, Section 4.2] and therefore we only explain how to apply this approach here: All of our cubes in the $r = 2^{-k}$ grid are of type A since we can freely move points outside of Q_0. We would like to apply analogy of [18, Lemma 4.9] for $M = 0$ and $\varepsilon = \frac{1}{10}$. The only difference is that in [18, Lemma 4.9] they choose

$$[v_1, v_2] \in I_\varepsilon = \left\{[\tilde{v}_1 + t, \tilde{v}_2 + t] : |t| \leqslant \varepsilon 2^{-k}\right\}$$

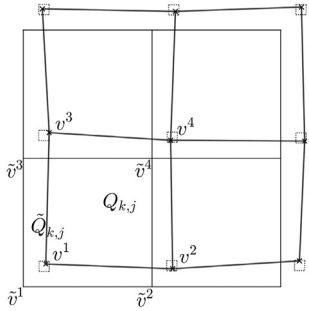


Fig. 2. Given a dyadic cube $\tilde{Q}_{k,j}$ with vertices $\tilde{v}^1, \tilde{v}^2, \tilde{v}^3, \tilde{v}^4$ we construct a quadrilateral $Q_{k,j}$ with vertices v^1, v^2, v^3, v^4. Each v^i is close to $\tilde{v}^i$, it is slightly shifted to the top and to the right from $\tilde{v}^i$.

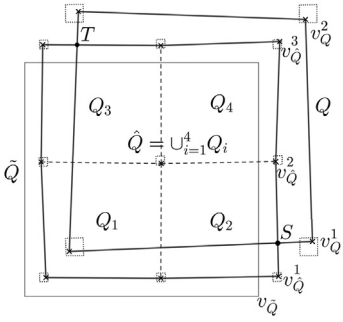


Fig. 3. Boundaries of Q and $\hat{Q} = \cup_{i=1}^4 Q_i$ intersect at two points S and T. Note here that $Q_1, \dots, Q_4$ refer to quadrilaterals which form the set $\hat{Q}$ which is the (almost square) octagon in the middle.

but we would like to make this choice in the subset of I_ε (of length 1/8 times the original length)

$$[v_1, v_2] \in I = \left\{[\tilde{v}_1 + t, \tilde{v}_2 + t] : t \in [\tfrac{1}{10}2^{-k} - \tfrac{1}{40}2^{-k}, \tfrac{1}{10}2^{-k}]\right\}.$$

This does not change anything substantial in the proof there, it only affects some multiplicative constants - use $8^2\frac{25}{\varepsilon r}$ instead of $\frac{25}{\varepsilon r}$ in the definition of $\Gamma(A, B, M)$ and then the proof carries through with obvious minor modifications. Then we can finish this step by applying analogy of [18, Lemma 4.13 and Lemma 4.16] (again with slightly increased multiplicative constant) to get our (5.2).

(2): This is easy to see from the definition of vertices of $\mathbb{Q}_{k,j}$ in step (1) (see Fig. 2).

(3): Let Q and $\hat{Q} = \bigcup_{i=1}^4 Q_i$ be as in the statement part (3) (see Fig. 3).

Let us define notation for certain vertices here, consult Fig. 3 for specific positions. Here $v_{\tilde{Q}}$ is a vertex of $\tilde{Q}$, v^1_Q and v^2_Q are vertices of Q and $v^1_{\hat{Q}}$, $v^2_{\hat{Q}}$, $v^3_{\hat{Q}}$ are vertices of $\hat{Q}$ (in fact the corresponding part of $\hat{Q}$ is given by two segments $v^1_{\hat{Q}}v^2_{\hat{Q}}$ and $v^2_{\hat{Q}}v^3_{\hat{Q}}$). From (5.1) we obtain for the x-coordinates of these points that

$$(v^1_Q)_1 - (v_{\tilde{Q}})_1, (v^2_Q)_1 - (v_{\tilde{Q}})_1 \in \left[\frac{2^{-k}}{10} - \frac{2^{-k}}{40}, \frac{2^{-k}}{10}\right]$$

and similarly from (5.1) for the choice of $\mathcal{D}_{k+1}$

$$(v^1_{\hat{Q}})_1 - (v_{\tilde{Q}})_1, (v^2_{\hat{Q}})_1 - (v_{\tilde{Q}})_1, (v^3_{\hat{Q}})_1 - (v_{\tilde{Q}})_1 \in \Big[\frac{2^{-(k+1)}}{10} - \frac{2^{-(k+1)}}{40}, \frac{2^{-(k+1)}}{10}\Big].$$

It follows that the distance of this side of Q (=segment $v^1_Q v^2_Q$) and this side of $\hat{Q}$ (=union of segments $v^1_{\hat{Q}} v^2_{\hat{Q}}$ and $v^2_{\hat{Q}} v^3_{\hat{Q}}$) is at least $\frac{2^{-k}}{10} - \frac{2^{-k}}{40} - \frac{2^{-(k+1)}}{10} = \frac{2^{-k}}{40}$ and thus these two sides do not intersect. By a similar reasoning on other sides we obtain that ∂Q and $\partial \hat{Q}$ intersect at exactly two points S and T as in Fig. 3.

Let us also note that the distance of S and v^1_Q (and similarly distance of S and $v_{\hat{Q}_1}$) is at least $\frac{2^{-k}}{40}$ and thus these intersection points are not too close to the vertices of ∂Q and $\partial \hat{Q}$. □

Definition 5.2. Note that conditions (1)-(3) above do not involve the boundary map φ. Hence we may define that any set $\mathcal{D}_k$ of quadrilaterals $Q_{k,j}$ satisfying the conditions (1)-(3) is called a *good modification* of the standard dyadic decomposition of Q_0.

Proof of Theorem 1.3. Note that the statement is obvious if $p \geqslant q$ as we can use the trivial radial extension. In the following we thus assume that $p < q$.

Given a homeomorphism $\varphi \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^2, \mathbb{R}^2)$ we were able to find in Lemma 5.1 a good modification $\mathcal{D}_k$ of the dyadic grid so that (5.2) holds. We could start with a homeomorphism $\varphi \in W^{1,p}(\mathbb{S}, \mathbb{S})$ and some analogy of dyadic grid on $\mathbb{S}$. Analogously to the proof of Lemma 5.1 we can find a good modification $\mathcal{D}_k$ of this grid on $\mathbb{S}$ so that an analogy of (5.2) holds for φ. In fact the whole statement can be also obtained locally using a bilipschitz change of variables. Given k, our dyadic grid $\mathcal{D}_k$ contains bi-Lipschitz squares of diameter $\approx 2^{-k}$ and of perimeter $\mathcal{H}^1(\partial Q_{k,j}) \approx 2^{-k}$. Moreover, there are approximately 2^{2k} such squares, let us denote by n_k here the total amount of bi-Lipschitz squares in $\mathcal{D}_k$.

In view of Theorem 1.2 it is now enough to show finiteness of (1.3). Using Hölder's inequality, (5.2), $q/p \geqslant 1$ and $p > \frac{2}{3}q$ we obtain

$$\begin{aligned}
\sum_{k=1}^{\infty}\sum_{j=1}^{n_k} 2^{-(3-q)k}\mathscr{H}^1(\varphi(\partial Q_{k,j}))^q &\leqslant \sum_{k=1}^{\infty}\sum_{j=1}^{n_k} 2^{-(3-q)k}\Big(\int\limits_{\partial Q_{k,j}} |D\varphi|\Big)^q \\
&\leqslant \sum_{k=1}^{\infty}\sum_{j=1}^{n_k} 2^{-(3-q)k}\Big(\Big(\int\limits_{\partial Q_{k,j}} |D\varphi|^p\Big)^{\frac{1}{p}}(2^{-k})^{1-\frac{1}{p}}\Big)^q \\
&\leqslant C\sum_{k=1}^{\infty} 2^{-(3-q)k}2^{-k(q-\frac{q}{p})}\sum_{j=1}^{n_k}\Big(\Big(2^k\int\limits_{2Q_{k,j}} |D\varphi|^p\Big)^{\frac{1}{p}}\Big)^q \\
&\leqslant C\sum_{k=1}^{\infty} 2^{-k(3-\frac{q}{p})}2^{k\frac{q}{p}}\sum_{j=1}^{n_k}\int\limits_{2Q_{k,j}} |D\varphi|^p
\end{aligned}$$

$$\leqslant C \sum_{k=1}^{\infty} 2^{-k(3-2\frac{q}{p})} < \infty. \quad \square$$

The aim of the next lemma is to consider the modified dyadic grid given by Lemma 5.1. For each level k, we then look at the image of the grid of level k under φ (specifically the set $\varphi(\cup_j \partial Q_{k,j})$). The aim is to modify this "image grid" so that instead of general Jordan curves it consists of curves which are piecewise linear. It is necessary to preserve both the topology of the image grid and the lengths of the image curves. This piecewise linear approximation will simplify future computations.

Lemma 5.3. *Let $p \geqslant 1$ and $\varphi : \overline{Q_0} \to \overline{Q_0}$ be a homeomorphism in the space $\varphi \in W^{1,p}(Q_0, \mathbb{R}^2)$. Let $\mathcal{D}_k$ be the set of modified dyadic quadrilaterals given by Lemma 5.1. In particular, the Jordan curves $\varphi(\partial Q_{k,j})$ for each $Q_{k,j} \in \mathcal{D}_k$ each have finite length. Then for each quadrilateral $Q_{k,j}$ there exists a corresponding closed Jordan curve $\Gamma_{k,j} \subset \overline{Q_0}$ on the image side such that.*

(1) Each of the curves $\Gamma_{k,j}$ is piecewise linear.
(2) Each point on the curve $\Gamma_{k,j}$ is of distance at most 2^{-k} from the set $\varphi(\partial Q_{k,j})$.
(3) The inequality $\mathscr{H}^1(\Gamma_{k,j}) \leqslant \mathscr{H}^1(\varphi(\partial Q_{k,j}))$ holds.
(4) $\Gamma_{k,j}$ passes through the four points $\varphi(v)$, where v ranges over the four vertices of the quadrilateral $Q_{k,j}$. These four points are called the vertices of $\Gamma_{k,j}$.
(5) If two quadrilaterals $Q_{k,j}, Q_{k,j'} \in \mathcal{D}_k$ share a common side with endpoints v_1, v_2, then the subarcs of their corresponding image curves $\Gamma_{k,j}, \Gamma_{k,j'}$ with endpoints at the common vertices $\varphi(v_1)$ and $\varphi(v_2)$ are the same.
(6) Apart from the cases where two curves $\Gamma_{k,j}, \Gamma_{k,j'}$ at the same level k share either a single vertex or a single subarc between two vertices as before, these Jordan curves are mutually disjoint (for each fixed level k).
(7) For every $Q_{k,j} \in \mathcal{D}_k$ and $Q_{k+1,j'} \in \mathcal{D}_k$ (see Fig. 3) we know that

$$\Gamma_{k,j} \cap \Gamma_{k+1,j'} = \varphi(\partial Q_{k,j}) \cap \varphi(\partial Q_{k+1,j'}).$$

That is each $\Gamma_{k,j}$ passes not only through its vertices but also through its intersection with grids of step $k+1$ and $k-1$, i.e. images of boundaries of $\mathcal{D}_{k+1}$ and $\mathcal{D}_{k-1}$.

Proof. In this proof we use ideas of [12] and [18] (see also [20] and [25]) where a similar piecewise linear approximation of curves was used. The idea is to do this approximation in three steps: First we linearize around vertices of the image grid, secondly linearize between intersection points of levels k and $k+1$ (to ensure that (7) is satisfied), and lastly to linearize the remaining non-intersecting curves.

Step 1. Linearization near vertices: Fix k for a moment, and denote by

$$\mathcal{V}_k = \{\varphi(v) : \ v \text{ is a vertex of some } Q_{k,j}\}$$

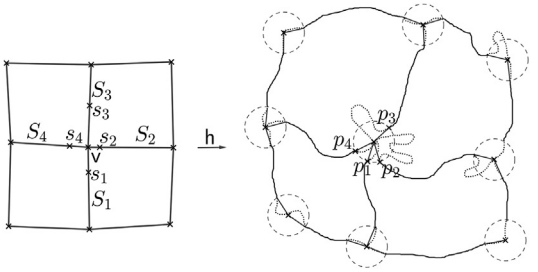


Fig. 4. We replace original curve near vertices (see dotted curves) by segments near vertices.

the set of images of vertices of $\mathcal{D}_k$. Let us also define the image grids

$$\mathcal{G}_0 = \emptyset \text{ and } \mathcal{G}_k = \bigcup_j \varphi(\partial Q_{k,j}).$$

Let $\mathcal{W}_k = \mathcal{G}_k \cap \mathcal{G}_{k+1}$ denote the set of intersection points between image grids of successive levels. Analogously to the reasoning in the proof of Lemma 5.1 (3), we see that both $\mathcal{V}_k$ and $\mathcal{W}_k$ are finite.

We now choose a collection of small balls $\mathcal{B}_k$ with centres at each point in $\mathcal{V}_k$. More precisely, for each vertex v of some $Q_{k,j}$ we choose $r > 0$ small enough so that the balls $B(\varphi(v), 2r)$ are pairwise disjoint and that these balls do not contain any of the points in $\mathcal{W}_k$ or $\mathcal{V}_{k+1}$. Due to the latter property we may also assume that the balls in $\mathcal{B}_k$ and the balls in $\mathcal{B}_{k+1}$ do not intersect either, as for each k we may first choose the balls in $\mathcal{B}_k$ and then later choose the balls in $\mathcal{B}_{k+1}$ small enough to not intersect the previous set of balls.

Furthermore, we may use the uniform continuity of φ^{-1} and φ to assume that

$$|\varphi(x) - \varphi(v)| < 2^{-k}, \qquad \forall x \in B\big(v, \operatorname{diam}(\varphi^{-1}(B(\varphi(v), r))\big). \tag{5.3}$$

For each vertex v of the grid $\mathcal{D}_k$ we have four sides S_1, S_2, S_3 and S_4 of some $Q_{k,j}$ that have v as their endpoint (see Fig. 4). On each of these sides we choose points $s_i \in S_i$ so that $p_i = \varphi(s_i) \in \partial B(\varphi(v), r)$ and so that s_i is furthest away from v with this property (e.g. on S_3 in Fig. 4 we have three points whose image intersects $\partial B(\varphi(v), r)$). Now we replace φ on each segment $[s_i, v]$ by a segment $[p_i, \varphi(v)]$ and we leave φ the same outside of these four segments (see Fig. 4). In this way we replace $\varphi(\partial Q_{k,j})$ by a curve $\Gamma^{(*)}_{k,j}$ which is piecewise linear close to the vertices.

It is easy to see that this new curve $\Gamma^{(*)}_{k,j}$ satisfies an analogy of (2) by (5.3) and it is not difficult to see that these new curves are one-to-one (see Fig. 4), i.e. they intersect only at original vertices v. These new curves have also length shorter or equal to the original $\mathscr{H}^1(\varphi(\partial Q_{k,j}))$.

We proceed to do the linearization process of this step on each level $k = 1, 2, 3, \ldots$, replacing the collection of all curves $\varphi(\partial Q_{k,j})$ by a new set of curves $\Gamma^{(*)}_{k,j}$. To reiterate,

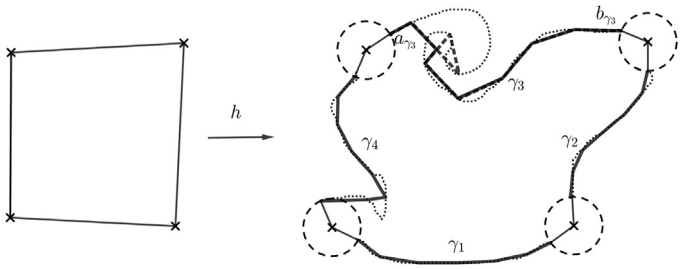


Fig. 5. We replace curves γ_m on the sides (see dotted curves) by piecewise linear curves. We may need to choose a one-to-one shortening of these replacements, i.e. we ignore some dashed part of the replacement of γ_3.

on each level k these curves are now linear around the points $\mathcal{V}_k$, but were unchanged near the set of intersection points $\mathcal{W}_k$. The properties $(2)-(7)$ are preserved in this process and we may continue the linearization to achieve (1) later.

