SIAM J. IMAGING SCIENCES
Vol. 16, No. 1, pp. 501–567

A Nonlocal Graph-PDE and Higher-Order Geometric Integration for Image Labeling*

Dmitrij Sitenko†, Bastian Boll†, and Christoph Schnörr†

Abstract. This paper introduces a novel nonlocal partial difference equation (G-PDE) for labeling metric data on graphs. The G-PDE is derived as a nonlocal reparametrization of the assignment flow approach that was introduced in [*J. Math. Imaging Vision,* 58 (2017), pp. 211–238]. Due to this parameterization, solving the G-PDE numerically is shown to be equivalent to computing the Riemannian gradient flow with respect to a nonconvex potential. We devise an entropy-regularized difference of convex (DC) functions decomposition of this potential and show that the basic geometric Euler scheme for integrating the assignment flow is equivalent to solving the G-PDE by an established DC programming scheme. Moreover, the viewpoint of geometric integration reveals a basic way to exploit higher-order information of the vector field that drives the assignment flow, in order to devise a novel accelerated DC programming scheme. A detailed convergence analysis of both numerical schemes is provided and illustrated by numerical experiments.

Key words. assignment flows, image labeling, replicator equation, nonlocal graph-PDE, geometric integration, DC programming, information geometry

MSC codes. 34B45, 34C40, 62H35, 68U10, 68T05, 90C26, 91A22

DOI. 10.1137/22M1496141

1. Introduction.

1.1. Overview, motivation.

Nonlocal iterative operations for data processing on graphs constitute a basic operation that underlies many major image and data processing frameworks, including variational methods and PDEs on graphs for denoising, morphological processing, and other regularization-based methods of data analysis [1, 2, 3, 4, 5]. This includes deep networks [6] and time-discretized neural ODEs [7] whose layers generate sequences of nonlocal data transformations.

Among the extensions of such approaches to *data labeling* on graphs, that is, the assignment of an element of a finite set of labels to data points observed at each vertex, one may distinguish approaches whose mathematical structure is directly dictated by the labeling task

*Received by the editors May 13, 2022; accepted for publication (in revised form) November 15, 2022; published electronically March 30, 2023.
https://doi.org/10.1137/22M1496141
Funding: The work of the authors was supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy grant EXC-2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). The work of the second author was also supported by DFG grant SCHN 457/17-1, within the priority programme SPP 2298: "Theoretical Foundations of Deep Learning."

†Image and Pattern Analysis Group, Heidelberg University, Heidelberg, 69120 Germany (http://ipa.math.uni-heidelberg.de, dmitrij.sitenko@iwr.uni-heidelberg.de, bastian.boll@iwr.uni-heidelberg.de, schnoerr@math.uni-heidelberg.de).

and approaches that combine traditional data processing with a subsequent final discretization step:

- Examples of the former class are discrete graphical models [8, 9] that encode directly the combinatorial label assignment task, as a basis for the design of various sequential nonlocal processing steps performing approximate inference, like belief propagation. However, the intrinsic nonsmoothness of discrete graphical models constitutes a major obstacle for the design of hierarchical models and for efficient parameter learning. Graphical models, therefore, have been largely superseded by deep networks during the last decade.
- Examples of the latter class include the combination of established PDE-based diffusion approaches and threshold operations [10, 11, 12]. The mathematical formulations inherit the connection between total variation–based variational denoising, mean curvature motion, and level set evolution [13, 14, 15, 16], and they exhibit also connections to gradient flows in terms of the Allen–Cahn equation with respect to the Ginzburg–Landau functional [11, 15]. Regarding data labeling, however, a conceptual shortcoming of these approaches is that they do not provide a direct and natural mathematical problem formulation. As a consequence, this renders it difficult to cope with the assignment of dozens or hundreds of labels to data, and to efficiently learn parameters in order to tailor regularization properties to the problem and the class of data at hand.

Assignment flows [17, 18] constitute a mathematical approach tailored to the data labeling problem, aimed at overcoming the aforementioned shortcomings. The basic idea is to represent label assignments to data by a *smooth* dynamical process, based on the Fisher–Rao geometry of discrete probability distributions and on a weighted (parametrized) coupling of local flows for label selection across the graph. As a result, no extrinsic thresholding or rounding is required since the underlying geometry enables one to perform both spatial diffusion for assignment regularization and rounding to an integral solution just by integrating the assignment flow.

Stability and convergence to integral solutions of assignment flows hold under mild conditions [19]. A wide range of numerical schemes exist [20] for integrating geometrically assignment flows with GPU-conforming operations. Generalized assignment flows for unsupervised and self-supervised scenarios [21, 22] are more involved computationally but do not essentially change the overall *mathematical* structure.

Assignment flows *regularize* the assignment of labels to data by parameters Ω that couple the local flows at edges across the graph. These parameters can be determined either directly in a data-driven way as demonstrated in Figure 4 or learned offline in a supervised way. Learning the parameters of assignment flows from data can be accomplished using symplectic numerical integration [23] or, alternatively and quite efficiently, using exponential integration of linearized assignment flows [24, 25]. In particular, deep parametrizations of assignment flows do not at all change the mathematical structure, which enables one to exploit recent progress on PAC-Bayes bounds in order to compute a statistical performance certificate of classifications performed by deep linearized assignment flows in applications [26]. The assignment flow approach is introduced in section 2.2.

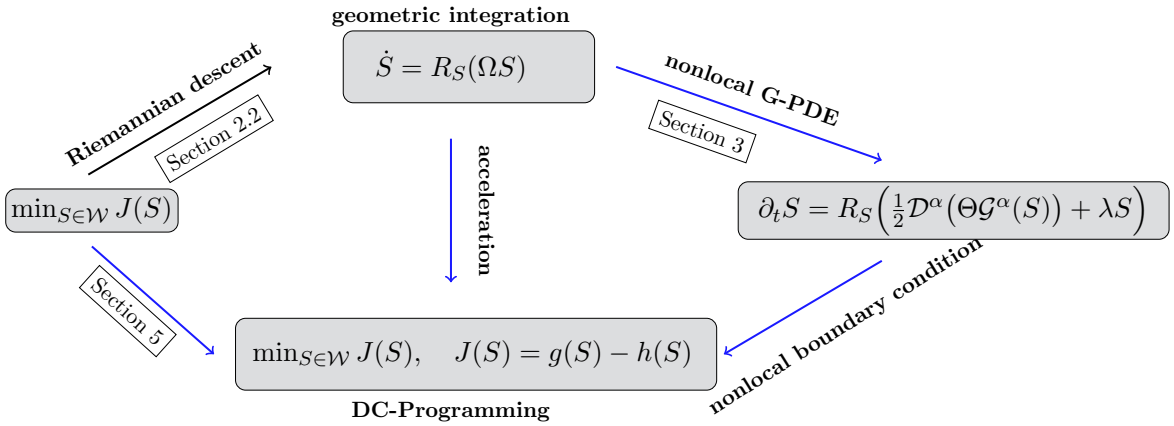


Figure 1. ***Summary of results.*** *The starting point (section 2.2) is a particular formulation of the assignment flow ODE (top) that represents the Riemannian gradient descent of a functional J (left). The first main contribution of this paper is an equivalent alternative representation of the assignment flow equation in terms of a PDE on the underlying graph (right), with a nonlocal data-driven diffusion term in divergence form and further terms induced by the information-geometric approach to the labeling problem. The second major contribution concerns a DC-decomposition of the nonconvex functional J (bottom) and a novel accelerated minimization algorithm using a second-order tangent space parametrization of the assignment flow.*