Step 2. Linearization at the intersection points $\mathcal{W}_k$: In the previous step, we avoided making any changes near the set of intersection points $\mathcal{W}_k$ between curves of level k and $k+1$. In this step we will, for each level k, linearize the curves $\Gamma_{k,j}^{(*)}$ around the points $\mathcal{W}_k$.

This process can be done quite analogously to Step 1. We choose a new set of balls $\mathcal{B}'_k$ which are centred around points in $\mathcal{W}_k$, and may again assume that the balls within each collection and between each successive collection ($\mathcal{B}'_k$ and $\mathcal{B}'_{k+1}$) are disjoint.

We then apply the same linearization process of Step 1 in each of these balls, linearizing each of the four parts (two from level k and two from $k+1$) which meet at the centre of each ball. This replaces the curves $\Gamma_{k,j}^{(*)}$ by another set of curves $\Gamma_{k,j}^{(**)}$ which are now also piecewise linear near the points in $\mathcal{W}_k$. In this modification the properties $(2)-(7)$ are again preserved for the whole collection of curves.

Step 3. Linearization of sides: Now we need to linearize the curves $\Gamma_{k,j}^{(**)}$ in the remaining parts which consist of simple Jordan curves between the balls in $\mathcal{B}_k$ and $\mathcal{B}'_k$. We call $\gamma_{k,m}$ the parts of $\Gamma_{k,j}^{(**)}$ where our curve is not piecewise linear yet, these correspond to image by φ of segments of $Q_{k,j}$ (minus some small segments near vertices of $\mathcal{D}_k$ and intersection points of $\mathcal{D}_k$ and $\mathcal{D}_{k+1}$).

These $\gamma_{k,m}$ are pairwise disjoint and we can choose $0 < \delta < 2^{-k}$ so that $\gamma_{k,m} + B(0,2\delta)$ are pairwise disjoint. Furthermore, we may choose δ small enough so that the sets $\gamma_{k,m} + B(0,2\delta)$ do not contain points from the curves $\gamma_{k+1,m'}$ by the fact that the sets of curves $\gamma_{k,m}$ and $\gamma_{k+1,m'}$ are mutually disjoint.

We choose enough division points in $\gamma_{k,m}$ and we connect them by segments (see Fig. 5) so that the union of these segments approximates the original curve. We definitely include two endpoints $a_{\gamma_{k,m}}$ and $b_{\gamma_{k,m}}$ in these division points and we assume that we have so many division points so that the union of these segments lies inside $\gamma_{k,m} + B(0,\delta)$. It follows that these segments for different $\gamma_{k,m}$ do not intersect.

However, it may happen that they intersect (see γ_3 in Fig. 5) for a given $\gamma_{k,m}$. In this case we simply choose a shortest path in the union of these segments between the

endpoints $a_{\gamma_{k,m}}$ and $b_{\gamma_{k,m}}$ and we replace the union of these segment by this shortest path (see the right side of Fig. 5). It is not difficult to see that by this replacement we get a one-to-one piecewise linear curve that replaces $\gamma_{k,m}$. Now we call $\Gamma_{k,j}$ the corresponding piecewise linear approximation of $\Gamma_{k,j}^{(**)}$. It is now easy to see that we have (1), (2) (using $\delta < 2^{-k}$), (3), (4), (5) and (6) for our $\Gamma_{k,j}$. Property (7) comes from our treatment of intersection points in Step 2, and the fact that in this step we chose δ small enough to not intersect the curves $\gamma_{k+1,m'}$. □

Parametrization of $\Gamma_{k,j}$: We have constructed a piecewise linear curve $\Gamma_{k,j}$ that approximated $\varphi(\partial Q_{k,j})$ and passes through the same image vertices $\mathcal{V}_k$ and intersection points $\mathcal{W}_k = \mathcal{G}_k \cap \mathcal{G}_{k+1}$. We know that there are four $y \in \mathcal{V}_k$ such that $y = \varphi(v)$ for some vertex of $Q_{k,j}$. Further, there are at most 8 points in

$$\mathcal{G}_{k+1} \cap \varphi(\partial Q_{k,j}) = \mathcal{G}_{k+1} \cap \Gamma_{k,j}$$

as on the image of each side of $Q_{k,j}$ there are at most two (see Fig. 3 and the proof of Lemma 5.1 (3)). Furthermore, we have at most two points in $\mathcal{G}_{k-1} \cap \varphi(\partial Q_{k,j})$, see Lemma 5.1 (3). Note also that analogously to the proof of Lemma 5.1 (3), there is $C > 0$ with such that

$$|\varphi^{-1}(y) - \varphi^{-1}(z)| \geq C 2^{-k}, \tag{5.4}$$

for any two distinct points $y, z \in \mathcal{V}_k \cup \mathcal{W}_k \cup \mathcal{W}_{k-1}$. Thus the distance between the preimages of these points is comparable to the sidelength of $Q_{k,j}$, i.e. 2^{-k}.

Now we divide $\Gamma_{k,j}$ into at most $4 + 8 + 2 = 14$ pieces P_i by these points in $\mathcal{V}_k \cup \mathcal{W}_k \cup \mathcal{W}_{k-1}$. For points $x \in \varphi^{-1}(\mathcal{V}_k \cup \mathcal{W}_k \cup \mathcal{W}_{k-1})$ we define $p(x) = \varphi(x)$ so that our parametrization p has the same value as original mapping φ on these "vertices" and intersection points. We parametrize the pieces P_i by a constant speed parametrization p there, i.e. on each of those pieces it has constant speed which might be different for each piece. Since the length of these pieces is bounded by $\mathscr{H}^1(\varphi(Q_{k,j}))$, we obtain using (5.4) that

$$|Dp| \leqslant C \frac{\mathscr{H}^1(\varphi(Q_{k,j}))}{2^{-k}} \text{ on the whole } Q_{k,j}.$$

6. The 2D extension

Let S be the square with vertices at $\{(1,0),(0,1),(-1,0),(0,-1)\}$ and $\mathbb{Y}$ be a Jordan domain with piecewise linear boundary. Suppose that a boundary homeomorphism $\varphi : \partial S \to \partial \mathbb{Y}$ is given. We now describe a way to extend φ as a homeomorphism of $\overline{S}$ to $\overline{\mathbb{Y}}$ with Lipschitz-continuity controlled by the boundary map.

First, we describe an extension H_φ of φ which is a *monotone map* from $\overline{S}$ to $\overline{\mathbb{Y}}$, meaning it is continuous and the preimage of every point is connected. The final homeomorphic extension will be obtained via an arbitrarily small modification of H_φ as we are able to

describe the points where it fails to be injective and fix them accordingly. However, this modification will be done only later in Section 8.

The extension H_φ will also be called the *shortest curve extension* of φ. To define H_φ, we let l_s denote the horizontal line segment which is obtained as the intersection between the line $\{(x, y) : y = s\}$ and S. This segment l_s has two endpoints a_s and b_s (from left to right) on ∂S. We let $A_s = \varphi(a_s)$, $B_s = \varphi(b_s)$, and define L_s as the shortest curve in $\varphi(\overline{S})$ which connects A_s to B_s.

The map H_φ is now given by defining it to map each horizontal segment l_s to the corresponding shortest curve L_s via constant speed parametrization. It is simple to verify that this mapping is continuous.

Lemma 6.1. *If $\varphi : \partial S \to \partial \mathbb{Y}$ is Lipschitz with constant L, then the shortest curve extension H_φ is also Lipschitz with constant at most CL for a uniform constant C.*

Proof. *Case 1.* Lipschitz continuity in the horizontal direction.

We show that H_φ satisfies the required Lipschitz-continuity on each of the horizontal segments l_s. For this, note that the constant speed parametrization on each of these segments implies that we only need to show that $|L_s| \leqslant 2L|l_s|$, where $|\cdot|$ denotes length. The endpoints of l_s separate ∂S into two connected components, the shorter of which we may call γ_s. Since L_t is the shortest curve from A_s to B_s, we find that $|\varphi(\gamma_s)| \geqslant |L_s|$. However, due to the Lipschitz-continuity of φ we must have that $|\varphi(\gamma_s)| \leqslant L|\gamma_s|$. Thus

$$|L_s| \leqslant |\varphi(\gamma_s)| \leqslant L|\gamma_s| \leqslant 2L|l_s|,$$

where the last inequality is due to the fact that l_s is the hypotenuse of a right-angled triangle with sides given by γ_s.

Case 2. Lipschitz continuity in the vertical direction.

Let us fix $s \in (-1, 1)$ and pick a point $z \in l_s$. For small δ we let $z_\delta = z + (0, \delta)$ and our aim is to show that $|H_\varphi(z_\delta) - H_\varphi(z)| \leqslant CL\delta$. As Lipschitz-continuity is a local property, we may assume that δ is arbitrarily small. In fact, to simplify calculations we assume that δ is very small compared to $|l_s|$, which lets us assume that the trapezium bounded by the segments l_s and $l_{s+\delta}$ is actually a rectangle with longer sides of length $|l_s|$ due to the fact that these two shapes are bilipschitz-equivalent with a uniform constant (say 2) for small enough δ.

Consider the curves L_s and $L_{s+\delta}$. By choosing δ small enough, we may assume that the endpoints A_s and $A_{s+\delta}$ lie on the same line segment of the piecewise linear boundary $\partial \mathbb{Y}$. The same may be assumed for B_s and $B_{s+\delta}$. Now basic geometry dictates that the curves L_s and $L_{s+\delta}$ must each consist of three parts as follows (for a detailed argument, see [18]). See also Fig. 6.

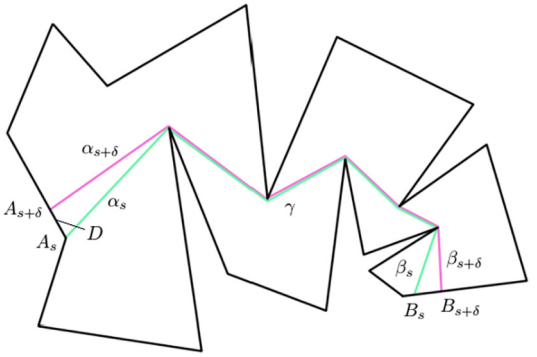


Fig. 6. The shortest curves L_s and $L_{s+\delta}$, split into three parts.

(1) α_s and $\alpha_{s+\delta}$: Curves which start from A_s and $A_{s+\delta}$ and do not intersect except at their common other endpoint. In fact, if δ is assumed small enough these curves may be assumed to be line segments.
(2) A common part of l_s and L_s, which is a piecewise linear curve we denote by γ.
(3) β_s and $\beta_{s+\delta}$: Analogously to the first part, these can be assumed to be line segments from B_s and $B_{s+\delta}$ respectively which meet at a common point (the other endpoint of γ).

We may assume that $H_\varphi(z)$ lies on either α_s or γ as the case where it lies on β_s is handled by symmetry. Let D denote the line segment between A_s and $A_{s+\delta}$. Then since φ is L-Lipschitz-continuous on ∂S, we find that $|D| \leqslant L\delta$. By the triangle inequality we obtain that $||\alpha_s| - |\alpha_{s+\delta}|| \leqslant L\delta$ and using the same argument for the β-curves gives $||L_s| - |L_{s+\delta}|| \leqslant 2L\delta$. Let also d denote the distance between z and a_s, which is also the distance from z_δ to $a_{s+\delta}$.

Suppose first that $H_\varphi(z)$ lies on γ. The length of the part of L_s between A_s and $H_\varphi(z)$ may now be calculated in two ways. The constant speed parametrization tells us that it is equal to $|L_s|d/|l_s|$. On the other hand, it is also equal to $|\alpha_s| + |\gamma'|$, where γ' denotes the part of γ between α_s and $H_\varphi(z)$. Thus

$$|\alpha_s| + |\gamma'| = \frac{|L_s|d}{|l_s|}.$$

If Γ denotes the part of $L_{s+\delta}$ between $H_\varphi(z)$ and $H_\varphi(z_\delta)$, then we may calculate the length of the part of $L_{s+\delta}$ between $a_{s+\delta}$ and $H_\varphi(z_\delta)$ in two ways similarly as above to obtain that

$$|\alpha_{s+\delta}| + |\gamma'| \pm |\Gamma| = \frac{|L_{s+\delta}|d}{|l_s|}.$$

The $\pm$ in this equation is there to account for the two cases on which side of $L_{s+\delta}$ the point $H_\varphi(z_\delta)$ lies in comparison to $H_\varphi(z)$. In either case, we find by combining the above two equalities that

$$\begin{aligned}|\Gamma| &\leqslant ||\alpha_s| - |\alpha_{s+\delta}|| + ||L_s| - |L_{s+\delta}||\frac{d}{|l_s|}\\ &\leqslant L\delta + 2L\delta.\end{aligned}$$

This shows that $|H_\varphi(z_\delta) - H_\varphi(z)| \leqslant 3L\delta$.

Suppose then that $H_\varphi(z)$ lies on α_s. The length of the part of α_s from A_s to $H_\varphi(z)$ must then be equal to $|L_s|d/|l_s|$ by constant speed parametrization. Let ω be a point on $\alpha_{s+\delta}$ of distance at most $|D|$ from $H_\varphi(z)$, which is possible to choose due to the concavity of $\alpha_{s+\delta}$ and α_s towards each other (more precisely, concavity towards the interior of the region defined by them and D). Let γ^* denote the part of $\alpha_{s+\delta}$ between $A_{s+\delta}$ and ω, and Γ the part of $L_{s+\delta}$ between ω and $H_\varphi(z_\delta)$. Both the part of α_s from A_s to $H_\varphi(z)$ and the curve γ^* are shortest curves between their respective endpoints, and since the endpoints are connected by curves of length at most $|D| \leqslant L\delta$, we get by triangle inequality that

$$\left||\gamma^*| - \frac{|L_s|d}{|l_s|}\right| \leqslant 2L\delta.$$

Thus we find that

$$\begin{aligned}|\Gamma| &\leqslant \left|\frac{|L_{s+\delta}|d}{|l_s|} - |\gamma^*|\right|\\ &\leqslant ||L_s| - |L_{s+\delta}||\frac{d}{|l_s|} + 2L\delta\\ &\leqslant 4L\delta.\end{aligned}$$

This shows that $|H_\varphi(z_\delta) - H_\varphi(z)| \leqslant 4L\delta$ and proves our claim.

As a clarifying remark, note that as we approach the top and bottom vertices $(0, \pm 1)$ on ∂S the corresponding shortest curves shrink to a single point. In this case the above estimates still go through with even some further simplification.

Note: We will use the following consequence of this proof repeatedly in multiple other parts of the paper. Given a Jordan domain $\mathbb{Y}$ with a piecewise linear boundary and points $A_1, A_2, B \in \partial\mathbb{Y}$, suppose that the part of $\partial\mathbb{Y}$ between A_1 and A_2 which does not contain B has length δ'. Then if $\varphi_1, \varphi_2 : [0,1] \to \overline{\mathbb{Y}}$ are the two shortest curves in $\overline{\mathbb{Y}}$ from B to A_1 and A_2 respectively, parametrized with constant speed, then $|\varphi_1(x) - \varphi_2(x)| \leqslant C\delta'$ for all $x \in [0,1]$. This claim follows from the above proof, notably the only difference is that we start from the same point B instead of two points B_s and $B_{s+\delta}$ but this case is even simpler. □

6.1. Lipschitz-continuity in the time variable

Our next aim is to look at a situation where instead of a single given boundary map φ, we are given a continuous sequence of boundary homeomorphisms $\varphi_t : S \to \mathbb{R}^2, t \in [0,1]$

(not necessarily to the same target domain). The aim is to show that if the dependence on t is Lipschitz, meaning that

$$|\varphi_{t_1}(z) - \varphi_{t_2}(z)| \leqslant L|t_1 - t_2| \qquad \text{for } z \in \partial S, \tag{6.1}$$

then the same estimate holds (up to a uniform constant) for the extensions H_{φ_t} as well. We expect this to be true in the general case, but for our purposes we will only need to prove such a result in a few simple cases which are easier to explain. Let us denote by $S_- := \{(x, y) \in S : x \leqslant 0\}$ the union of the two left sides of S and by S_+ the union of the two right sides.

Lemma 6.2. *Suppose that $\mathbb{Y} \subset \mathbb{C}$ is a piecewise linear Jordan domain and $\varphi_t : \partial S \to \partial \mathbb{Y}$ are given boundary homeomorphisms so that (6.1) is valid. Suppose also that the maps $\varphi_t(z)$ are equal on one half of S, say $\varphi_t(z) = \varphi_0(z)$ for all $z \in S_+$. Then $H_{\varphi_t}(z)$ is CL-Lipschitz in (z, t), where C is a uniform constant.*

Proof. Let $z \in \overline{S}$. We consider the horizontal segment l passing through z and its two endpoints a and b. Fixing the point $t_1 \in (0, 1)$, by continuity we choose $t_2 \in (0, 1)$ close enough to t_1 so that $\varphi_{t_1}(b)$ and $\varphi_{t_2}(b)$ lie on the same segment on $\partial \mathbb{Y}$. By our assumptions also $\varphi_{t_1}(a) = \varphi_{t_2}(a)$. For φ_{t_1}, we let L^{t_1} denote the shortest curve from $\varphi_{t_1}(a)$ to $\varphi_{t_1}(b)$ in $\overline{\mathbb{Y}}$. Similarly L^{t_2} is the shortest curve from $\varphi_{t_1}(a)$ to $\varphi_{t_2}(b)$. Then $H_{\varphi_{t_1}}(z)$ lies on L^{t_1} and $H_{\varphi_{t_2}}(z)$ lies on L^{t_2} and the exact positioning of these points on these curves is again determined by the constant-speed parametrization on the horizontal segment l. But this situation is essentially exactly the same as in the second case of the proof of Lemma 6.1 (see note at the end of that proof), and we may apply the same proof to show that

$$|H_{\varphi_{t_1}}(z) - H_{\varphi_{t_2}}(z)| \leqslant 4L|t_1 - t_2|. \quad \square$$

We now address how the shortest curve extension behaves with respect to a changing target boundary.

Definition 6.3. Let us first define that a *simple modification* of a piecewise linear Jordan curve $\varphi : S \to \partial \mathbb{Y}$ is any other piecewise linear curve φ^* obtained as follows. Let P be a vertex of φ and let $P_1 = \varphi(s_1)$ and $P_2 = \varphi(s_2)$ be its two neighbouring vertices. We pick another point Q on the ray $\overrightarrow{P_1P}$ and define φ^* as the piecewise linear curve through the vertices of φ with P replaced with Q.

Regarding parametrization we require that $\varphi(s) = \varphi^*(s)$ for all s except for those in the preimage of the segments P_1P and PP_2 under φ, and moreover that these preimages are either both contained in S_+ or both in S_-.

Next, a homotopy $\varphi_t : \partial S \to \mathbb{R}^2, t \in [0, 1]$ of piecewise linear Jordan curves is called a *simple homotopy* if for all t_1 and $t_2 > t_1$ sufficiently close to t_1, the curve φ_{t_2} may be obtained from φ_{t_1} via a simple modification as described above.

Lemma 6.4. *If a homotopy $\varphi_t : S \to \mathbb{R}^2, t \in [0,1]$ of piecewise linear Jordan curves is simple and Lipschitz-continuous in (z,t) with constant L, then the shortest curve extensions H_{φ_t} are also Lipschitz-continuous in (z,t) with constant CL for a uniform constant C.*

Proof. We aim to use the same types of arguments as in the proof of Lemma 6.1 to obtain Lipschitz estimates for H_{φ_t} in t, but we must elaborate further as we are dealing with two shortest curves within two different domains. However, it is enough to show Lipschitz-estimates locally and hence we are able to use the condition of φ_t being a simple homotopy to deduce that on the given time interval the two image domains are similar apart from one added or removed triangle (this triangle is ΔPQP_2 in Definition 6.3).

Let thus $z \in S$ and $t_1, t_2 \in [0,1]$. Let l be the horizontal segment in S which passes through z and let a and b be its endpoints from left to right. Suppose that we are in the case where the mappings φ_{t_i} are equal on S_-, so that $\varphi_{t_1}(a) = \varphi_{t_2}(a)$. Let $\mathbb{Y}_{t_i}$ be the Jordan domain bounded by $\varphi_{t_i}(\partial S)$ and L^{t_i} be the shortest curve within the closure of $\mathbb{Y}_{t_i}$ between $\varphi_{t_i}(a)$ and $\varphi_{t_i}(b)$. We also let $p_{t_i} := \varphi_{t_i}(b)$.

By assumption of simpleness of φ_t the only difference between the boundaries of $\mathbb{Y}_{t_1}$ and $\mathbb{Y}_{t_2}$ is the addition or removal of a triangle ΔPQP_2. The curve $\partial\mathbb{Y}_{t_1}$ traverses straight from P to P_2 while $\partial\mathbb{Y}_{t_2}$ goes through the point Q inbetween.

Due to some distinct geometrical possibilities here, we split the argument into cases as follows. Recall that the shortest curves L^{t_1} and L^{t_2} have one common endpoint and their non-common endpoints are p_{t_1} and p_{t_2}. We split into cases based on whether one of these points p_{t_i} belongs to the part of the boundary being changed or not.

Case 1. If p_{t_1} does not lie on the segment of $\partial\mathbb{Y}_{t_1}$ between P and P_2.

In this case, p_{t_1} lies on the common boundary of $\mathbb{Y}_{t_1}$ and $\mathbb{Y}_{t_2}$. We now define another map on the horizontal segment l by considering the shortest curve from $\varphi_{t_1}(a)$ to p_{t_1}, but this time within the closure of $\mathbb{Y}_{t_2}$. Let this map be called $\Phi : l \to \overline{\mathbb{Y}_{t_2}}$ and parametrize it in constant speed also. Then the result of Lemma 6.1 shows that $|H_{\varphi_{t_2}}(z) - \Phi(z)|$ may be estimated from above in terms of a constant times the length of the boundary of $\mathbb{Y}_{t_2}$ between p_{t_1} and p_{t_2}. But the boundary estimates from before show that this length may be estimated from above by $CL|t_1 - t_2|$.

Hence due to the triangle inequality we have

$$|H_{\varphi_{t_2}}(z) - H_{\varphi_{t_1}}(z)| \leqslant |H_{\varphi_{t_2}}(z) - \Phi(z)| + |\Phi(z) - H_{\varphi_{t_1}}(z)|.$$

It remains to consider the quantity $|\Phi(z) - H_{\varphi_{t_1}}(z)|$. This quantity depends on the curves L^{t_1} and $\Phi(l)$. These curves are both shortest curves from $\varphi_{t_1}(a)$ to p_{t_1}. However, one is within the domain $\mathbb{Y}_{t_1}$ and the other is within the domain $\mathbb{Y}_{t_2}$. Thus we are to investigate how this change of domain affects the behaviour of the shortest curve. We split again into cases based on a few different geometrical possibilities.

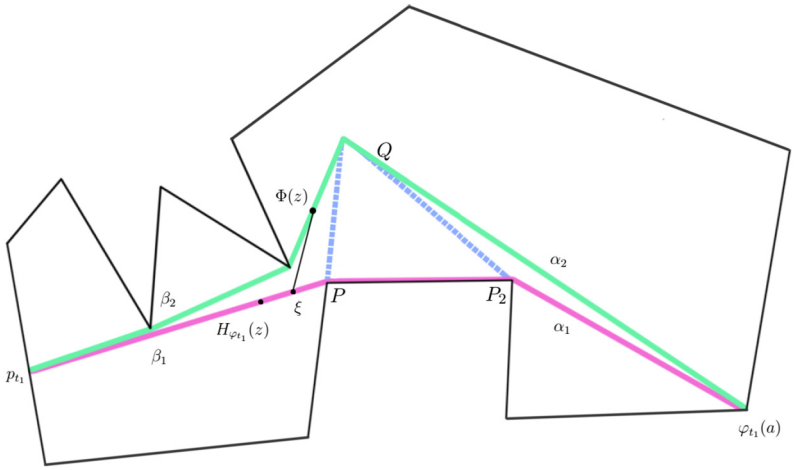


Fig. 7. Case 1: Two shortest curves between $\varphi_{t_1}(a)$ to p_{t_1} in different domains. Here $\partial\mathbb{Y}_{t_1}$ is denoted by the black piecewise linear curve, while $\mathbb{Y}_{t_2}$ is created from $\mathbb{Y}_{t_1}$ by adding a triangle ΔPQP_2. Note that the segment PP_2 is part of α_1.

Case 1a. Suppose that the curve L^{t_1} does not touch the segment PP_2.

Since L^{t_1} is the shortest curve between $\varphi_{t_1}(a)$ and $\varphi_{t_1}(b)$ in $\mathbb{Y}_{t_1}$, if $\mathbb{Y}_{t_2} \subset \mathbb{Y}_{t_1}$ then $\Phi(l)$ (the shortest curve between the same points in $\mathbb{Y}_{t_2}$) must be at least as long as L^{t_1}. But since L^{t_1} does not intersect PP_2 we must have $L^{t_1} \subset \mathbb{Y}_{t_2}$ and thus $L^{t_1} = \Phi(l)$. If $\mathbb{Y}_{t_2}$ is not contained in $\mathbb{Y}_{t_1}$, which is when Q lies outside of $\mathbb{Y}_{t_1}$, then it still must hold that $L^{t_1} = \Phi(l)$ because the shortest curve $\Phi(l)$ cannot pass through the interior the triangle ΔPQP_2 as it can only enter and exit through the segment PP_2. Thus there is nothing to prove in this case. *Case 1b.* Suppose that $P \in L^{t_1}$ and $Q \in \Phi(l)$.

Let the part of L^{t_1} between $\varphi_{t_1}(a)$ and P be called α_1 and the part from P to p_{t_1} be called β_1. Similarly, let the part of $\Phi(l)$ from $\varphi_{t_1}(a)$ to Q be α_2 and from Q to p_{t_1} be β_2. Let $|P - Q| = \delta$.

Let us say that a simple curve in $\overline{\mathbb{Y}_1}$ does not cross the segment PQ if that curve is a uniform limit of curves within $\mathbb{Y}_1 \setminus PQ$, parametrizations may be taken in arc length here. Note that none of the curves $\alpha_1, \alpha_2, \beta_1$ and β_2 pass through the interior of the triangle ΔPQP_2 and also do not cross the segment PQ. Hence within the class of curves in $\overline{\mathbb{Y}_1}$ which do not cross the segment PQ, these curves are also the shortest curves between their respective endpoints.

We suppose that $\Phi(z)$ is on β_2. The case where it is on α_2 is proven similarly. We define a point $\xi \in \beta_1$ as the intersection point of β_1 with the line passing through $\Phi(z)$ and parallel to PQ (see Fig. 7). Due to the fact that β_1 and β_2 are shortest curves in $\overline{\mathbb{Y}_1}$ which do not cross the segment PQ, the segment from $\Phi(z)$ to ξ lies entirely between these two curves and has length smaller than δ - this can be argued similarly as the convexity part in Case 2 of Lemma 6.1. Let β_2^* be the part of β_2 from p_{t_1} to $\Phi(z)$ and β_1^* be the part of β_1 from p_{t_1} to ξ. Then a simple shortest curve estimate shows that

$$||\beta_2^*| - |\beta_1^*|| \leqslant |\Phi(z) - \xi| \leqslant \delta. \tag{6.2}$$

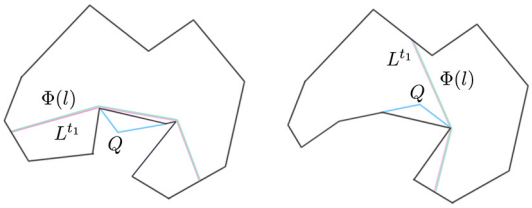


Fig. 8. Case 1d: Reduces to pictured possibilities in which the curves L^{t_1} and $\Phi(l)$ are the same.

Similarly, we may find that

$$\begin{aligned} ||\beta_2| - |\beta_1|| &\leqslant \delta \\ ||\alpha_1| - |\alpha_2|| &\leqslant \delta. \end{aligned} \tag{6.3}$$

Now consider the length of the part of $H_{\varphi_1}(l)$ between p_{t_1} and $H_{\varphi_{t_1}}(z)$, call this length τ. Due to constant speed parametrization, if the distance from a to z is x, we find that $\tau = (|\alpha_1| + |\beta_1|)x/|l|$. But since $x \leqslant |l|$ and the estimates (6.3), we find that

$$|\tau - |\beta_2^*|| = \left|\tau - \frac{(|\alpha_2| + |\beta_2|)x}{|l|}\right| \leqslant 2\delta.$$

However, (6.2) then implies that $|\tau - |\beta_1^*|| \leqslant 3\delta$. This essentially says that the part of the curve β_1 between ξ and $H_{\varphi_{t_1}}(z)$ has length at most 3δ, and hence we also have the Euclidean distance estimate $|\xi - H_{\varphi_{t_1}}(z)| \leqslant 3\delta$ and finally also $|\Phi(z) - H_{\varphi_{t_1}}(z)| \leqslant 4\delta$ from (6.2). Since $\delta \leqslant CL|t_1 - t_2|$ this is enough.

Case 1c. Suppose $Q \in \Phi(l)$, $P \notin L^{t_1}$ but either L^{t_1} passes through PQ or through QP_2.

If L^{t_1} passes through PQ, let the intersection point of PQ and L^{t_1} be X. This case can be handled the same way as Case 1b, with X taking the role of P. The case where L^{t_1} passes through QP_2 can be handled symmetrically.

Case 1d. Suppose that $Q \notin \Phi(l)$.

This case appears either when the point Q is outside the domain $\mathbb{Y}_{t_1}$ or when L^{t_1} only intersects the triangle ΔPQP_2 at one of the vertices P or P_2 (see Fig. 8). In all of these cases the curves L^{t_1} and $\Phi(l)$ are the same, and there is nothing to prove. This handles all the possible options and finishes the proof of Case 1.

Case 2. If p_{t_2} lies on a part of $\partial\mathbb{Y}_{t_2}$ which is not on the segments PQ or QP_2. This case may be treated with the same arguments as Case 1, with t_1 and t_2 interchanged.

This covers the cases where either p_{t_1} or p_{t_2} lies outside the triangle ΔPQP_2, leaving the case where both points lie on respective sides of this triangle.

Case 3. We suppose that p_{t_1} lies on the segment PP_2 and p_{t_2} on either PQ or QP_2.

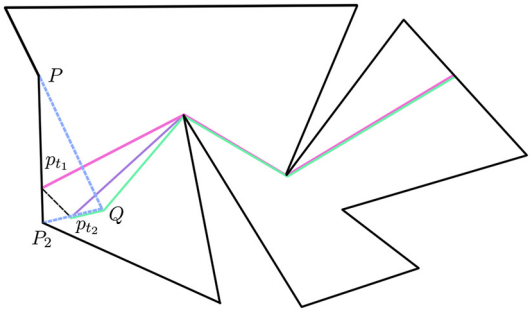


Fig. 9. Case 3a: Shortest curves to p_{t_1} and p_{t_2} when T is inside of $\mathbb{Y}_{t_1}$. In this case, $\hat{\mathbb{Y}}$ is obtained by taking $\partial\mathbb{Y}_1$ and replacing $p_{t_1}P_2$ with $p_{t_1}p_{t_2}$ and $p_{t_2}P_2$. Again $\mathbb{Y}_{t_2}$ is created from $\mathbb{Y}_{t_1}$ by adding a triangle ΔPQP_2.

By symmetry, we may suppose that p_{t_2} lies on QP_2 and that $t_1 < t_2$. We now consider the triangle $T = \Delta PQP_2$, but must split into cases depending on if this triangle is inside or outside of $\mathbb{Y}_{t_1}$.