1.2. Contribution, organization. This paper makes two contributions, illustrated by Figure 1:

(a) Given an undirected weighted regular grid graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \Omega)$, we show that solving a particular parametrization of the assignment flow is equivalent to solving the *nonlocal nonlinear partial difference equation* (*G-PDE*) on the underlying graph $\mathcal{G}$,

$$\partial_t S(x,t) = R_{S(x,t)}\left(\frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(S)\right) + \lambda S\right)(x,t) \qquad \text{on } \mathcal{V}\times\mathbb{R}_+, \tag{1.1a}$$

$$S(x,t) = 0 \qquad \text{on } \mathcal{V}_{\mathcal{I}}^\alpha \times \mathbb{R}_+, \tag{1.1b}$$

$$S(x,0) = \overline{S}(x)(0) \qquad \text{on } \overline{\mathcal{V}}\times\mathbb{R}_+, \tag{1.1c}$$

where the vector field S takes values at $x \in \mathcal{V}$ in the relative interior of the probability simplex that is equipped with the Fisher–Rao metric. $\mathcal{D}^\alpha$ and $\mathcal{G}^\alpha$ are nonlocal divergence and gradient operators based on established calculus [27, 28]. The linear mapping $R_{S(x),t}$ is the inverse metric tensor corresponding to the Fisher–Rao metric, expressed in ambient coordinates.

The G-PDE (1.1) confirms and provides a generalized *nonlocal* formulation of a PDE that was heuristically derived by [29, section 4.4] in the continuous-domain setting. In particular, (1.1) addresses the data labeling problem *directly* without any further pre- or postprocessing step and thus contributes to the line of PDE-based research of image analysis initiated by Alvarez et al. [30] and Weickert [31].

(b) The particular parametrization of the assignment flow that we show in this paper to be equivalent to (1.1), constitutes a Riemannian gradient flow with respect to a

non-convex potential [29, section 3.2]. We consider a *difference of convex (DC)* function decomposition [32] of this potential and show

(i) that the simplest first-order geometric numerical scheme for integrating the assignment flow can be interpreted as basic two-step iterative method of DC programming [33];
(ii) that a corresponding tangent-space parametrization of the assignment flow and second-order derivatives of the tangent vector field can be employed to *accelerate* the basic DC iterative scheme.

Due to result (a), both schemes (i) and (ii) also solve the G-PDE (1.1). In addition, we point out that while a rich literature exists about accelerated *convex* optimization (see, e.g., [34, 35, 36] and references therein), methods for accelerating *nonconvex* iterative optimization schemes have been less explored.

Organization. Our paper is organized as follows. Section 2 introduces nonlocal calculus and the assignment flow, respectively. The equivalence of the assignment flow and the G-PDE (1.1) is derived in section 3, together with a tangent space parametrization as the basis for the development of iterative numerical solvers, and with a balance law that reveals how spatial diffusion interacts with label assignment by solving (1.1). Section 4 is devoted to explicitly working out common aspects and differences of (1.1) to related work:

– continuous-domain nonlocal diffusion [37],
– nonlocal variational approaches to image analysis [3], and
– nonlocal G-PDEs on graphs [2, 5].

As summarized by Figure 8 and Table 1, these approaches can be regarded as special cases from the mathematical viewpoint. They differ, however, regarding the scope and the class of problems to be solved: the approach (1.1) is only devoted to the data *labeling* problem, which explains its mathematical form. Finally, we show how our work extends the result of [29]. Section 5 details contribution (b) on DC programming from the viewpoint of geometric integration. The corresponding convergence analysis is provided in section 6. Numerical results that illustrate our findings are reported in section 7. We conclude in section 8.

2. Preliminaries.

This section contains basic material required in the remainder of this paper. A list of symbols and their meanings follows.

Symbol	Description
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \Omega)$	A graph with vertex set $\mathcal{V}$, edge set $\mathcal{E}$, and weights Ω.
$\mathcal{V}$	Set of vertices representing the discrete domain $\mathcal{V} \subset \mathbb{Z}^d$.
n	Total number $n = \lvert\mathcal{V}\rvert$ of nodes in the graph $\mathcal{G}$.
d	Dimension of the discrete domain associated with $\mathcal{V}$.
Ω	Weighted symmetric adjacency matrix of the graph $\mathcal{G}$.
$\mathcal{N}(x)$	Neighborhood of $x \in \mathcal{V}$ induced by Ω.
E	Subset of a Euclidean space.
$\mathcal{F}_{\mathcal{V}}, \mathcal{F}_{\mathcal{V},E}$	Space of one-point functions defined on $\mathcal{V}$, taking values in $\mathbb{R}$, resp., E.
$\mathcal{F}_{\mathcal{V}\times\mathcal{V}}, \mathcal{F}_{\mathcal{V}\times\mathcal{V},E}$	Space of two-point functions defined on $\mathcal{V} \times \mathcal{V}$, taking values in $\mathbb{R}$, resp., E.
$\alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$	Antisymmetric mapping that defines the interaction of nodes $x, y \in \mathbb{Z}^d$.
$\Theta \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$	Nonnegative scalar-valued symmetric mapping that parametrizes the introduced nonlocal diffusion process.

$\mathcal{V}_{\mathcal{I}}^{\alpha}$	Nonlocal interaction domain which represents the connectivity of nodes $x \in \mathcal{V}$ to nodes $y \in \mathbb{Z}^d \setminus \mathcal{V}$.
$\overline{\mathcal{V}}$	Extension of the discrete domain associated with $\mathcal{V}$ by the nodes in $\mathcal{V}_{\mathcal{I}}^{\alpha}$.
$\mathcal{D}^{\alpha}, \mathcal{G}^{\alpha}$	Nonlocal divergence and gradient operators parametrized by the mapping α.
$\mathcal{N}^{\alpha}$	Nonlocal interaction operator parametrized by the mapping α.
$\mathcal{L}_{\omega}$	Nonlocal Laplacian with weight function ω.
$\mathcal{X}_n$	Data on the graph $\mathcal{G}$ taking values in a metric space $\mathcal{X}$.
$X(x)$	Data point $X \in \mathcal{X}_n$ given at $x \in \mathcal{V}$.
$\mathcal{X}^*$	Set of labels $\{X_j^* : j \in \mathcal{J}\} \subset \mathcal{X}$.
c	Number of labels $c = \lvert\mathcal{J}\rvert$, one of which is uniquely assigned to each data point.
Δ_c	Probability simplex in $\mathbb{R}^c$ of dimension $c-1$.
$\mathcal{S}$	Relative interior of the probability simplex Δ_c, forming the factors of the product manifold $\mathcal{W}$.
T_0	Tangent space corresponding to $\mathcal{S}$.
$\mathcal{W}$, $\mathcal{T}_0$	Assignment manifold and the corresponding tangent space at the barycenter $\mathbb{1}_{\mathcal{W}}$.
$S, W \in \mathcal{W}$	Points on the assignment manifold taking values $S(x), W(x) \in \mathcal{S}$ at $x \in \mathcal{V}$.
$S^*, W^* \in \overline{\mathcal{W}} \setminus \mathcal{W}$	Integral vectors on the boundary of $\mathcal{W}$.
$V \in \mathcal{T}_0$	Points in the tangent space taking values $V(x) \in T_0$ at $x \in \mathcal{V}$.
Π_0	Orthogonal projection onto the tangent space $\mathcal{T}_0$.
R_S	Replicator map at $S \in \mathcal{W}$.
$\odot$	Hadamard product (componentwise multiplication)

2.1. Nonlocal calculus. Following [27], we collect some basic notions of nonlocal calculus which will be used throughout this paper. See [38] for a detailed exposition.

Let $(\mathcal{V}, \mathcal{E}, \Omega)$ be an undirected weighted regular grid graph with

$$n = |\mathcal{V}|, \qquad \mathcal{V} \subset \mathbb{Z}^d, \qquad 2 \leq d \in \mathbb{N}, \tag{2.1}$$

nodes, with edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ that has no self-loops, and with the weighted adjacency matrix Ω that satisfies

$$0 \leq \Omega(x,y) \leq 1, \qquad \Omega(x,y) = \Omega(y,x) \qquad \forall x, y \in \mathcal{V}. \tag{2.2}$$

Ω defines the neighborhoods

$$\mathcal{N}(x) := \{y \in \mathcal{V} \colon \Omega(x,y) > 0\}, \qquad x \in \mathcal{V}, \tag{2.3}$$

and serves as a function $\Omega \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ measuring the similarity of adjacent nodes.

We define the function spaces

$$\mathcal{F}_{\mathcal{V}} := \{f \colon \mathcal{V} \to \mathbb{R}\}, \qquad \mathcal{F}_{\mathcal{V}\times\mathcal{V}} := \{F \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}\}, \tag{2.4a}$$

$$\mathcal{F}_{\mathcal{V},E} := \{F \colon \mathcal{V} \to E\}, \qquad \mathcal{F}_{\mathcal{V}\times\mathcal{V},E} := \{F \colon \mathcal{V} \times \mathcal{V} \to E\}, \tag{2.4b}$$

where E denotes a (possibly improper) subset of a Euclidean space. The spaces $\mathcal{F}_{\mathcal{V}}$ and $\mathcal{F}_{\mathcal{V}\times\mathcal{V}}$, respectively, are equipped with the inner products

$$\langle f, g \rangle_{\mathcal{V}} := \sum_{x \in \mathcal{V}} f(x) g(x), \qquad \langle F, G \rangle_{\mathcal{V}\times\mathcal{V}} := \sum_{(x,y) \in \mathcal{V}\times\mathcal{V}} F(x,y) G(x,y). \tag{2.5}$$

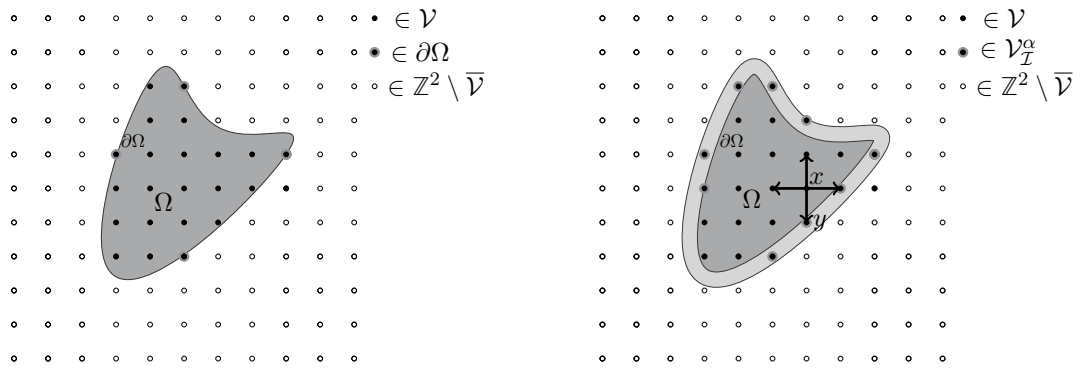


Figure 2. ***Schematic visualization of a nonlocal boundary.*** *Left: A bounded open domain $\Omega \subset \mathbb{R}^2$ with local boundary $\partial\Omega$ overlaid by the grid $\mathbb{Z}^2$. Right: A bounded open domain Ω with nonlocal boundary (light gray). Nodes $\bullet$ and $\bullet$, respectively, are vertices on the graph $\mathcal{V}$ and on the interaction domain $\mathcal{V}_{\mathcal{I}}^{\alpha}$ given by (2.8).*

We set

$$\overline{\mathcal{V}} := \mathcal{V} \dot{\cup} \mathcal{V}_{\mathcal{I}}^{\alpha} \qquad \text{(disjoint union)}, \tag{2.6}$$

where the *nonlocal interaction domain* $\mathcal{V}_{\mathcal{I}}^{\alpha}$ *with respect to* an *antisymmetric* mapping

$$\alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \qquad \alpha(x,y) = -\alpha(y,x) \qquad \forall x, y \in \overline{\mathcal{V}} \tag{2.7}$$

is defined as

$$\mathcal{V}_{\mathcal{I}}^{\alpha} := \{x \in \mathbb{Z}^d \setminus \mathcal{V} : \alpha(x,y) \neq 0 \text{ for some } y \in \mathcal{V}\}. \tag{2.8}$$

$\mathcal{V}_{\mathcal{I}}^{\alpha}$ serves discrete formulations of conditions on nonlocal boundaries with positive measure in a Euclidean domain. Such conditions are distinct from traditional conditions imposed on boundaries that have measure zero. Figure 2 displays a possible nonlocal boundary configuration.