Case 3a. If T is inside of $\mathbb{Y}_{t_1}$. The shortest curve L^{t_1} must pass through T before it reaches its endpoint at p_{t_1} (see Fig. 9). Moreover, the part of L^{t_1} inside the closure of T must be a single segment since T is convex. Now, the point p_{t_2} splits the union of the segments PQ and QP_2 into two parts. Let $\hat{\gamma}$ be the part which does not intersect L^{t_1}.

The idea now is to create a new domain $\hat{\mathbb{Y}}$. We take the Jordan curve $\partial\mathbb{Y}_{t_1}$, add the union of $p_{t_1}p_{t_2}$ and $\hat{\gamma}$ to it, and remove the segment of $\partial\mathbb{Y}_{t_1}$ which has the same endpoints as this union does (either we remove $p_{t_1}P_2$ or $p_{t_1}P$). This Jordan curve now defines $\hat{\mathbb{Y}}$. An equivalent definition is to cut off from $\mathbb{Y}_{t_1}$ a region bounded by $p_{t_1}p_{t_2}$ and $\hat{\gamma}$. The key point is that by this construction the curve L^{t_1} still lies in the closure of $\hat{\mathbb{Y}}$. In fact, the curve L^{t_1} is still the shortest curve from $\varphi_{t_1}(a)$ to p_{t_1} within the new domain $\hat{\mathbb{Y}}$. This is due to the fact that the shortest curve from $\varphi_{t_1}(a)$ to p_{t_1} does not change if we remove a region of the domain which does not intersect this shortest curve to begin with.

Let now $\Phi : l \to \overline{\hat{\mathbb{Y}}}$ denote the shortest curve from $\varphi_{t_1}(a)$ to p_{t_2} in the closure of $\hat{\mathbb{Y}}$, parametrized with constant speed. Now we split our estimates via the triangle inequality

$$|H_{\varphi_{t_2}}(z) - H_{\varphi_{t_1}}(z)| \leqslant |H_{\varphi_{t_2}}(z) - \Phi(z)| + |\Phi(z) - H_{\varphi_{t_1}}(z)|.$$

The quantity $|\Phi(z) - H_{\varphi_{t_1}}(z)|$ may now be estimated via the arguments of Lemma 6.1, since both Φ and $H_{\varphi_{t_1}}$ map the horizontal segment l to a shortest curve within $\hat{\mathbb{Y}}$, and the distance between their endpoints p_{t_2} and p_{t_1} is estimated from above by $CL|t_1 - t_2|$.

The quantity $|H_{\varphi_{t_2}}(z) - \Phi(z)|$ is dealt with the same arguments as Case 1, since Φ and $H_{\varphi_{t_2}}$ map the horizontal segment l to shortest curves from $\varphi_{t_1}(a)$ to p_{t_2}, however in different domains $\hat{\mathbb{Y}}$ and $\mathbb{Y}_{t_2}$. The difference between these domains is, again, small.

Case 3b. If T is outside of $\mathbb{Y}_{t_1}$. This case is handled much the same as the previous one, only now we create $\hat{\mathbb{Y}}$ from $\mathbb{Y}_{t_2}$ by adding $p_{t_1}p_{t_2}$ and the part of PP_2 which does not intersect L^{t_2}. We also remove either $p_{t_2}P_2$ or the two segments of $\partial\mathbb{Y}_2$ between p_{t_2} and

P to create the Jordan curve that bounds $\hat{\mathbb{Y}}$. Now the situation is dealt with the same arguments as the previous case. □

Lemma 6.5. *Suppose that $\varphi_0, \varphi_1 : \partial S \to \mathbb{R}^2$ are two piecewise linear embeddings of the square ∂S into $\mathbb{R}^2$. Let $\mathbb{Y}_0$ and $\mathbb{Y}_1$ be the Jordan domains bounded by the respective image curves $\varphi_0(\partial S)$ and $\varphi_1(\partial S)$. Suppose that $\varphi_0(z) = \varphi_1(z)$ for all $z \in \partial S_-$ and both maps have constant speed on S_+. Suppose that the curves $\varphi_0(S_+)$ and $\varphi_1(S_+)$ do not intersect except for their endpoints. Suppose also that both embeddings φ_0 and φ_1 are Lipschitz-continuous with constant L. Then there exists a homotopy φ_t, $t \in [0,1]$ of piecewise linear curves which is simple, CL-Lipschitz in (z,t), and φ_t lies within the region bounded by φ_0 and φ_1.*

Proof. Let $\gamma_0 = \varphi_0(S_+)$ and $\gamma_1 = \varphi_1(S_+)$. We first describe a homotopy γ_t between these two curves, which will then be used to construct φ_t by setting $\varphi_t(S_+) = \gamma_t$ and fixing a parametrization. On S_- we naturally set $\varphi_t \equiv \varphi_0$.

The curve γ_t is defined as follows. Let the mutual endpoints of γ_0 and γ_1 be A and B and the domain between these curves be denoted by $\hat{\mathbb{Y}}$. Let $\gamma_{1/2}$ be the shortest path from A to B within the closure of $\hat{\mathbb{Y}}$. We now need to only describe how to deform γ_0 to $\gamma_{1/2}$ as the case from $\gamma_{1/2}$ to γ_1 will be handled in the same way.

For $t \in [0, 1/2]$, note that $2t$ varies from 0 to 1. We choose γ_t as follows. First, travel along γ_0 starting from A until we have travelled a curve of length $2t|\gamma_0|$. We have arrived at a point of γ_0 which we shall call P_t. For the remainder of the parametrization, we take the shortest curve from P_t to B within the closure of $\hat{\mathbb{Y}}$. This defines γ_t up to parametrization, and the exact parametrization of γ_t will be defined now.

We divide the time interval $[0, 1/2]$ into intervals $[t_n, t_{n+1})$ so that for all $t \in [t_n, t_{n+1})$ the curve φ_t is obtained from φ_{t_n} via simple modification, at least as long as we now guarantee that the parametrization aligns with the requirements in Definition 6.3. For a fixed parameter t, the curves γ_t and γ_0 agree on the initial part of γ_0 of length $2t|\gamma_0|$. For those $s \in S_+$ for which $\varphi_0(s)$ is on this initial part, we also set $\varphi_t(s) = \varphi_0(s)$. Let $s_{2t} \in S_+$ be defined so that $P_t = \varphi_t(s_{2t})$, and recall that the curves γ_{t_n} and γ_t for $t \in [t_n, t_{n+1})$ only differ by moving P_{t_n} to P_t. Let $t' > t$ be so that $P' = \varphi_{t_n}(s_{2t'})$ is the next vertex after P_{t_n} on this curve, so that P' is also the next vertex after P_t for φ_t. Now if the angle $\angle P_t P_{t_n} P'$ is concave (above π) towards the interior, then the curves γ_{t_n} and γ_t are the same and we may also set the parametrizations φ_{t_n} and φ_t to be exactly the same.

In the case where the angle is convex (less than π), we set $\varphi_t(s) = \varphi_{t_n}(s)$ for all $s \geqslant 2t'$. It remains to define φ_t on $(2t, 2t')$ assuming by induction that φ_{t_n} is given. Let the union of the segments $P_t P_{t_n}$ and $P_{t_n} P'$ be U_1 and let U_2 denote the segment $P_t P'$. We choose a constant speed map $\Psi_t : U_1 \to U_2$, and this constant is smaller than one because U_2 is shorter than U_1. Then we define $\varphi_t(s) = \Psi_t(\varphi_{t_n}(s))$ for $s \in (2t, 2t')$.

This shows that the Lipschitz constant of φ_t in s decreases as t increases. It remains to obtain estimates in t. It is enough to show that $|\varphi_t(s) - \varphi_{t_n}(s)| \leqslant CL|t - t_n|$ for

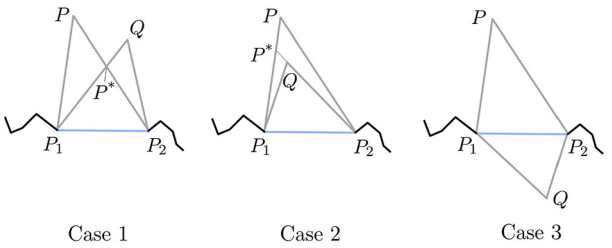


Fig. 10. Moving the point P to Q through a point P^* via two simple modifications.

$s \in (2t, 2t')$. For this, through some simple geometry we see that the distance between the points $\varphi_t(s)$ and $\varphi_{t_n}(s)$, which lie on the sides of the triangle $\Delta P_t P_{t_n} P'$, can be estimated from above by the length of the side $P_t P_{t_n}$. But $|P_t - P_{t_n}| = |\varphi_{t_n}(2t) - \varphi_{t_n}(2t_n)| \leqslant 2L|t - t_n|$, which finishes the proof. □

Definition 6.6. A homotopy $\varphi_t : \partial S \to \mathbb{R}^2, t \in [0,1]$ of piecewise linear Jordan curves is called a *2-simple homotopy* if for all t_1 and $t_2 > t_1$ sufficiently close to t_1, the curve φ_{t_2} may be obtained from φ_{t_1} via two successive simple modifications on the same vertex P.

The difference between one and two simple modifications is that in a simple modification the point P is only moving along the ray $\overrightarrow{P_1 P}$, while after two simple modifications the point P may technically move to any other in the plane. In our case, some further restrictions will apply as we must also maintain injectivity during this process.

Lemma 6.7. *If $\varphi_t : S \to \mathbb{R}^2, t \in [0,1]$ is a 2-simple homotopy of piecewise linear Jordan curves and Lipschitz-continuous in (z,t) with constant L, then the shortest curve extensions H_{φ_t} are also Lipschitz-continuous in (z,t) with constant CL for a uniform constant C.*

Proof. Fix t_1 and let $t_2 > t_1$ be close to t_1. Then Definition 6.6 implies that there is a simple modification which turns φ_{t_1} into another curve φ^* and another simple modification which turns φ^* into φ_{t_2}. It is enough to show that we may choose φ^* so that the estimate $|\varphi_{t_2}(s) - \varphi^*(s)| \leqslant CL|t_1 - t_2|$ is satisfied, as then the two simple modifications $\varphi_{t_1} \mapsto \varphi^*$ and $\varphi^* \mapsto \varphi_{t_2}$ can be seen to be CL-Lipschitz-continuous in (z,t) and we may finish by applying the proof of Lemma 6.4 to obtain the desired result.

Let P be the vertex on the curve φ_{t_1} being moved to the vertex Q on φ_{t_2}, and let P_1 and P_2 be their shared neighbouring vertices. Let us pick t_2 close enough to t_1 so that P and Q are on the same side of the segment P_1P_2, eliminating Case 3 in Fig. 10. We may assume that the ray $\overrightarrow{P_1 Q}$ intersects the segment P_2P at a point P^* (otherwise we consider the intersection of $\overrightarrow{P_2 Q}$ and P_1P, or switch the roles of P and Q). Now due to the assumption that the homotopy φ_t is Lipschitz continuous in t with constant L, we have that $\operatorname{dist}(Q, P_1P \cup PP_2) \leqslant L|t_1 - t_2|$ (at least for t_2 close enough to t_1 so that there is no interference from the rest of the curve). Due to some elementary geometry

the distance $|QP^*|$ from Q to P^* must be comparable to the distance from Q to the segments P_1P and PP_2, giving that $|QP^*| \leqslant CL|t_1 - t_2|$. Now let us compose φ_{t_2} with a piecewise linear map which is otherwise the identity but sends the segments P_1Q and QP_2 to P_1P^* and P^*P_2 respectively. This is a simple modification of φ_{t_2} which we call φ^*. Each point on the curve φ_{t_2} is moved at most a distance of $|QP^*|$, which gives the desired estimate $|\varphi_{t_2}(s) - \varphi^*(s)| \leqslant |QP^*| \leqslant CL|t_1 - t_2|$. Moreover, it is clear that φ_{t_1} is a simple modification of φ^* as P^* lies on P_2P. Thus the proof is complete. □

7. The 3D extension

We now proceed to the construction of the extension h into the upper half space, continuing the proof of Theorem 1.2 along the lines described at the start of Section 4. The main goal here is to define h precisely on each $U_{k,j}$. Recall the definition of the sets $U_{k,j}, \tilde{Q}_{k,j}$ and curves $\Gamma_{k,j}$ from Section 4.

Step 1. We define h on the sides of the top and bottom faces of $U_{k,j}$. We wish to map the top sides $\partial\tilde{Q}_{k,j} \times \{2^{-(k-1)}\}$ to the Jordan curve $\Gamma_{k,j}$ and the bottom sides $\partial\tilde{Q}_{k,j} \times \{2^{-k}\}$ to $\hat{\Gamma}_{k,j}$. Note that here and what follows we abuse ∂ to mean the 1D boundary of these sets rather than taking the topological boundary of the sets in 3D space.

Step 2. We define h on the top and bottom faces of $U_{k,j}$. To simplify notation, we set $\mathcal{U}_t = \tilde{Q}_{k,j} \times \{t\}$. Furthermore, let $top := 2^{-(k-1)}$ and $bot := 2^{-k}$ so that $\mathcal{U}_{top}$ is the top face and $\mathcal{U}_{bot}$ is the bottom one. Similarly we set $\varphi_t = h|_{\partial\mathcal{U}_t}$ and $h_t = h|_{\mathcal{U}_t}$, although only φ_{top} and φ_{bot} have been defined so far. On $\mathcal{U}_{top}$, we simply define h_{top} as the shortest curve extension of φ_{top}. Note that this choice also forces us to define h_{bot} on $\mathcal{U}_{bot}$ in a specific way to avoid discontinuity. Indeed, the bottom side $\mathcal{U}_{bot}$ is in fact the union of four top sides of dyadic cubes of the form $U_{k+1,j'}$ on the next level. Thus on $\mathcal{U}_{bot}$ the map h_{bot} is defined separately in each of the four squares as the shortest curve extension of the corresponding boundary values.

Step 3. Let $mid := 2^{-k}+2^{-k-1}$ be the middle point of $[2^{-k}, 2^{-(k-1)}]$ so that $\mathcal{U}_{mid}$ is the middle level of the cube $U_{k,j}$. On the sides of $\mathcal{U}_{mid}$ and for every parameter $t \in [bot, mid]$, we define φ_t equal to φ_{bot}. On $\mathcal{U}_{mid}$ we define h_{mid} as the shortest curve extension of φ_{mid}. Hence for $t \in [bot, mid]$, the mapping h_t has the same boundary values on each level $\mathcal{U}_t$ but is a different map on the faces $\mathcal{U}_{mid}$ and $\mathcal{U}_{bot}$. We return to this part in a later step and describe how to define h_t for $t \in (bot, mid)$ to give the correct isotopy between the maps h_{mid} and h_{bot}.

Step 4. For $t \in [mid, top]$, we will define h_t as the shortest curve extension of φ_t. However, we have not yet defined φ_t for these parameters. Note that the image of φ_{top} is $\Gamma_{k,j}$ and the image of φ_{mid} is $\hat{\Gamma}_{k,j}$. Thus we must define a homotopy φ_t between these two curves which is what we will do now.