We state the following identity induced by (2.7):

$$\sum_{x,y\in\overline{\mathcal{V}}} (F(x,y)\alpha(x,y) - F(y,x)\alpha(y,x)) = 0 \qquad \forall F \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}. \tag{2.9}$$

The *nonlocal divergence operator* $\mathcal{D}^{\alpha}$ and the *nonlocal interaction operator* $\mathcal{N}^{\alpha}$ are defined by

$$\mathcal{D}^{\alpha} : \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}} \to \mathcal{F}_{\overline{\mathcal{V}}}, \quad \mathcal{D}^{\alpha}(F)(x) := \sum_{y\in\overline{\mathcal{V}}} (F(x,y)\alpha(x,y) - F(y,x)\alpha(y,x)), \qquad x \in \overline{\mathcal{V}}, \tag{2.10a}$$

$$\mathcal{N}^{\alpha} : \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}} \to \mathcal{F}_{\mathcal{V}_{\mathcal{I}}^{\alpha}}, \quad \mathcal{N}^{\alpha}(F)(x) := -\sum_{y\in\overline{\mathcal{V}}} (F(x,y)\alpha(x,y) - F(y,x)\alpha(y,x)), \quad x \in \mathcal{V}_{\mathcal{I}}^{\alpha}. \tag{2.10b}$$

Based on the mapping α given by (2.7), the operator (2.10b) is nonzero in general and accounts for the density of a *nonlocal flux* from the entire domain $\overline{\mathcal{V}}$ to nodes $x \in \mathcal{V}_{\mathcal{I}}^{\alpha}$ [38].

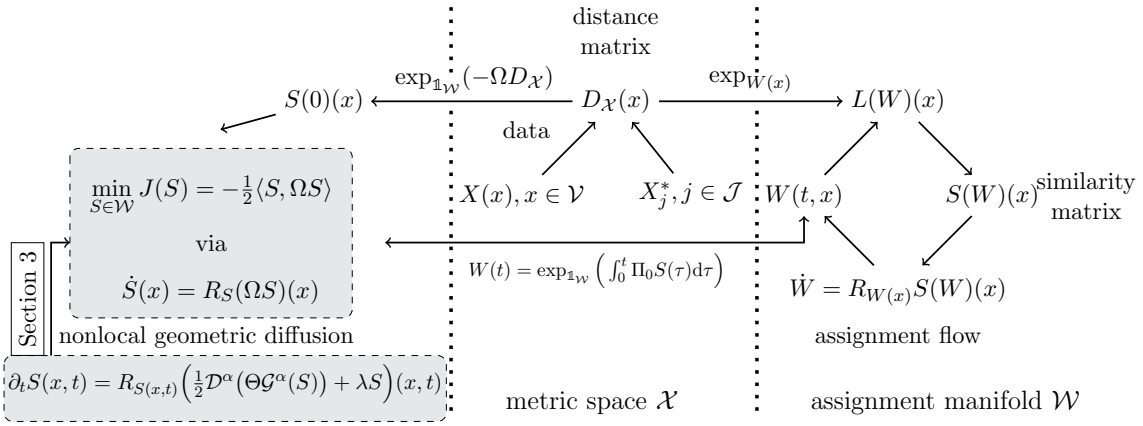


Figure 3. ***Inference of label assignments via assignment flows.*** *Center column: Application task of assigning data to prototypes in a metric space. Right column: Overview of the geometric approach* [17]. *The data are represented by the distance matrix $D_\mathcal{X}$ and by the likelihood vector field $L(W)$ on the assignment manifold $\mathcal{W}$. The similarity vectors $S(W)(x)$, determined through geometric averaging of the likelihood vectors, drive the assignment flow whose numerical geometric integration results in spatially coherent and unique label assignment to the data. Left column: Alternative equivalent reformulation of the assignment flow* [29] *which separates* (i) *the influence of the data that only determine the initial point of the flow (cf.* (2.38a)*), and* (ii) *the influence of the parameters Ω that parametrize the vector field which drives the assignment flow. This enables us to derive the novel nonlocal geometric diffusion equation in section* 3.

This generalizes the notion *local flux* density $\langle q(x), n(x)\rangle$ on continuous domains $\Omega\subset\mathbb{R}^d$ with outer normal vector field $n(x)\in\mathbb{R}^d$ on the boundary $\partial\Omega$, and with a vector-valued function $q(x)$ on $\partial\Omega$ that typically stems from an underlying constitutive physical relation. Due to the identity (2.9), the operators (2.10) satisfy the *nonlocal Gauss theorem*

$$\sum_{x\in\mathcal{V}} \mathcal{D}^\alpha(F)(x) = \sum_{y\in\mathcal{V}_\mathcal{I}^\alpha} \mathcal{N}^\alpha(F)(y). \tag{2.11}$$

The operator $\mathcal{D}^\alpha$ maps two-point functions $F(x,y)$ to $\mathcal{D}^\alpha(F)\in\mathcal{F}_{\overline{\mathcal{V}}}$, whereas $\mathcal{N}^\alpha(F)$ is defined on the domain $\mathcal{V}_\mathcal{I}^\alpha$ given by (2.8) where nonlocal boundary conditions are imposed.

The adjoint mapping $(\mathcal{D}^\alpha)^*$ with respect to the inner product (2.5) is determined by the relation

$$\langle f, \mathcal{D}^\alpha(F)\rangle_{\overline{\mathcal{V}}} = \langle (\mathcal{D}^\alpha)^*(f), F\rangle_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}} \qquad \forall f\in\mathcal{F}_{\overline{\mathcal{V}}}, \qquad \forall F\in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \tag{2.12}$$

which yields the operator

$$(\mathcal{D}^\alpha)^*\colon \mathcal{F}_{\overline{\mathcal{V}}}\to\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \qquad (\mathcal{D}^\alpha)^*(f)(x,y) := -(f(y)-f(x))\alpha(x,y) \qquad \forall f\in\mathcal{F}_{\overline{\mathcal{V}}}. \tag{2.13}$$

The *nonlocal gradient operator* is defined as

$$\mathcal{G}^\alpha\colon \mathcal{F}_{\overline{\mathcal{V}}}\to\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \qquad \mathcal{G}^\alpha(f)(x,y) := -(\mathcal{D}^\alpha)^*(f)(x,y) \qquad \forall f\in\mathcal{F}_{\overline{\mathcal{V}}}. \tag{2.14}$$

For *vector-valued* mappings, the operators (2.10) and (2.13) naturally extend to $\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}},E}$ and $\mathcal{F}_{\overline{\mathcal{V}},E}$, respectively, by acting *componentwise*.

Using the mappings (2.13), (2.14), the nonlocal Gauss theorem (2.11) implies *Green's nonlocal first identity*

(2.15)
$$\sum_{x\in\mathcal{V}} u(x)\mathcal{D}^\alpha(F)(x) - \sum_{x\in\overline{\mathcal{V}}}\sum_{y\in\overline{\mathcal{V}}} \mathcal{G}^\alpha(u)(x,y)F(x,y) = \sum_{x\in\mathcal{V}_{\mathcal{I}}^\alpha} u(x)\mathcal{N}^\alpha(F)(x), \qquad \begin{matrix} u & \in\mathcal{F}_{\overline{\mathcal{V}}}, \\ F & \in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}. \end{matrix}$$

Given a function $f\in\mathcal{F}_{\overline{\mathcal{V}}}$ and a symmetric mapping

$$\Theta\in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}} \quad \text{with} \quad \Theta(x,y)=\Theta(y,x), \tag{2.16}$$

we define the *linear nonlocal diffusion operator*

$$\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(f)\right)(x) = 2\sum_{y\in\overline{\mathcal{V}}} \mathcal{G}^\alpha(f)(x,y)\Theta(x,y)\alpha(x,y), \qquad f\in\mathcal{F}_{\overline{\mathcal{V}}}. \tag{2.17}$$

For the particular case with no interactions, i.e., $\alpha(x,y)=0$ if $x\in\mathcal{V}$ and $y\in\mathcal{V}_{\mathcal{I}}^\alpha$, expression (2.17) reduces with $\Theta(x,y)=1, x,y\in\mathcal{V}$ to

$$\mathcal{L}_\omega f(x) \overset{(2.13)}{=} \sum_{y\in\mathcal{N}(x)} \omega(x,y)\left(f(y)-f(x)\right), \qquad \omega(x,y)=2\alpha(x,y)^2, \tag{2.18}$$

which coincides with the *combinatorial Laplacian* [39, 40] after reversing the sign.

The next remark provides an intuition for appropriate setup of parameters $\alpha,\Theta\in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$.

Remark 2.1 (role of parameters in modeling nonlocal diffusion processes). In our work we differentiate the parameters α,Θ by their role in modeling nonlocal diffusion processes of the form (2.17). More precisely, we use the antisymmetric mapping $\alpha\in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ for definition of first-order derivative operators $\mathcal{D}^\alpha,\mathcal{G}^\alpha,\mathcal{N}^\alpha$ and the symmetric mapping $\Theta\in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ for specifying the constitutive function at each $x\in\mathcal{V}$ that controls the smoothing properties of operator (2.18). Instances of α,Θ along with an analytical ablation study will be presented in section 4.

2.2. The assignment flow approach. We summarize the assignment flow approach introduced by [17] and refer to [18] for more background and a review of related work. Figure 3 illustrates on the left the assignment flow approach and on the right its extension presented in section 3 (left panel).

2.2.1. Assignment manifold. Let $(\mathcal{X},d_{\mathcal{X}})$ be a metric space and

$$\mathcal{X}_n=\{X(x)\in\mathcal{X}\colon x\in\mathcal{V}\} \tag{2.19}$$

be given data on a graph $(\mathcal{V},\mathcal{E},\Omega)$ as specified in section 2.1. We encode assignments of data $X(x)$, $x\in\mathcal{V}$, to a set

$$\mathcal{X}^*=\{X_j^*\in\mathcal{X}, j\in\mathcal{J}\}, \qquad c:=|\mathcal{J}|, \tag{2.20}$$

of predefined prototypes by *assignment vectors*

$$W(x)=(W_1(x),\ldots,W_c(x))^\top\in\mathcal{S}, \tag{2.21}$$

where $\mathcal{S} = \operatorname{rint}\Delta_c$ denotes the relative interior of the probability simplex $\Delta_c \subset \mathbb{R}^c_+$ that we turn into a Riemannian manifold $(\mathcal{S}, g)$ with the Fisher–Rao metric g from information geometry [41, 42] at each $p \in \mathcal{S}$,

$$g_p(u,v) = \sum_{j\in\mathcal{J}} \frac{u_j v_j}{p_j} = \langle u, v\rangle_p, \qquad u, v \in T_0, \tag{2.22}$$

with tangent space T_0 given by (2.24). The *assignment manifold* $(\mathcal{W}, g)$ is defined as the product space $\mathcal{W} = \mathcal{S} \times \cdots \times \mathcal{S}$ of $n = |\mathcal{V}|$ such manifolds. Points on the assignment manifold row-stochastic matrices with full support are denoted by

$$W = (\ldots, W(x), \ldots)^\top \in \mathcal{W} \subset \mathbb{R}^{n\times c}_{++}, \qquad x \in \mathcal{V}. \tag{2.23}$$

The assignment manifold has the trivial tangent bundle $T\mathcal{W}$ with $T_W\mathcal{W} = \mathcal{T}_0 \, \forall W \in \mathcal{W}$ and tangent space

$$\mathcal{T}_0 = T_0 \times \cdots \times T_0, \qquad T_0 = \{v \in \mathbb{R}^c \colon \langle \mathbb{1}_c, v\rangle = 0\}. \tag{2.24}$$

The metric (2.22) naturally extends to

$$g_W(U,V) = \sum_{x\in\mathcal{V}} g_{W(x)}\left(V(x), U(x)\right), \qquad U, V \in \mathcal{T}_0. \tag{2.25}$$

The orthogonal projection onto T_0 is given by

$$\Pi_0 \colon \mathbb{R}^c \to T_0, \qquad \Pi_0(u) = u - \langle \mathbb{1}_{\mathcal{S}}, u\rangle \mathbb{1}_c, \qquad \mathbb{1}_{\mathcal{S}} := \frac{1}{c}\mathbb{1}_c. \tag{2.26}$$

The orthogonal projection onto $\mathcal{T}_0$, also denoted by Π_0 for simplicity, is

$$\Pi_0 \colon \mathbb{R}^{n\times c} \to \mathcal{T}_0, \qquad \Pi_0 D = (\ldots, \Pi_0 D(x), \ldots)^\top. \tag{2.27}$$

2.2.2. Assignment flows. Based on the given data and prototypes, we define the distance vector field on $\mathcal{V}$ by

$$D_{\mathcal{X}}(x) = (d_{\mathcal{X}}(X(x), X_1^*), \ldots, d_{\mathcal{X}}(X(x), X_c^*))^\top, \qquad x \in \mathcal{V}. \tag{2.28}$$

This data representation is lifted to $\mathcal{W}$ to obtain the *likelihood vectors*

$$L(x) \colon \mathcal{S} \to \mathcal{S}, \qquad L(W)(x) = \frac{W(x) \odot e^{-\frac{1}{\rho} D_{\mathcal{X}}(x)}}{\langle W(x), e^{-\frac{1}{\rho} D_{\mathcal{X}}(x)}\rangle}, \qquad x \in \mathcal{V}, \quad \rho > 0, \tag{2.29}$$

where the exponential function applies componentwise and $\odot$ denotes the componentwise multiplication

$$(p \odot q)_j = p_j q_j, \quad j \in [c], \qquad p, q \in \mathcal{S}, \tag{2.30}$$

of vectors p, q. Accordingly, we denote componentwise division of vectors by

$$\frac{v}{p} = \left(\frac{v_1}{p_1}, \ldots, \frac{v_c}{p_c}\right)^\top, \qquad p \in \mathcal{S}, \tag{2.31}$$

for strictly positive vectors p.