The left part of Fig. 11 depicts the curves $\Gamma_{k,j}$ and $\hat{\Gamma}_{k,j}$. Since the curves $\Gamma_{k,j}$ (respectively $\hat{\Gamma}_{k,j}$) form a grid topologically equivalent with a dyadical grid, we may abuse terminology here and talk about vertices and edges of $\Gamma_{k,j}$ when considered as a topo-

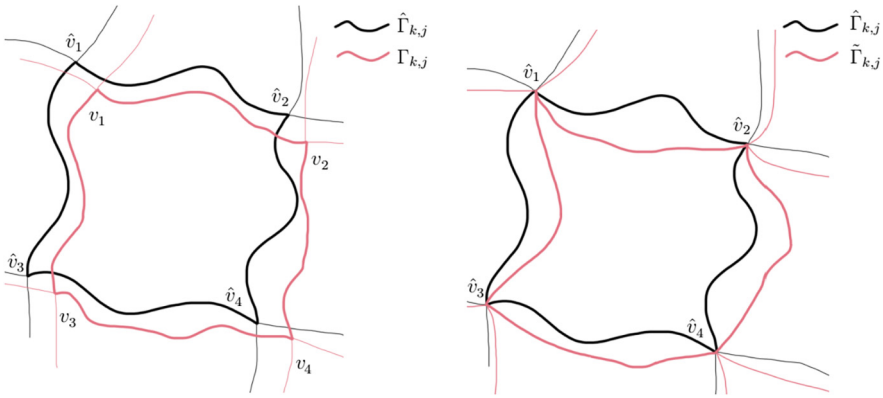


Fig. 11. On the left, the curve $\Gamma_{k,j}$ and its corresponding curve $\hat{\Gamma}_{k,j}$ on the next level. On the right, $\Gamma_{k,j}$ has been modified to $\tilde{\Gamma}_{k,j}$. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

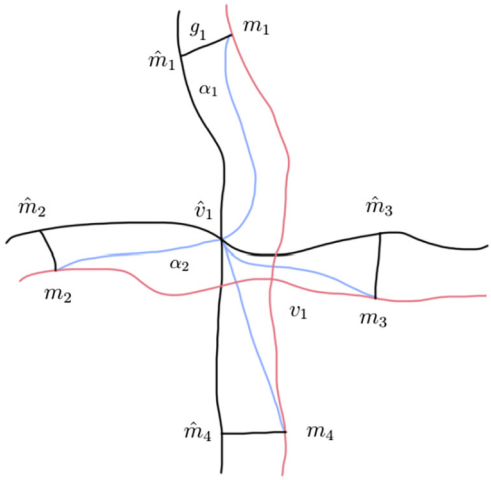


Fig. 12. The plus-shaped region whose boundary consists of two crosses and curves from the points m_i to $\hat{m}_i$.

logical square. As in the figure, let us label the vertices of these curves by v_j and $\hat{v}_j$, $j = 1, 2, 3, 4$ in corresponding order. We pick one pair of such vertices, say v_1 and $\hat{v}_1$. The vertex v_1 is the intersection point of two edges of $\Gamma_{k,j}$ as well as two other edges in the same grid, for a total of four. We let the midpoint of the edges meeting at v_1 be m_j, $j = 1, 2, 3, 4$, see Fig. 12. We similarly define four points $\hat{m}_j$ as the midpoints of the edges in the grid formed by the curves $\tilde{\Gamma}_{k,j}$ which meet at $\hat{v}_1$, numbered correspondingly to the points m_j. We now connect each of the points m_j with $\hat{m}_j$ through a piecewise linear curve g_j which does not intersect either of the grids and has length comparable to the infimal length of such curves.

Our aim now is to deform the cross formed by the curves with endpoints at $m_1, \ldots, m_4$ and intersecting at v_1, to a cross with the same endpoints but middle point at $\hat{v}_1$ instead.

Naturally we wish to introduce no new intersection points during this homotopy and keep the deformation within the plus-shaped region pictured in Fig. 12. At each point in time the cross we are considering meets four different dyadic regions in the image side, and we wish to create this deformation between crosses in a way where we can apply Lemma 6.7 for each of these four regions to obtain the required interior Lipschitz-estimates. Thus it is necessary to form the homotopy in a way that with respect to each four regions the part of the border that is deforming behaves as a 2-simple homotopy (see Definition 6.6). A fixed number of reparametrizations of curves is also needed in the arguments used here, but we recall that Lemma 6.2 allows us to do so while still maintaining the required interior estimates.

We first connect the points m_1 and $\hat{v}_1$ with a piecewise linear Jordan curve α_1 which does not intersect any of the other considered curves and has distance comparable to the sum of the length of the curve g_1 from m_1 to $\hat{m}_1$ and the curve from $\hat{m}_1$ to $\hat{v}_1$ which is part of $\hat{\Gamma}_{k,j'}$ for some j'. This can be done for example by choosing a curve sufficiently close to those two curves but not intersecting them or itself. Similarly, we define a curve α_2 from m_2 to $\hat{v}_1$, see again Fig. 12.

Let ψ_0 be the union of the curves from m_1 to v_1 and from v_1 to m_2, parametrized on $[0,1]$. Similarly, let ψ_1 be the union of α_1 and α_2. We may assume that $\psi_0(1/2) = v_1$ and $\psi_1(1/2) = \hat{v}_1$. Using the method of Lemma 6.5 we connect ψ_0 to ψ_1 via a homotopy ψ_t. We define a curve from v_1 to $\hat{v}_1$ by $\Psi(t) = \psi_t(1/2)$.

This homotopy from ψ_0 to ψ_1 gives one part of the sought homotopy between the two crosses. Let β_1 denote the curve from m_3 to v_1 and β_2 the curve from m_4 to v_1. We denote by ψ_0^* the union of β_1 and β_2, parametrized again on $[0,1]$ with $\psi_0^*(1/2) = v_1$. We wish to construct another simple homotopy ψ_t^* with $\psi_t^*(0) = m_3$, $\psi_t^*(1/2) = \psi_t(1/2) = \Psi(t)$, $\psi_t^*(1) = m_4$, and so that the curve ψ_t^* has no additional intersection points with ψ_t.

At each time t we must find curves from m_3 and m_4 to $\Psi(t)$. In order to do this we first describe the properties of the curve $\Psi(t)$, as this curve may not be injective. Following the construction done in Lemma 6.5, the domain bounded by the two curves ψ_t and ψ_1 is decreasing as a function of t. Thus it is not possible for the curve $\Psi(t)$ to form a proper loop to intersect itself, but a priori it can be constant on some interval and it can also travel backwards along itself. For the moment, let us describe the construction of ψ_t^* while assuming that $\Psi(t)$ does not intersect itself or ψ_0^*.

The idea of the construction of the homotopy ψ_t^* is to add to the initial curve ψ_0^* a part which follows close to the curve Ψ to a certain point and then returns back along another path close to Ψ. At $t = 1$ we will travel the full length of the curve Ψ to the point $\hat{v}_1$ and back.

Let us suppose that the homotopy ψ_t^* has been defined up to a point t_n where $P_n := \Psi(t_n)$ is a vertex on the piecewise linear curve given by Ψ. Let P_{n+1} be the next vertex after P_n on Ψ, and let P_{n-1}^1 and P_{n-1}^2 denote the two neighbouring vertices of P_n on the curve $\psi_{t_n}^*$. The aim now is to "open up" a part of the segment P_nP_{n+1} into two segments $P_n^1Q_t$ and $P_n^2Q_t$, but some care must be made to not cause intersections, see the rightmost part of Fig. 13 to illustrate this process.

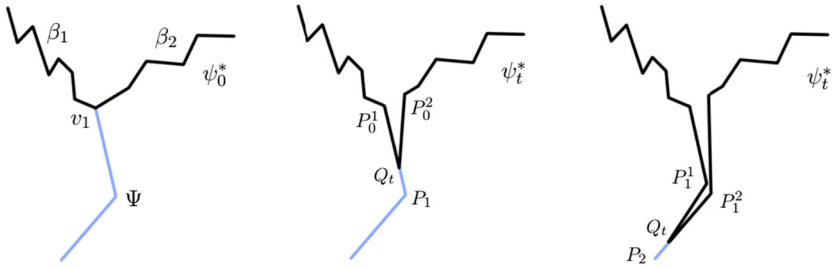


Fig. 13. Opening up the curve Ψ to create a homotopy of Jordan curves.

More precisely, let us suppose that the angle $\angle P_{n-1}^1 P_n P_{n+1}$ (interpreted as the smaller angle of the two choices) is smaller or equal than $\angle P_{n-1}^2 P_n P_{n+1}$ (again, the smaller choice). We pick another point P_n^1 on the segment $P_n P_{n-1}^1$ which may be chosen arbitrarily close to P_n. We may let $P_n^2 := P_n$ in this case, if the size of the two angles $\angle P_{n-1}^1 P_n P_{n+1}$ and $\angle P_{n-1}^2 P_n P_{n+1}$ is reversed then so is the role of P_n^1 and P_n^2.

For a point $t_{n+1} > t_n$ to be chosen later, we will now define ψ_t^* for $t \in (t_n, t_{n+1}]$. For $t \in (t_n, t_{n+1}]$ let X_t denote a point parametrized linearly on $P_n P_{n+1}$ so that $X_{t_n} = P_n$ and $X_{t_{n+1}} = P_{n+1}$. For each $t \in (t_n, t_{n+1}]$ we now define ψ_t^* by mapping the preimage of the segment $P_n^1 P_n$ to $P_n^1 X_t$ and the preimage of $P_n^2 P_n$ to $P_n^2 X_t$. This simply corresponds to moving the point P_n along the segment $P_n P_{n+1}$ to the point X_t while keeping the parametrization consistent, see Fig. 13. By choosing P_n^1 close enough to P_n we can guarantee that no new intersection points are created during this process (since by assumption Ψ does not intersect itself), and that the added length is comparable to the length of Ψ.

Let us elaborate a bit further on the parametrization of the curves ψ_t^* used here. We pick one constant speed parametrization Θ from $I := [1/4, 3/4]$ to the final curve between P_0^1 and P_0^2 defined by the process above. This final curve travels arbitrarily close to Ψ all the way up to $\hat{v}_1$ and then back along a similar curve to P_0^2. Let us first reparametrize the initial curve ψ_0^* in order to guarantee that a small part is not mapped to Θ in the end. We choose ψ_0^* to map the intervals $[1/4, 1/2]$ and $[1/2, 3/4]$ to the two segments $P_0^1 P_0$ and $P_0 P_0^2$, keeping the relation $\psi_0^*(1/2) = P_0 = v_1$. The exact parametrization can be inherited backwards from the final parametrization Θ, so that the preimage of the segments $P_n^1 X_t$ and $X_t P_n^2$ under each curve ψ_t^* for $t \in [t_n, t_{n+1})$ is the same set as the preimage of the part of Θ between P_n^1 to P_n^2. As the latter image curve is longer we may guarantee that the Lipschitz-constant of ψ_t^* on $[1/4, 3/4]$ is controlled by the length of Θ.

The parametrization in the time variable t can also be chosen based on Θ. In fact, as long as we pick the time intervals $[t_n, t_{n+1})$ to have comparable length to the total length of the preimage of the segments $P_{n-1}^1 P_n^1$ and $P_{n-1}^2 P_n^2$ under Θ, the Lipschitz constant in the time direction will be bounded from above by a constant times the length of Θ.

Thus the boundary curves ψ_t^* have the correct Lipschitz bounds, and we turn our attention to interior estimates. Note that there are four different regions meeting at the

cross with centre $\Psi(t)$. Let us denote the region which only meets ψ_t by V_t^1, the region which only meets ψ_t^* by V_t^4, and let V_t^2 and V_t^3 be the two regions which meet one half of both of these curves. We let U_t^i denote the corresponding dyadic squares on the domain side (which, if interpreted as planar sets, are the same set for each t), whose boundaries are all identified with S for the sake of constructing the shortest curve extension to V_t^i.

In each of the sets V_t^i, one part of the boundary is fixed while the deformation of the other part is dictated by the homotopies ψ_t and ψ_t^*. Whichever domain V_t^i is chosen, locally in t the deformation only consists of moving around the single vertex $\Psi(t)$. Hence as long as the preimage (in U_t^i) of the part being deformed corresponds to being either contained completely in S_+ or completely in S_-, this homotopy induces a homotopy on ∂V_t^1 which is at worst a 2-simple homotopy (see Definition 6.6). The preimage being contained entirely in S_+ or S_- happens exactly when the preimage of v_1 happens to be identified with the vertices $(0, \pm 1)$ on S, while the opposite is true when $\Psi(t)$ is identified with $(\pm 1, 0)$.

If the homotopy of ∂V_t^i is indeed 2-simple, then Lemma 6.7 implies that the shortest curve extension satisfies the required interior Lipschitz bounds. We need hence address the case where $\Psi(t)$ is identified with $(\pm 1, 0)$. Note that in the definition of the shortest curve extension which is now applied inside the diamond shaped domain U_t^i, there is an implicit choice of horizontal/vertical direction based on which two opposing vertices we pick as the top and bottom vertices. If we choose the direction where the horizontal lines point towards the preimage of $\Psi(t)$, then the condition of the deformation being contained inside S_+ or S_- in Definition 6.3 is satisfied. Naturally we cannot a priori choose the orientation to always satisfy this condition as exactly two of the vertices of U_t^i satisfy this condition and two do not, and eventually we will need to repeat this argument with respect to crosses with centres at each of the four vertices of U_t^i.

We take care of this issue with the following trick. Let ρ denote a bilipschitz map from the square domain bounded by S to the unit disk, and let $\nu_t(z) = e^{i\pi t/2} z$ denote a rotation map on the unit disk. Let $\tilde{H}_0 : U_t^i \to V_t^i$ denote the shortest curve extension of a boundary map $\tilde{\varphi}_0 : \partial U_t^i \to \partial V_t^i$. We then define a new map $\tilde{H}_t$ on U_t^i by making a change of variables on the domain side in U_t^i (identified with S) via the map $\rho^{-1} \circ \nu_t \circ \rho$, and instead of extending $\tilde{\varphi}$ from ∂U_t^i we extend the map $\tilde{\varphi}_t := \tilde{\varphi}_0 \circ \rho^{-1} \circ \nu_{-t} \circ \rho$ via shortest curve extension. Thus $\tilde{H}_t$ and $\tilde{H}_0$ have the same boundary values but differ in the interior. In essence, $\tilde{H}_t$ corresponds to "rotating" the horizontal lines in U_t^i by an angle $\pi t/2$ and constructing the shortest curve extension based on these new curves. But we only need to know that for $t = 1$ the map $\tilde{H}_1$ corresponds to constructing the shortest curve extension with the horizontal lines in U_t^i replaced by vertical lines, which can be done by choosing the bilipschitz map ρ accordingly. The homotopy $\tilde{H}_t$ can be seen to be Lipschitz continuous in (z, t) with constant CL, where L is the Lipschitz constant of $\tilde{\varphi}_0$. This follows from the Lipschitz continuity of ρ, ν_t and their inverses, and an application of Lemma 6.2 since $\tilde{\varphi}_t$ satisfies the correct bounds in (z, t). The homotopy $\tilde{H}_t$ can be used to temporarily change the direction of horizontal lines in U_t^i to suit our purposes,

showing that we may reduce to the previous case where the homotopy on the boundary is 2-simple.

Let us now address the fact that in general the curve Ψ may intersect itself. Perhaps the easiest way to deal with this is to make a slight modification on the construction of Lemma 6.5, as the homotopy of curves γ_t parametrized on $[0,1]$ constructed in that lemma defines Ψ by the relation $\gamma_t(1/2) = \Psi(t)$. We will now make a slight perturbation of the curves γ_t to make them mutually non-intersecting, which will guarantee that $\Psi(t)$ becomes injective.

Note that if two of the curves γ_t do intersect, they in particular intersect at a vertex P of $\partial\hat{\mathbb{Y}}$, where $\hat{\mathbb{Y}}$ denotes the Jordan domain bounded by the curves γ_0 and γ_1. At any such vertex P we attach to it a small segment PV_P facing the interior of $\hat{\mathbb{Y}}$ and bisecting the angle of $\partial\hat{\mathbb{Y}}$ at P.

Now for each such segment we consider all the curves γ_t which pass through PV_P and let the intersection point of γ_t with this segment be P_t. Thus for those parameters t the map $t \to P_t$ defines either an increasing or decreasing parametrization of PV_P, which is not strictly monotone as some interval of parameters is sent to the point P. However, we may make an arbitrarily small modification to this parametrization to make it strictly monotone, replacing each point P_t with another point P_t^* on PV_P.