The map (2.29) is based on the affine e-connection of information geometry [41, 42]. The scaling parameter $\rho > 0$ normalizes the a priori unknown scale of the components of $D_{\mathcal{X}}(x)$. Likelihood vectors are spatially regularized by the *similarity map* and the *similarity vectors*, respectively, given for each $x \in \mathcal{V}$ by

$$S(x)\colon \mathcal{W} \to \mathcal{S}, \qquad S(W)(x) = \operatorname{Exp}_{W(x)}\left(\sum_{y \in \mathcal{N}(x)} \Omega(x,y) \operatorname{Exp}_{W(x)}^{-1}\left(L(W)(y)\right) \right), \tag{2.32}$$

where

$$\operatorname{Exp}\colon \mathcal{S} \times T_0 \to \mathcal{S}, \qquad \operatorname{Exp}_p(v) = \frac{p \odot e^{\frac{v}{p}}}{\langle p, e^{\frac{v}{p}} \rangle}, \qquad \frac{v}{p} = \left(\frac{v_1}{p_1}, \ldots, \frac{v_c}{p_c} \right)^\top \tag{2.33}$$

is the exponential map corresponding to the e-connection. If the exponential map of the Riemannian (Levi-Civita) connection were used instead, then the term in parentheses in (2.32) would be the optimality condition for the weighted Riemannian mean of the vectors $\{L(W)(y)\colon y \in \mathcal{N}(x)\}$ [43, Lemma 6.9.4]. Using the exponential map of the e-connection enables one to evalute the right-hand side of (2.32) in closed form and to define the similarity vectors as geometric means of the likelihood vectors [18].

The weights $\Omega(x,y)$ determine the regularization properties of the similarity map; cf. Remark 2.2 below. They satisfy (2.2) and the additional constraint

$$\sum_{y \in \mathcal{N}(x)} \Omega(x,y) = 1 \quad \forall x \in \mathcal{V}. \tag{2.34}$$

The *assignment flow* is induced on the assignment manifold $\mathcal{W}$ by solutions $W(t,x) = W(x)(t)$ of the system of nonlinear ODEs

$$\dot{W}(x) = R_{W(x)} S(W)(x), \qquad W(0,x) = W(x)(0) \in \mathbb{1}_{\mathcal{S}}, \quad x \in \mathcal{V}, \tag{2.35}$$

where the map

$$R_p = \operatorname{Diag}(p) - pp^\top, \qquad p \in \mathcal{S}, \tag{2.36}$$

corresponds to the inverse metric tensor expressed in the embedding coordinates of the ambient Euclidean space $\mathbb{R}^c$, which turns the right-hand side into the tangent vector field

$$\mathcal{V} \ni x \mapsto R_{W(x)} S(W)(x) = \operatorname{Diag}\left(W(x)\right) S(W)(x) - \langle W(x), S(W)(x) \rangle W(x) \in T_0. \tag{2.37}$$

Integrating the system (2.35) numerically [20] yields integral assignment vectors $W(t,x)$, $x \in \mathcal{V}$, for $t \to \infty$, that uniquely assign a label from the set $\mathcal{X}^*$ to each data point $X(x)$ [19].

Remark 2.2(regularization). From the viewpoint of variational imaging, *regularization* of the assignment flow has to be understood in a broad sense: The parameters Ω define by (2.32), at each location x and locally within neighborhoods $\mathcal{N}(x)$, what similarity of the collection of likelihood vectors $L(W)(y)$, $y \in \mathcal{N}(x)$, which represent the input data, really means in terms

of a corresponding geometric average, called similarity vector $S(W)(x)$. Unlike traditional variational approaches where regularization affects the primary variables directly, regularization of the assignment flow is accomplished more effectively by affecting *velocities* that *generate* the primary assignment variables: the vector field $S(W)$ drives the assignment flow (2.35). Figure 4 illustrates two applications of the assignment flow approach using data-driven nonlocal regularization. Learning the regularization parameters Ω from data was studied by [23, 25].

2.2.3. S-flow parametrization. We adopt from [29, Proposition 3.6] the *S-parametrization* of the assignment flow system (2.35)

$$\dot{S} = R_S(\Omega S), \qquad S(0) = \exp_{\mathbb{1}_\mathcal{W}}(-\Omega D_\mathcal{X}), \tag{2.38a}$$

$$\dot{W} = R_W(S), \qquad W(0) = \mathbb{1}_\mathcal{W}, \quad \mathbb{1}_\mathcal{W}(x) = \mathbb{1}_\mathcal{S}, \quad x \in \mathcal{V}, \tag{2.38b}$$

where both S and W are points on $\mathcal{W}$ and hence have the format (2.23) and

$$R_S(\Omega S)(x) = R_{S(x)}\left((\Omega S)(x)\right), \quad (\Omega S)(x) = \sum_{y \in \mathcal{N}(x)} \Omega(x,y) S(y), \tag{2.39}$$

$$\exp_{\mathbb{1}_\mathcal{W}}(-\Omega D_\mathcal{X}) := \left(\ldots, \mathrm{Exp}_{\mathbb{1}_\mathcal{S}} \circ R_{\mathbb{1}_\mathcal{S}}(-(\Omega D_\mathcal{X})(x)), \ldots\right)^\top \in \mathcal{W}, \quad x \in \mathcal{V}, \tag{2.40}$$

with the mappings Exp_p, R_p, $p \in \mathcal{S}$ defined by (2.33) and (2.36), respectively. In view of (2.40), we define the *lifting map*

$$\exp_p \colon T_0 \to \mathcal{S}, \qquad \exp_p(v) := \mathrm{Exp}_p \circ R_p v = \frac{p \odot e^v}{\langle p, e^v \rangle}, \qquad p \in \mathcal{S},\ v \in T_0, \tag{2.41}$$

which satisfies

$$\exp_{\exp_p(v)}(v') = \exp_p(v + v'), \qquad p \in \mathcal{S}, \quad v, v' \in T_0. \tag{2.42a}$$