This gives us a way to replace each of the piecewise linear curves γ_t by another curve γ_t^* which, for each segment PV_P that intersects γ_t, passes through the point P_t^* instead of P_t. As this modification may be done in an arbitrarily small way we may assume that the Lipschitz estimates we obtained before for φ_t and for H_{φ_t} also hold after the modification up to a multiplicative constant arbitrarily close to 1. Thus although the new homotopy induced by the curves γ_t^* is not necessarily simple, it gives the desired Lipschitz-estimates inside and all of the curves γ_t^* are mutually nonintersecting. For further details also see Section 8 where a similar construction is explained in more depth.

This concludes the construction of the homotopy of the two crosses with centres v_1 and $\hat{v}_1$. After doing this process for every vertex v_j and every curve $\Gamma_{k,j}$ on level k, we have replaced the curve $\Gamma_{k,j}$ with another curve $\tilde{\Gamma}_{k,j}$ with the same vertices as $\hat{\Gamma}_{k,j}$ but not intersecting it, see Fig. 11. The homotopy between $\tilde{\Gamma}_{k,j}$ and $\hat{\Gamma}_{k,j}$ is now easy to construct. Between each pair of neighbouring vertices, say $\hat{v}_1$ and $\hat{v}_2$, we deform the part of $\tilde{\Gamma}_{k,j}$ into $\hat{\Gamma}_{k,j}$ via the method explained in Lemma 6.5. After deforming each four parts in succession we have deformed $\tilde{\Gamma}_{k,j}$ into $\hat{\Gamma}_{k,j}$.

Still in the situation of Fig. 11, we provide a few more details regarding parametrization and estimates happening here. We may divide the interval $[mid, top]$ into two halves, on one of which we deform $\Gamma_{k,j}$ into $\tilde{\Gamma}_{k,j}$ and on the other $\tilde{\Gamma}_{k,j}$ into $\hat{\Gamma}_{k,j}$. To offer more details on what happens in the first half, we divide the first half further into four intervals so that on each we move one of the vertices v_j to the corresponding point $\hat{v}_j$, $j = 1, 2, 3, 4$.

In the first half, the length of the relevant curves is always controlled from above by $|\Gamma_{k,j}| + |\hat{\Gamma}_{k,j}|$, plus the same quantity over the neighbours of $\Gamma_{k,j}$. As the initial curves are parametrized with constant speed we know by Lemma 6.4 that the Lipschitz-constant of

the shortest curve extension h in the (z,t)-variables is thus controlled by $2^k(|\Gamma_{k,j}|+|\hat{\Gamma}_{k,j}|)$ added with this quantity over the neighbours.

In the second half, each part of $\tilde{\Gamma}_{k,j}$ having two of the $\hat{v}_i$ as endpoints is deformed to the part of $\hat{\Gamma}_{j,k}$ with the same endpoints. Here we are again using Lemma 6.4 and therefore the Lipschitz-constant is estimated from above by $2^k(|\Gamma_{k,j}| + |\hat{\Gamma}_{k,j}|)$.

Step 5. For $t \in [bot, mid]$, the situation is as follows. The maps h_{mid} and h_{bot} have already been defined. We interpret these maps as planar maps, identifying the horizontal sections $\mathcal{U}_t$ of the cube $U_{k,j}$ on the domain side with the same square domain which we call $\mathcal{U}$. Both maps h_{mid} and h_{bot} are hence interpreted to be defined on $\mathcal{U}$ and as they have the same boundary map $\varphi_{mid} = \varphi_{bot}$, we may interpret them to map $\mathcal{U}$ into the same target domain $\mathcal{V}$ bounded by the piecewise linear Jordan curve $\varphi_{mid}(\partial\mathcal{U})$. The difference between these two maps is that h_{mid} is defined by the shortest curve extension of φ_{mid} and h_{bot} is defined as the shortest curve extension of its boundary values in each of the four child squares of $\mathcal{U}$.

Let us denote by $\mathcal{C}$ the cross formed by the two segments between opposing midpoints of the sides of $\mathcal{U}$. Hence the way h_{mid} maps $\mathcal{C}$ is determined by the shortest curve extension and we denote the image cross by $T_{mid} = h_{mid}(\mathcal{C})$. The way h_{bot} maps $\mathcal{C}$ is predetermined by the piecewise linear approximations of the original boundary map defined in Section 5. We denote $T_{bot} = h_{bot}(\mathcal{C})$.

A key point to note is the following. Let $\mathcal{U}'$ denote one of the four children of $\mathcal{U}$. Then we claim that h_{mid} restricted to $\mathcal{U}'$ is actually the shortest curve extension of its boundary value on $\partial\mathcal{U}'$. Let ℓ denote one of the horizontal line segments inside $\mathcal{U}'$ (the meaning of 'horizontal' here is as it was used in the definition of the shortest curve extension), with a and b being its endpoints. Then ℓ is part of a horizontal segment of $\mathcal{U}$ and is mapped to a curve under h_{mid} which is the shortest such curve between its endpoints. This must mean also that the curve is the shortest curve from $h_{mid}(a)$ to $h_{mid}(b)$ inside $\overline{\mathcal{U}'}$. Moreover, since h_{mid} maps each horizontal segment in $\mathcal{U}$ to its target curve with constant speed, h_{mid} must also have constant speed on ℓ. This cements the fact that h_{mid} on $\mathcal{U}'$ is the shortest curve extension of its boundary values.

However, the above argument has the following minor defect. In Section 6, the shortest curve extension was defined for a boundary map from a square to a piecewise linear Jordan domain. But the map h_{mid} might not map the two line segments making up $\mathcal{C}$ to true Jordan curves as the shortest curve extension may fail to be injective and thus the image cross T_{mid} may touch the boundary in $\overline{\mathcal{V}}$. Nevertheless, these curves are still piecewise linear and are given by a uniform limit of Jordan curves. There is no issue defining the notion of shortest curves and shortest curve extensions to areas bounded by such degenerate Jordan curves as well, and the estimates we have established before in results such as Lemma 6.2 and Lemma 6.4 extend naturally to this setting as well. This can be seen by verifying that the proofs go through in the degenerate case as well.

From now the strategy to define a homotopy h_t for $t \in [bot, mid]$ is as follows. For each such t, the map h_t on $\partial\mathcal{U}$ will have the same boundary values φ_{mid}. Moreover, we

will define a homotopy of crosses T_t between the two crosses T_{mid} and T_{bot}. Once such a homotopy has been defined and parametrized as a map $\Phi_t : \mathcal{C} \to T_t$, for each child $\mathcal{U}'$ of $\mathcal{U}$ we define h_t on $\mathcal{U}'$ as the shortest curve extension of its boundary values on $\partial\mathcal{U}'$. Thus h_t will be equal to φ_{mid} on $\partial\mathcal{U}$ and to Φ_t on $\mathcal{C}$.

To construct the homotopy between the two crosses, we would like to apply the same argument from Step 4 which was used to create a homotopy between the crosses depicted in Fig. 12. However, in the argument from Step 4 it was essential that the two crosses only had two intersection points (on the curves between v_1, m_1 and v_1, m_2). In our case, the crosses T_{mid} and T_{bot} may have arbitrarily many intersection points. To address this issue, we define another cross T_{fix} which satisfies this property respective to both the crosses T_{mid} and T_{bot}, and then simply deform first T_{mid} to T_{fix} and then to T_{bot}. Due to Lemma 6.2, the exact nature of the parametrization Φ_t does not play a role here and we may assume for example that on each of the four arms of $\mathcal{C}$ the parametrization always has constant speed.

Before defining T_{fix}, we make a small modification to T_{mid} in order to replace it with a cross T_{mid*} which does not intersect the boundary except at the four endpoints. Since the cross T_{mid} consists of piecewise linear curves, this modification can be done by moving each of its vertices that touch the boundary (except for the four endpoints) by an arbitrarily small amount towards the interior of $\mathcal{V}$ so that the resulting cross does not intersect itself nor $\partial\mathcal{V}$. This modification provides a homotopy from T_{mid} to T_{mid*} which we may, for example, dedicate the first quarter of the interval $[bot, mid]$ towards in t. The fact that this modification to the cross may be done in an arbitrarily small way guarantees that the Lipschitz estimates (in t) both on $\mathcal{C}$ and for the shortest curve extensions to the four regions of $\mathcal{V}$ can be controlled by above with a constant of our choice.

It now remains to define T_{fix}. Since neither of the crosses T_{mid*} and T_{bot} touch the boundary $\partial\mathcal{V}$ except at their common four endpoints, we may choose T_{fix} for example as follows. We pick a point P in $\mathcal{V}$ close enough to an image point of a corner of $\mathcal{U}$ under φ_{mid} so that P belongs to $h_{mid*}(\mathcal{U}') \cap h_{bot}(\mathcal{U}')$ for one of the children $\mathcal{U}'$ of $\mathcal{U}$. Then we connect P to the four endpoints of T_{mid*} via piecewise linear curves to form the cross T_{fix}. These curves, if chosen to run sufficiently close along the boundary $\partial\mathcal{V}$, may be assumed to satisfy the necessary properties of not intersecting themselves or each other. Moreover, they can be chosen so that two of them intersect T_{mid*} and T_{bot} exactly once and two of them do not intersect these crosses (apart from the endpoints). See Fig. 14. This means that the crosses T_{fix} and T_{mid*} are in the same configuration as the crosses in Step 4, and the same goes for T_{fix} and T_{bot}. Hence we may repeat the argument to find a homotopy between these crosses, and extend the boundary values defined by this via the shortest curve extension to the whole of $\mathcal{U}$. For each t, we lift the copy of $\mathcal{U}$ and the map h_t to the appropriate horizontal section at height t in $U_{k,j}$ and $V_{k,j}$ in order to fully define our extension there.

We have thus defined the extension h as a monotone map on each set $U_{k,j}$ to the image set $V_{k,j}$. We now return to our original goal of controlling the Lipschitz constant of h in

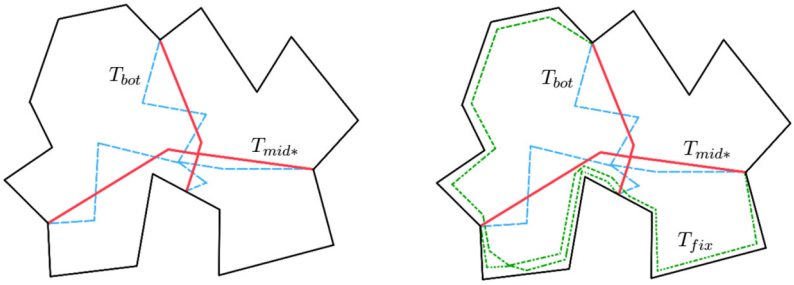


Fig. 14. Constructing an intermediate cross T_{fix}. The original crosses T_{mid*} and T_{bot} are denoted in red and blue colour and they intersect a lot. Thus we construct a new intermediate cross T_{fix} denoted in green which does not intersect T_{mid*} and T_{bot} too much.

$U_{k,j}$. For the readers convenience, we recall that the goal here amounts to showing that the Lipschitz constant of h in $U_{k,j}$ is controlled from above by $2^k|\Gamma_{k,j}|$ plus possibly the same quantity over the dyadic neighbours and children of $U_{k,j}$. Note that the quantity $2^k|\hat{\Gamma}_{k,j}|$ is equivalent with the Lipschitz constant of a constant speed parametrization of $\hat{\Gamma}_{k,j}$ over the boundary of the dyadic square on generation k.

To justify that this bound is maintained throughout $U_{k,j}$, we explain as follows. In Step 4, the Lipschitz constant of the boundary value isotopy φ_t is controlled by above (in both the space and t variable) by the lengths of the corresponding boundary curves and possibly the lengths of the neighbouring curves. Lemma 6.4 then shows that this implies the correct Lipschitz estimates for h in the region where $t \in [mid, top]$. In the region $t \in [bot, mid]$, the map h is defined piecewise as the shortest curve extension yet again, so to obtain the correct Lipschitz estimates one needs only estimate the length of the boundary curves on the image side. These consist of the original boundary curve $\partial\mathcal{V}$ and the lengths of the crosses T_{mid}, T_{fix} and T_{bot}. The first two can be bounded from above by a constant times the length of $\partial\mathcal{V}$ (which is the length of $\hat{\Gamma}_{k,j}$, while the last one is bounded by the lengths of the image curves of the children $\hat{\Gamma}^{(m)}_{k,j}$. Thus we get the desired estimate that yields a bound on the $W^{1,q}$-norm of h in terms of the quantity on the left hand side of (1.3).

8. Making it all injective

Let $\varphi : \partial S \to \partial\mathbb{Y}$ be a homeomorphic boundary map to a Jordan domain $\mathbb{Y}$ with piecewise linear boundary. We now describe how to tackle the issue that the shortest curve extension H_φ is not injective but rather a monotone map. The main issue is that the images of two horizontal segments l_{s_1} and l_{s_2} of S may intersect each other or intersect the boundary of the image domain $\partial\mathbb{Y}$. However, the saving grace is that these images are shortest curves between their respective endpoints and thus do not cross, allowing us to make a minor modification to the curves so that they do not intersect each other or touch the boundary and therefore create a homeomorphic extension H^*_φ of φ. This modification is not too difficult for a single map and was done already in [18]. However, in our case more details are needed as we need to make this modification

consistent in a way that if φ_t is a continuous family of boundary maps, not necessarily to the same image domain, then the modified extensions $H^*_{\varphi_t}$ need to be continuous in t and the modification must be done in a way to preserve the Lipschitz estimates in terms of φ_t.

We will now describe a precise way of constructing the injectification of a single shortest curve extension $H_\varphi : S \to \mathbb{Y}$. One may imagine here that $H_\varphi = H_{\varphi_t}$ for some homotopy of maps φ_t but with a specific fixed parameter t. We drop the subscript t for ease of presentation, however. We will define this injectification process with dependence on certain auxiliary parameters (such as D, defined later), and one should keep in mind that these parameters will need to be interpreted as functions of t later. By later fixing their dependence on t we will be able to argue that the process ensures both continuity in t for the extensions as well as the required Lipschitz-estimates.

Firstly, we may assume here that the boundary map φ is also piecewise linear, as such is the case in the whole construction done in previous sections, where φ is always defined piecewise as a constant speed map. When φ and $\partial\mathbb{Y}$ are piecewise linear, it is not difficult to check that then also the shortest curve extension H_φ becomes a piecewise affine map on $\overline{S}$.

The aim is to show that the modification from the shortest curve extension H_φ to its homeomorphic variant H^*_φ may be done in an arbitrarily small way in the following sense. On each horizontal segment l_s, the map H_φ maps l_s to a shortest curve L_s with constant speed. The map H^*_φ instead maps l_s to another piecewise linear curve L^*_s, also with constant speed, and so that L^*_s may be obtained from L_s by shifting each vertex of L_s by a small distance. We will show that such distances can be chosen to be arbitrarily small, controlled by a single constant per map, which means that the modified map $H^*_{\varphi_t}$ will also be arbitrarily close to H_φ which lets us obtain the same Lipschitz-estimates for it.

The idea behind modifying the curves L_s to the curves L^*_s is quite simple. At each vertex of $\partial\mathbb{Y}$ where L_s passes through, we move that vertex of L_s a little bit further away from the boundary. For curves $L_{s'}$ with $s' > s$, this movement should be a little bit larger for vertices on $\partial\mathbb{Y}$ on the image of the part of ∂S below l_s and a little smaller for vertices on $\partial\mathbb{Y}$ on the image of the part of ∂S above l_s. See Fig. 15.

We now begin the precise definitions. Let us define a number D as the minimal length between two sides of $\partial\mathbb{Y}$ which are not neighbours. Next, for any point $P \in \partial\mathbb{Y}$ we define the *inner normal* of P, denoted ℓ_P, as the ray which starts from the point P, points towards the interior of $\mathbb{Y}$ near P, and forms equal angles with $\partial\mathbb{Y}$ i.e. is an angle bisector for the angle of $\partial\mathbb{Y}$ formed at P.