In addition, one has (cf. (2.24), (2.26))

$$\exp_p(d) = \exp_p(\Pi_0 d) \quad \forall d \in \mathbb{R}^c. \tag{2.42b}$$

Analogous to (2.40), the lifting map (2.41) extends to

$$\exp_S \colon \mathcal{T}_0 \to \mathcal{W}, \qquad \exp_S(V) = \left(\ldots, \exp_{S(x)}(V(x)), \ldots\right) \tag{2.43a}$$

and the relations (2.42) extend to

$$\exp_{\exp_S(V)}(V') = \exp_S(V + V'), \qquad S \in \mathcal{W}, \quad V, V' \in \mathcal{T}_0, \tag{2.44a}$$

$$\exp_S(D) = \exp_S(\Pi_0 D) \qquad \forall D \in \mathbb{R}^{n \times c}. \tag{2.44b}$$

Parametrization (2.38) has the advantage that $W(t)$ depends on $S(t)$, but not vice versa. As a consequence, it suffices to focus on (2.38a) since its solution $S(t)$ determines the solution to (2.38b) by [19, Proposition 2.1.3],

$$W(t) = \exp_{\mathbb{1}_\mathcal{W}}\left(\int_0^t \Pi_0 S(\tau) \mathrm{d}\tau\right). \tag{2.45}$$

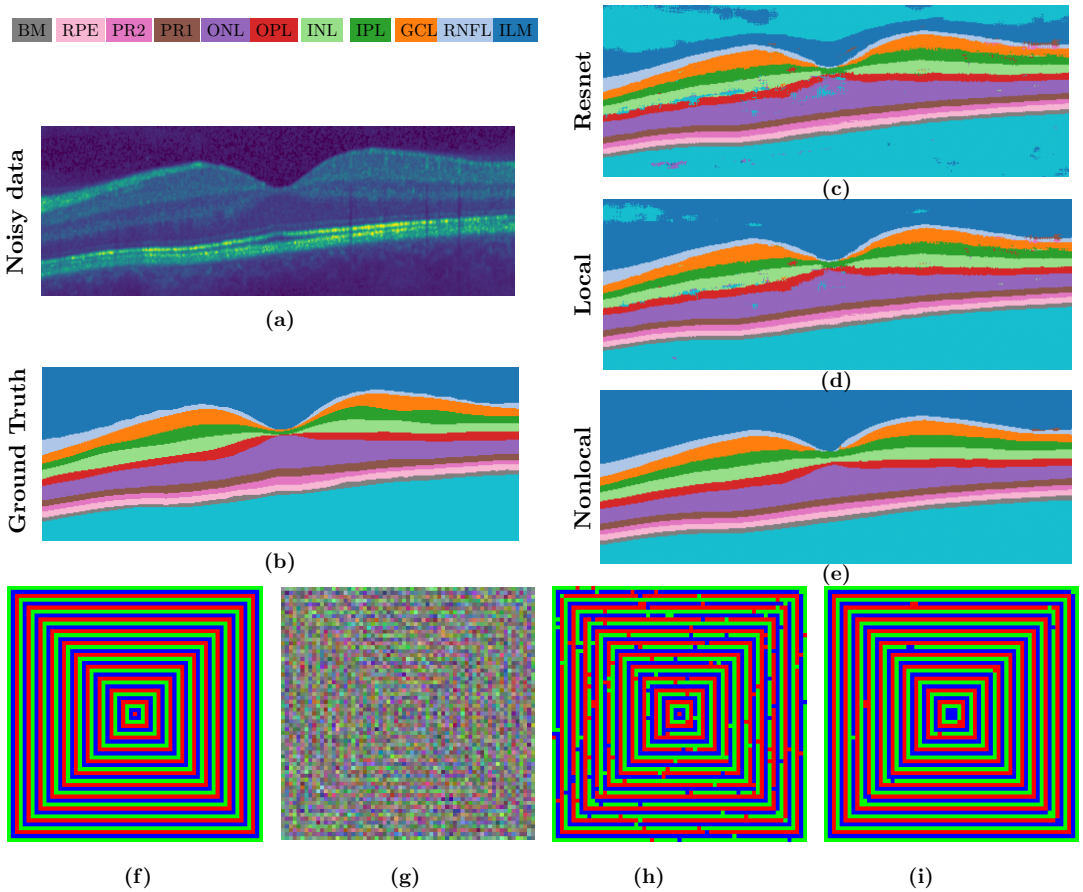


Figure 4. *Two image labeling scenarios demonstrating the influence of nonlocal regularization. Top: Application of assignment flows to a three-dimensional (3D) medical imaging problem for segmenting the human retina (see [44] for a detailed exposition).* (a) *A B-scan from a 3D OCT-volume showing a section of the human retina that is corrupted by speckle noise.* (b) *The corresponding ground truth labeling with ordered retina layers.* (c) *Output from a Resnet that serves as the distance matrix (2.28).* (d) *Result of applying assignment flow with local neighborhoods given by a 3D seven-point stencil.* (e) *Labeling obtained with nonlocal uniform neighborhoods of size* $|\mathcal{N}| = 11 \times 11 \times 11$. *Increasing the connectivity leads to more accurate labeling that satisfies the ordering constraint depicted in* (b). *Bottom: Labeling of noisy data by assignment flows with data-driven parameters* Ω *determined by nonlocal means [4] using patches of size* 7×7 *pixels.* (f) *Synthetic image with thin repetitive structure.* (g) *Severely corrupted input image to be labeled with* $\mathcal{X}^* = \{$●, ●, ●$\}$. (h), (i) *Labeling by the assignment flow that was regularized with neighborhood sizes* $|\mathcal{N}| = 3 \times 3$ *and* $|\mathcal{N}| = 11 \times 11$, *respectively. Enlarging the neighborhood size* $|\mathcal{N}|$ *increases labeling accuracy.*

In addition, (2.38a) was shown in [29] to be the *Riemannian gradient descent flow* with respect to the potential

$$J\colon \mathcal{W} \to \mathbb{R}, \qquad J(S) = -\frac{1}{2}\langle S, \Omega S\rangle = \frac{1}{4}\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{N}(x)} \Omega(x,y)\|S(x) - S(y)\|^2 - \frac{1}{2}\|S\|_F^2, \tag{2.46}$$

where $\|\cdot\|_F$ denotes the Frobenius (matrix) norm and the vector field $\mathcal{V} \ni x \mapsto S(x) \in \mathcal{S}$ is identified with the matrix

$$S = (S_j(x))_{x\in\mathcal{V},\, j\in[c]} \in \mathbb{R}^{n\times c}_{++} \tag{2.47}$$

such that (2.39) can be written as

$$((\Omega S)(x))_j = \sum_{y\in\mathcal{N}(x)} (\Omega(x,y)S(y))_j = \sum_{y\in\mathcal{N}(x)} \Omega(x,y)S(y,j) = (\Omega S)_{x,j}. \tag{2.48}$$

Convergence and stability results for the gradient flow (2.38a) have been established by [19].