For every vertex $P \in \partial\mathbb{Y}$, we pick a positive number $\epsilon_P < 1$ whose role will become apparent later in making the modification process continuous in t. We then define the point V_P as the point on ℓ_P which is of distance $\epsilon_P D/3$ away from P. By the definition of D, the point V_P must be at a distance of at least $2D/3$ away from any other side of

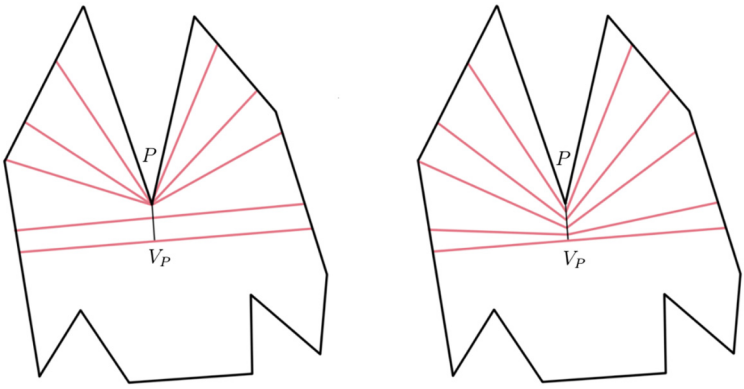


Fig. 15. Modifying the curves L_s on the segment PV_P.

$\partial\mathbb{Y}$ than the two P lies on. This means that apart from the point P, the segment PV_P cannot intersect $\partial\mathbb{Y}$ nor can it intersect any other such segment QV_Q for another vertex Q of $\partial\mathbb{Y}$.

Note that two of the shortest curves L_s may only intersect at points on $\partial\mathbb{Y}$. Since the point V_P is inside $\mathbb{Y}$, for each P there must be a unique parameter s_P for which L_{s_P} passes through V_P. We also define $\hat{s}_P$ as the parameter for which P is one of the endpoints of $L_{\hat{s}_P}$. Thus the curves L_s which intersect the segment PV_P are exactly those for which $s \in [s_P, \hat{s}_P]$. It can also be possible that $s_P = \hat{s}_P$, in which case the segment PV_P belongs fully to the curve L_{s_P}. This is also the only case in which a curve L_s intersects PV_P more than once. In this case we will not modify the curve $L_{\hat{s}_P}$ which is equivalent with setting $\epsilon_P = 0$.

Suppose that $s_P > \hat{s}_P$. For each $s \in [\hat{s}_P, s_P]$ there is a unique point X_s on PV_P which belongs to L_s. Let $f_P : [\hat{s}_P, s_P] \to [0, \epsilon_P D/3]$ denote the function which sends s to $|X_s - P|$. Now f_P is an increasing and surjective piecewise linear function, strictly increasing on the preimage of $(0, \epsilon_P D/3]$, but it is possible that f_P sends a nontrivial interval of parameters $[\hat{s}_P, x]$ to 0. In fact, this happens exactly in the case where there are multiple curves L_s that intersect at P.

The idea now is the following. We pick a strictly increasing surjective piecewise linear function $f_P^* : [\hat{s}_P, s_P] \to [0, \epsilon_P D/3]$ to act as an injective replacement for f_P. We wish to make a canonical choice here so for an increasing surjective function $f_P : [0,1] \to [0,1]$ for which $f^{-1}(\{0\}) = [0, A]$ we set

$$f_P^*(x) = \begin{cases} x/(2A) & \text{when } x \in [0, A], \\ (f_P(x) + 1)/2 & \text{otherwise.} \end{cases}$$

The way we will modify each curve L_s for $s \in [\hat{s}_P, s_P]$ is by moving the point X_s on L_s to a new point X_s^* on PV_P so that $|X_s^* - P| = f_P^*(s)$.

If $s_P < \hat{s}_P$, we do the exact same process as above only on the interval $[s_P, \hat{s}_P]$ on which the analogously defined function f_P will be decreasing instead of increasing.

Similarly we choose f_P^* as a strictly decreasing function. We now define the curves L_s^*. For each curve L_s, we make note of all the segments PV_P which this curve passes through. We only consider segments with $s_P \neq \hat{s}_P$ as to neglect cases where the segment PV_P is fully on L_s. On each of the applicable segments PV_P intersecting L_s we move the point X_s on the curve L_s to X_s^*. Note that the curves L_{s_P} and $L_{\hat{s}_P}$ are not modified with respect to the process specific to the segment PV_P (although they may be changed when we repeat this process on other segments QV_Q).

Step 1. Proving that the curves L_s^* do not intersect $\partial\mathbb{Y}$ except at their endpoints.

Fix s and consider the curve L_s. For each vertex P of $\partial\mathbb{Y}$, we consider the segments PV_P. We recall these segments are mutually disjoint. Considering the intersection points of L_s with all such segments PV_P, this splits the curve L_s into segments $Q_0Q_1, Q_1Q_2, \ldots, Q_{N-1}Q_N$ so that Q_0, Q_N are the endpoints of L_s and for each Q_j, there is a point P_j which is a vertex of $\partial\mathbb{Y}$ so that $Q_j \in P_jV_{P_j}$. Moreover, we assume that there are no other such points on L_s.

Consider now a segment Q_jQ_{j+1} with $0 < j < N-1$. During the deformation from L_s to L_s^*, the point Q_j is moved on the segment $P_jV_{P_j}$ to another point Q_j^*. Suppose for the contrary that the segment $Q_j^*Q_{j+1}^*$ intersects the boundary $\partial\mathbb{Y}$. Let $Q_j^r = (1-r)Q_j + rQ_j^*$. As neither Q_j^* or Q_{j+1}^* belong to $\partial\mathbb{Y}$, there must be a minimal number $0 < r < 1$ so that $Q_j^rQ_{j+1}^r$ intersects $\partial\mathbb{Y}$. We now consider two cases:

(1) If a vertex P of $\partial\mathbb{Y}$ intersects $Q_j^rQ_{j+1}^r$. Basic geometry dictates that such a vertex P cannot share a side with P_j or P_{j+1}. If P equals Q_j^r or Q_{j+1}^r, this contradicts the definition of D as then the distance from P to either P_j or P_{j+1} would be too small, seeing as $|Q_j^r - P_j| \leqslant D/3$ holds for all j and r due to $Q_j^r \in P_jV_{P_j}$. If P is strictly between Q_j^r and Q_{j+1}^r, then again a simple geometrical argument shows that there must be a non-endpoint of Q_jQ_{j+1} which is on PV_P, a contradiction with the definition of the points Q_j.
(2) If a point X of $\partial\mathbb{Y}$ which is not a vertex intersects $Q_j^rQ_{j+1}^r$. We obtain a similar contradiction as above if X is either of Q_j^r or Q_{j+1}^r. In the case where X is strictly inside $Q_j^rQ_{j+1}^r$, the segment of $\partial\mathbb{Y}$ on which X is on must be parallel to $Q_j^rQ_{j+1}^r$, as otherwise a smaller choice of r would result in the segments still intersecting but contradicting the minimality of r. But for any two segments which are parallel and intersect each other, one must contain an endpoint of the other one. Thus this reduces to one of the cases already considered.

Step 2. Proving that the curves L_s^* do not intersect each other.

If two of the curves L_s^* and $L_{s'}^*$ intersected each other with $s < s'$. Then for all $r \in (s, s')$ the curve L_r^* would also necessarily intersect both L_s^* and $L_{s'}^*$ or either it would provide a separation between them. But for r close enough to s, the curves L_r^* and L_s^* may not

intersect. This is due to the fact that these curves may be decomposed into the same number of segments I_j^r and I_j^s, $j = 1, \dots, N$, and so that $I_j^r \to I_j^s$ as $r \to s$. This convergence implies that for r close enough to s, the segment I_j^r cannot intersect $I_{j'}^s$, unless $j' \in \{j-1, j, j+1\}$. However, even in this case these segments may not intersect due to geometrical reasons, as the nature of the construction guarantees that I_j^r and I_j^s do not intersect.

Step 3. Uniform estimates in t.

We now describe the process of ensuring that the modification done in the previous steps stays continuous in t and has comparable Lipschitz-estimates in each dyadic set to the original extension. During the construction made in Section 7, we have created an extension $h : [0,1]^3 \to [0,1]^3$ of the boundary map φ so that each level $[0,1]^2 \times \{t\}$ is mapped to $\mathbb{R}^2 \times \{t\}$. For each t, such a level is divided into a number (depending on t) of dyadic squares whose boundaries are mapped to piecewise linear Jordan curves by h on the target side. Moreover, inside these squares the map h is defined by the shortest curve extension of its boundary values. For each dyadic level, there is a specific parameter t at which the construction changes from being based on those dyadic squares to being based on their children. The exact behaviour of h at this parameter was described in Step 5 of Section 7 at the parameter $t = mid$ in the cube $U_{k,j}$. We let the sequence of such parameters be denoted by $t_1 > t_2 > t_3 > \dots$ corresponding to each dyadic level.

We first describe how to modify the extension h inside each interval $I_j = (t_{j+1}, t_j]$ without paying mind to the continuity between successive intervals. We focus now on a fixed parameter t and a single dyadic square $\tilde{Q}_{k,j} \times \{t\}$ on the domain side and its target set, which we interpret as a planar Jordan domain $\mathbb{Y}_t$ with piecewise linear boundary. We may appeal to the fact that boundary of the domain $\mathbb{Y}_t$ deforms continuously in t and the fact that there is an upper bound on the number of vertices of each piecewise linear curve to deduce that the quantity $D = D(t)$ as defined earlier on $\mathbb{Y}_t$ has a uniform lower bound for $t \in I_j$. Here we recall that the quantity D and all other quantities introduced in the earlier description of the construction need to be interpreted as functions of t.

We now appeal to the behaviour of the piecewise linear curve $\partial \mathbb{Y}_t$. In a neighbourhood of parameters t where the number of vertices of $\partial \mathbb{Y}_t$ is constant, the domain $\mathbb{Y}_t$ changes in t only by moving these vertices around in a continuous way. There is hence a correspondence between the segments PV_P in t in this neighbourhood and thus to guarantee continuity of the modified extension we must simply ensure that the length of each such segment is a continuous function in t. This length of PV_P was defined as $\epsilon_P D/3$. Since D is locally bounded from below in t, ϵ_P can be chosen for each t in such a way as to make $\epsilon_P D$ a continuous function in t in such a neighbourhood. In fact, we choose $\epsilon_P D$ to be a piecewise linear function to maintain Lipschitz-continuity in t as well (we pay proper attention to estimates later).

We should pay some special attention here to shortest curves L_s which completely contain a segment PV_P. This happens only when the shortest curve $L_{\hat{s}_P}$ with endpoint

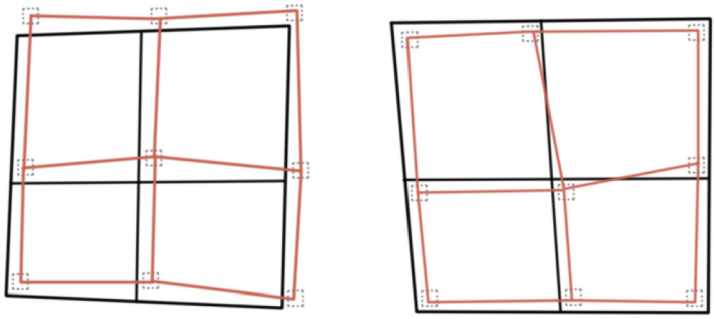
P bisects the angle of the boundary at P. In such a case no other curve $L_{s'}$ may pass through P as these curves have mutually disjoint endpoints, nor may it pass through PV_P as the shortest curves do not intersect in the interior. At any parameter t where this issue happens we may therefore set $\epsilon_P = 0$, essentially forgetting about the segment PV_P altogether, without losing injectivity of the modified extension at this parameter. To maintain the continuity of $\epsilon_P D$ in t near those parameters t for which $L_{\hat{s}_P}$ contains PV_P, we may for example take an already chosen function $\epsilon_P(t)$ and multiply it with a (piecewise linear) function $G(t)$ for which $G(t) \in [0,1)$ and $G(t) = 0$ exactly for these exceptional parameters t.

The number of vertices of $\partial \mathbb{Y}_t$ does not generally remain constant, as there may be new vertices appearing from an edge turning into two edges via a new angle being created at a given point P on that edge. The reverse may also happen to reduce the vertex count by one, but for the purposes of proving continuity both of these cases are symmetric to each other. Let us hence assume that at time T_0 the point $P = P(T_0)$ lies on an edge of $\partial \mathbb{Y}_{T_0}$, but on the interval (T_0, T_1) the point $P(t)$ is a true vertex of $\partial \mathbb{Y}_t$. In this case we do as before on (T_0, T_1), choosing $\epsilon_P D$ to be continuous in terms of t. Moreover, we choose ϵ_P in such a way that $\epsilon_P D \to 0$ as $t \to t_0$. This means that the segment PV_P shrinks to a point as $t \to T_0$, which guarantees continuity at this point also.

For a fixed parameter t, it is clear that as the numbers ϵ_P are chosen uniformly small enough, for example, by multiplying each with a small constant $\delta_t > 0$ independent of P, the modified extension H^*_φ is arbitrarily close to the original extension H_φ in the Lipschitz norm. Moreover as the quantities $\epsilon_P D$ were chosen to be Lipschitz continuous, choosing δ_t as a piecewise linear function in t with small enough Lipschitz norm guarantees that the map $(z,t) \to H^*_{\varphi_t}(z)$ may be chosen arbitrarily close to the original map h in the Lipschitz norm for $t \in (t_{k+1}, t_k]$. This shows that the Lipschitz estimates obtained in the previous section may be inherited by the modified extension as well.

Finally, we address the case of the parameters t_k where we switch from one dyadic level to another ($t = mid$ in $U_{k,j}$). We pick a parameter $t^*_k < t_k$ slightly below t_k so that on the level t^*_k the extension h is given by the shortest curve extension in the four dyadic children instead. Choosing t^*_k close enough to t_k lets us assume that the two maps levels t^*_k and t_k are arbitrarily close to each other in the Lipschitz norm. Moreover, due to this we may assume that the two modified maps are also as close in the Lipschitz norm as we want. For the sake of this argument we interpret these modified maps as planar maps $h_{t_k}, h_{t^*_k} : S \to \mathbb{Y}$ from a square to a piecewise Lipschitz Jordan domain, and recall that they have the same boundary values. As both of these maps are piecewise linear and homeomorphic, for t^*_k close enough to t_k we may assume that each of the maps $h^{(\tau)} := (1-\tau)h_{t_k} + \tau h_{t^*_k}$ is also homeomorphic for $\tau \in [0,1]$ due to the fact that the Jacobian determinant of $h^{(\tau)}$ must be bounded away from zero for all τ when h_{t_k} and $h_{t^*_k}$ are close enough in the Lipschitz norm.

We may then redefine the extension for parameters $t \in [t^*_k, t_k]$ by setting it equal to $h^{(\tau)}$ for $\tau = (t - t_k)/(t^*_k - t_k)$. Note that the Lipschitz norm in t may now be very large here due to the fact that the denominator $t^*_k - t_k$ may be arbitrarily small. To fix this,



The quadrilaterals on the side faces. The quadrilaterals on the top face.

Fig. 16. The construction of refined dyadic quadrilaterals on the side and top faces.

we rescale the parametrization on the interval $(t_{k+1}, t_k]$ on the domain and target side so that if M denotes the midpoint of this interval, we scale $(t_k^*, t_k]$ to $(M, t_k]$ and $(t_{k+1}, t_k^*]$ to $(t_{k+1}, M]$. The length of the interval $(M, t_k]$ is hence comparable to 2^{-k}, which means that the Lipschitz constant of the map for parameters $t \in (M, t_k]$ on $U_{k,j}$ is controlled by $2^k|\hat{\Gamma}_{k,j}|$ as we have wanted. This finishes the construction and the proof.

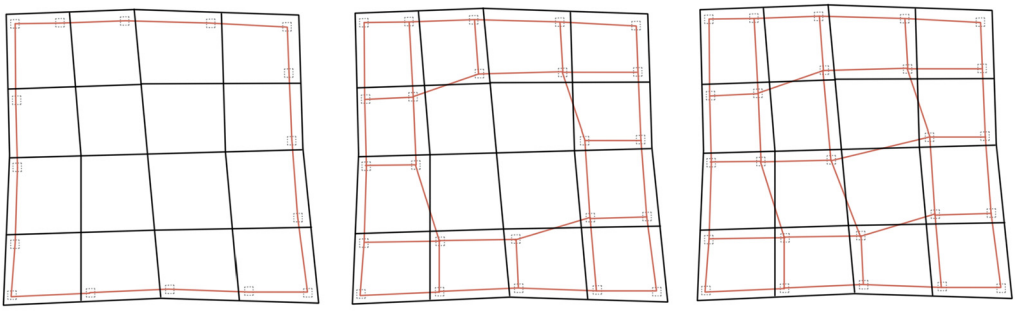
9. Extending a boundary map of the sphere

In this section we describe how to modify the local extension method constructed in Sections 5 to 8 to obtain a proof of Theorem 1.2. We go through the arguments in order and explain the changes needed in each part.

Proof of Theorem 1.2. First we must define a dyadic decomposition of the unit sphere. For this purpose we embed the boundary of the unit cube smoothly onto the sphere and inherit the dyadic decomposition from each face of the unit cube. Thus the dyadic decomposition of the sphere splits into six dyadic decompositions of squares, which correspond to the six faces of the unit cube, and we may label the respective sets on the sphere as four side faces and one top and bottom face.

The key difference in the spherical case lies in Lemma 5.1, where the dyadic decomposition is refined on each level. The main issue is that in Lemma 5.1 the vertices of the refined quadrilaterals $Q_{k,j}$ were positioned in the same direction (to the right and up) with respect to the original dyadic squares, whereas no such uniform direction can be chosen on the sphere. Instead we do as follows. For each dyadic quadrilateral $Q_{k,j}$ belonging to one of the side faces, we apply the same arguments as in Section 5 and choose the vertices of its four children in the direction of east and north on the sphere, see Fig. 16. Thus on the side faces the construction can proceed as usual.

We turn our attention to the top face. Let us fix a dyadic level k and suppose that the choice of quadrilaterals $Q_{k,j}$ has been made. Let us denote by $\{v_m\}$ the collection of points that are either vertices of the quadrilaterals $Q_{k,j}$, midpoints of their sides, or intersections of two segments between opposing midpoints. The points $\{\hat{v}_m\}$ will denote



Stage 1: Vertices $\hat{v}_m \in \hat{O}$ are fixed due to the construction on the side faces.

Stage 2: Picking neighbouring vertices.

Stage 3: The rest may be chosen arbitrarily.

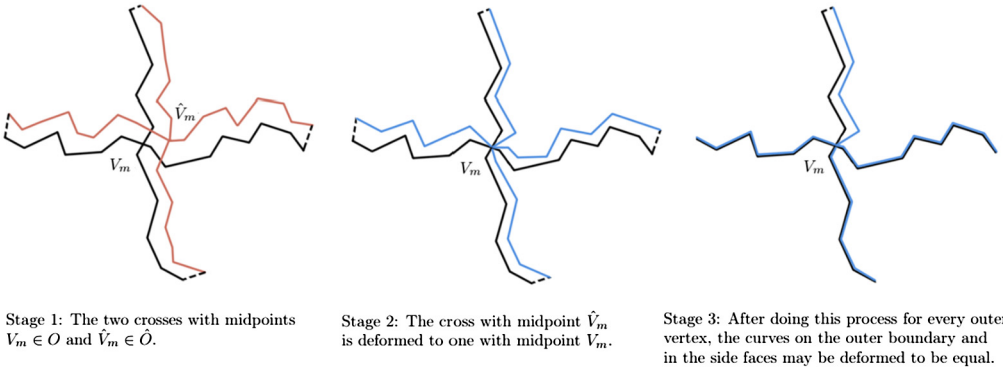
Fig. 17. Picking vertices $\hat{v}_m$ in three stages.

vertices of the quadrilaterals $Q_{k+1,j}$ which we must now choose. Let us define two sets of vertices O and $\hat{O}$ by saying that $v_m \in O$ if the vertex v_m is on the outer boundary of the union of all $Q_{k,j}$ on the top face, and likewise $\hat{v}_m \in \hat{O}$ if $\hat{v}_m$ is on the outer boundary of the union of all $Q_{k+1,j}$ on the top face. The choices made on the side faces already fix the points $\hat{v}_m \in \hat{O}$ and imply that on each dyadic level k, vertices $\hat{v}_m \in \hat{O}$ are closer to the north pole than the vertices $v_m \in O$, and thus belong inside the union of all $Q_{k,j}$, see Fig. 16. On the bottom face this relation is reversed, but these two cases are analogous enough that we only need to describe the construction on the top face and the other case is done with similar arguments.

For vertices $v_m \notin O$, we must pick one of four possible directions in which to choose $\hat{v}_m$ in, corresponding to the four dyadic quadrilaterals meeting at v_m. Supposing that $k \geqslant 2$, we pick the vertices as follows. For each vertex v_m for which v_m has a neighbour $v_{m'} \in O$, we choose $\hat{v}_m$ to lie inside the same quadrilateral as $\hat{v}_{m'}$, see Fig. 17. There are four vertices near the corners where the choice of $v_{m'}$ is not unique and thus we have two quadrilaterals to choose from: one in the corner and one adjacent to it. In this case we pick $\hat{v}_m$ in the quadrilateral adjacent to the corner. For vertices v_m not having neighbours in O, we can pick the direction in which to choose $\hat{v}_m$ arbitrarily.

As we have now chosen the grids on the domain side, we proceed as usual to define curves $\Gamma_{k,j}$ on the image side as piecewise linear approximations of the image curves of $\partial Q_{k,j}$ under φ. Topological information can be preserved here since φ is a homeomorphism, which means that we can assume that the image grid formed by the $\Gamma_{k,j}$ is topologically equivalent to the domain grid. Hence on each dyadic level k the grid formed by the $\Gamma_{k,j}$ and the grid on the next level formed by the children $\hat{\Gamma}_{k,j}$ can be assumed to have topologically the same intersection points as the respective grids on the domain side.

Due to the appearance of some additional intersection points compared to the arguments in Section 7, we must explain how the homotopy between $\Gamma_{k,j}$ and $\hat{\Gamma}_{k,j}$ is defined in our case. Denote by V_m and $\hat{V}_m$ the vertices on the image side corresponding to v_m and $\hat{v}_m$, and abuse notation to define $V_m \in O$ if $v_m \in O$. First we note that due to the



Stage 1: The two crosses with midpoints $V_m \in O$ and $\hat{V}_m \in \hat{O}$.

Stage 2: The cross with midpoint $\hat{V}_m$ is deformed to one with midpoint V_m.

Stage 3: After doing this process for every outer vertex, the curves on the outer boundary and in the side faces may be deformed to be equal.

Fig. 18. Deforming the two crosses.

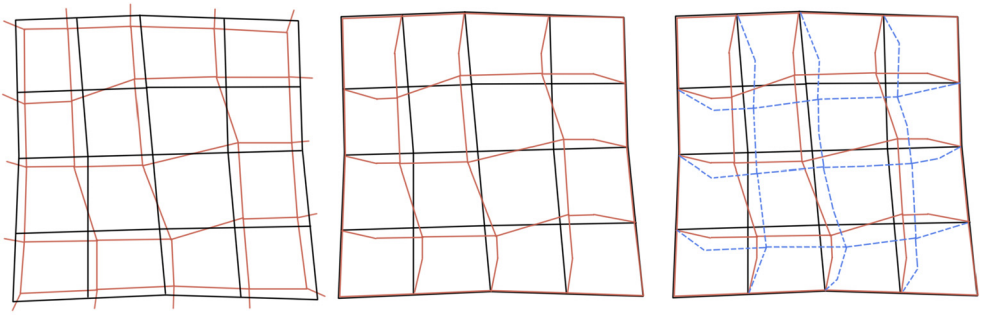
choice of the vertices $\hat{v}_m$ before, if $\hat{v}_m \in \hat{O}$ then at these points we are in the topologically correct situation to apply the homotopy construction from Section 7. As in the argument presented there, we may deform the cross with centre $\hat{V}_m$ into a cross with centre V_m and having the same endpoints, see Stage 2 in Fig. 18.

At the four vertices in the corners of the top face there is a special situation where only three curves meet at v_m and $\hat{v}_m$ instead of four, so technically we can not apply the previous homotopy argument between crosses here. But the "cross" consisting of three curves is only easier to deform than one with four. For example, one can add an auxiliary curve to both configurations, use the previous argument for four curves, and then forget about the auxiliary curves altogether.

Thus we may apply an initial homotopy at the points $\hat{V}_m \in \hat{O}$ and the side faces to replace the grid formed by the curves $\hat{\Gamma}_{k,j}$ with another grid $\hat{G}$ whose outer boundary curves and points align with the respective $\Gamma_{k,j}$ and V_m. See Stages 2 and 3 in Fig. 19. We must then describe how to deform the parts of the two grids left over inside the top face to each other despite the existence of some extra intersection points.

In order to do this we simply define an auxiliary grid with vertices at points we denote by W_m as follows. The points W_m will be chosen in the same direction with respect to both points $V_m \notin O$ and $\hat{V}_m \notin \hat{O}$. Precisely we mean that if the grid $\hat{G}$ is identified with a square grid of dimensions $2^k \times 2^k$, then each point W_m lies in the square to, say, the lower right of its respective point $\hat{V}_m \in G$. We may make this choice so that W_m also lies to the lower right with respect to V_m in the original grid G consisting of the curves $\Gamma_{k,j}$.

The points W_m can then be connected by piecewise linear Jordan curves with lengths comparable to the total length of the respective curves $\Gamma_{k,j}$ and $\hat{\Gamma}_{k,j}$. This may be justified for example by travelling sufficiently close to either of the given grids G and $\hat{G}$. These curves form an auxiliary grid $\tilde{G}$ containing the points W_m, and we can moreover pick this grid so that each of the curves in $\tilde{G}$ between neighbouring points W_m only intersects both grids G and $\hat{G}$ at most once.



Stage 1: The inital configuration.

Stage 2: The vertices $\hat{v}_m \in \hat{O}$ are moved to respective vertices $v_m \in O$.

Stage 3: Adding an auxiliary grid.

Fig. 19. Deformation between the two grids.

It then only remains to apply the arguments of Section 7 concerning the homotopy between grids to first deform G to $\tilde{G}$, and then $\tilde{G}$ to $\hat{G}$ as the grid $\tilde{G}$ is in the correct position w.r.t. the other two grids to apply the usual construction. The rest of the proof proceeds the same way. □

Data availability

No data was used for the research described in the article.

References

[1] L.V. Ahlfors, Extension of quasiconformal mappings from two to three dimensions, Proc. Natl. Acad. Sci. USA 51 (1964) 768–771.
[2] G. Alessandrini, M. Sigalotti, Geometric properties of solutions to the anisotropic p-Laplace equation in dimension two, Ann. Acad. Sci. Fenn., Math. 26 (2001) 249–266.
[3] S.S. Antman, Nonlinear Problems of Elasticity, Applied Mathematical Sciences, vol. 107, Springer-Verlag, New York, 1995.
[4] K. Astala, T. Iwaniec, G. Martin, Elliptic Partial Differential Equations and Quasiconformal Mappings in the Plane, Princeton University Press, 2009.
[5] K. Astala, T. Iwaniec, G.J. Martin, J. Onninen, Extremal mappings of finite distortion, Proc. Lond. Math. Soc. (3) 91 (3) (2005) 655–702.
[6] J.M. Ball, Convexity conditions and existence theorems in nonlinear elasticity, Arch. Ration. Mech. Anal. 63 (4) (1976/77) 337–403.
[7] A. Beurling, L. Ahlfors, The boundary correspondence under quasiconformal mappings, Acta Math. 96 (1956) 125–142.
[8] L. Carleson, The extension problem for quasiconformal mappings, in: Contributions to Analysis (a Collection of Papers Dedicated to Lipman Bers), Academic Press, New York, 1974, pp. 39–47.
[9] P.G. Ciarlet, Mathematical Elasticity vol. I. Three-Dimensional Elasticity, Studies in Mathematics and Its Applications, vol. 20, North-Holland Publishing Co., Amsterdam, 1988.
[10] M. Csörnyei, S. Hencl, J. Malý, Homeomorphisms in the Sobolev space $W^{1,n-1}$, J. Reine Angew. Math. 644 (2010) 221–235.
[11] P. Duren, Harmonic Mappings in the Plane, Cambridge University Press, Cambridge, 2004.
[12] S. Daneri, A. Pratelli, Smooth approximation of bi-Lipschitz orientation-preserving homeomorphisms, Ann. Inst. Henri Poincaré, Anal. Non Linéaire 31 (2014) 567–589.
[13] E. Gagliardo, Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili, Rend. Semin. Mat. Univ. Padova 27 (1957) 284–305.
[14] S. Hencl, Sobolev homeomorphism with zero Jacobian almost everywhere, J. Math. Pures Appl. (9) 95 (4) (2011) 444–458.

[15] S. Hencl, P. Koskela, Lectures on Mappings of Finite Distortion, Lecture Notes in Mathematics, vol. 2096, Springer, Cham, 2014.
[16] S. Hencl, P. Koskela, J. Malý, Regularity of the inverse of a Sobolev homeomorphism in space, Proc. R. Soc. Edinb., Sect. A 136 (6) (2006) 1267–1285.
[17] S. Hencl, P. Koskela, J. Onninen, A note on extremal mappings of finite distortion, Math. Res. Lett. 12 (2–3) (2005) 231–237.
[18] S. Hencl, A. Pratelli, Diffeomorphic approximation of $W^{1,1}$ planar Sobolev homeomorphisms, J. Eur. Math. Soc. 20 (3) (2018) 597–656.
[19] J. Hubbard, Teichmüller Theory and Applications to Geometry, Topology, and Dynamics, vol. 1, Matrix Editions, Ithaca, NY, 2006.
[20] T. Iwaniec, L.V. Kovalev, J. Onninen, Diffeomorphic approximation of Sobolev homeomorphisms, Arch. Ration. Mech. Anal. 201 (2011) 1047–1067.
[21] T. Iwaniec, G. Martin, Geometric Function Theory and Non-linear Analysis, Oxford Mathematical Monographs, Oxford University Press, 2001.
[22] A. Koski, J. Onninen, Sobolev homeomorphic extensions, J. Eur. Math. Soc. 23 (12) (2021) 4065–4089.
[23] R.S. Laugesen, Injectivity can fail for higher-dimensional harmonic extensions, Complex Var. Theory Appl. 28 (4) (1996) 357–369.
[24] J. Onninen, Regularity of the inverse of spatial mappings with finite distortion, Calc. Var. Partial Differ. Equ. 26 (3) (2006) 331–341.
[25] A. Pratelli, E. Radici, Approximation of planar BV homeomorphisms by diffeomorphisms, J. Funct. Anal. 276 (2019) 659–686.
[26] Y.G. Reshetnyak, Space Mappings with Bounded Distortion, American Mathematical Society, Providence, RI, 1989.
[27] W. Sickel, Sobolev Spaces of Fractional Order, Nemytskii Operators and Nonlinear Partial Differential Equations, de Gruyter, Berlin, 1996.
[28] W. Sickel, H. Triebel, Hölder inequalities and sharp embeddings in function spaces of B^s_{pq} and F^s_{pq} type, Z. Anal. Anwend. 14 (1) (1995) 105–140.
[29] D. Sullivan, Hyperbolic geometry and homeomorphisms, in: Geometric Topology, Proc. Georgia Topology Conf., Athens, Ga., 1977, Academic Press, New York-London, 1979, pp. 543–555.
[30] P. Tukia, J. Väisälä, Quasiconformal extension from dimension n to $n+1$, Ann. Math. (2) 115 (2) (1982) 331–348.
[31] G.C. Verchota, Harmonic homeomorphisms of the closed disc to itself need be in $W^{1,p}$, $p < 2$, but not $W^{1,2}$, Proc. Am. Math. Soc. 135 (3) (2007) 891–894.