3. Nonlocal graph-PDE.

In this section, we show that the assignment flow corresponds to a particular nonlocal diffusion process. This results in an equivalent formulation of the Riemannian gradient flow (2.38a) in terms of a suitable *nonlinear* extension of the nonlocal *linear* diffusion operator (2.17).

3.1. S-flow: Nonlocal PDE formulation.

We start with specifying a general class of parameter matrices Ω satisfying (2.2) and (2.34) in terms of antisymmetric and symmetric mappings $\alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ and $\Theta \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$, respectively.

Lemma 3.1. *Let*

$$\begin{array}{lll} \alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, & \alpha(y,x) = -\alpha(x,y), & \forall x,y \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \\ \Theta \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, & \Theta(x,y) = \Theta(y,x) \geq 0, & \forall x,y \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \end{array} \tag{3.1}$$

be antisymmetric and nonnegative symmetric mappings, respectively. Assume further that α satisfies

$$\alpha(x,y) = 0 \qquad \forall x,y \in \mathcal{V}_{\mathcal{I}}^{\alpha}. \tag{3.2}$$

Then, for neighborhoods $\mathcal{N}(x)$ defined by (2.3) and with parameter matrix

$$\Omega(x,y) = \begin{cases} \Theta(x,y)\alpha^2(x,y) & \text{if } x \neq y, \\ \Theta(x,x) & \text{if } x = y, \end{cases} \qquad x,y \in \mathcal{V}, \tag{3.3}$$

for each function $f \in \mathcal{F}_{\overline{\mathcal{V}}}$ with $f|_{\mathcal{V}_{\mathcal{I}}^{\alpha}} = 0$, the identity

$$\sum_{y\in\mathcal{V}} \Omega(x,y)f(y) = \frac{1}{2}\mathcal{D}^{\alpha}\left(\Theta\mathcal{G}^{\alpha}(f)\right)(x) + \lambda(x)f(x) \qquad \forall x \in \mathcal{V}, \quad \forall f \in \mathcal{F}_{\overline{\mathcal{V}}}\colon f|_{\mathcal{V}_{\mathcal{I}}^{\alpha}} = 0 \tag{3.4}$$

holds with $\mathcal{D}^{\alpha}, \mathcal{G}^{\alpha}$ given by (2.10), (2.14), and

$$\lambda(x) = \sum_{y\in\overline{\mathcal{V}}} \Theta(x,y)\alpha^2(x,y) + \Theta(x,x), \quad x \in \mathcal{V}. \tag{3.5}$$

In addition, if $\lambda(x) \leq 1$ in (3.5) $\forall x \in \mathcal{V}$, then Ω given by (3.3) satisfies (2.2), and equality $\lambda(x) = 1\, \forall x \in \mathcal{V}$ is achieved if property (2.34) holds.

Proof. See section A.1 for the proof.

Remark 3.2(comments). Lemma 3.1 characterizes a class of parameter matrices Ω whose action (3.4) admits a representation using the nonlocal operators from section 2.1.

Some comments follow on parameter matrices *not* covered by Lemma 3.1, due to the imposed constraints.

(i) By ignoring the *nonnegativity constraint* of (3.1) imposed on Ω through the mapping Θ, Lemma 3.1 additionally covers a class of nonlocal graph Laplacians proposed in [5] and [3] for the aim of image inpainting. We refer to section 4 for a more detailed discussion.

(ii) Due to assuming *symmetry* of the mapping Θ, formulation (3.3) does *not* cover nonlocal diffusion processes on *directed* graphs $(\mathcal{V}, \mathcal{E}, \Omega)$.

(iii) Imposing *zero nonlocal Dirichlet boundary conditions* is essential for relating assignment flows to the specific class of nonlocal PDEs related to (3.4); see Proposition 3.3 below.

As argued in [19] by a range of counterexamples, using nonsymmetric parameter matrices Ω compromises convergence of the assignment flow (2.38a) to integral solutions (labelings) and is therefore not considered. The study of more general parameter matrices is left for future work; see sections 8 and 4.1 for modifying the identity (3.4) in view of nonsymmetric parameter matrices Ω.

Next, we generalize the common *local* boundary conditions for PDEs to nonlocal *volume constraints* for *nonlocal* PDEs on discrete domains. Following [27], given an antisymmetric mapping α as in (2.8) and Lemma 3.1, the natural domains $\mathcal{V}^\alpha_{\mathcal{I}_N}, \mathcal{V}^\alpha_{\mathcal{I}_D}$ for imposing nonlocal *Neumann* and *Dirichlet* constraints are given by a disjoint decomposition of the interaction domain (2.8),

$$\mathcal{V}^\alpha_{\mathcal{I}} = \mathcal{V}^\alpha_{\mathcal{I}_N} \dot\cup \mathcal{V}^\alpha_{\mathcal{I}_D}. \tag{3.6}$$

The following proposition reveals how the flow (2.38a), with Ω satisfying the assumptions of Lemma 3.1, can be reformulated as a nonlocal partial difference equation with zero nonlocal Dirichlet boundary condition imposed on the entire interaction domain, i.e., $\mathcal{V}^\alpha_{\mathcal{I}} = \mathcal{V}^\alpha_{\mathcal{I}_D}$. Recall the definition of the manifold $\mathcal{S}$ of discrete probability vectors with full support in connection with (2.21).

Proposition 3.3 (S-flow as nonlocal G-PDE). *Let $\alpha, \Theta \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ be as in (3.2). Then the flow (2.38a) with Ω given through (3.3) admits the representation*

$$\partial_t S(x,t) = R_{S(x,t)}\left(\frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(S)\right) + \lambda S\right)(x,t) \qquad \text{on } \mathcal{V}\times\mathbb{R}_+, \tag{3.7a}$$

$$\overline{S}(x,t) = 0 \qquad \text{on } \mathcal{V}^\alpha_{\mathcal{I}}\times\mathbb{R}_+, \tag{3.7b}$$

$$S(x,0) = \overline{S}(x)(0) \qquad \text{on } \overline{\mathcal{V}}\times\mathbb{R}_+, \tag{3.7c}$$

where $\lambda = \lambda(x)$ is given by (3.2) and $\overline{S} \in \mathcal{F}_{\overline{\mathcal{V}},\mathbb{R}^c_+}$ denotes the zero extension of the $\mathcal{S}$-valued vector field $S \in \mathcal{F}_{\mathcal{V},\mathcal{S}}$ to the interaction domain $\overline{\mathcal{V}}^\alpha_{\mathcal{I}}$.

Proof. See section A.1 for the proof.

Proposition 3.3 states the equivalence of the potential flow (2.38a), with Ω defined by (3.3), and the nonlocal diffusion process (3.7) with zero nonlocal Dirichlet boundary condition. We now explain that the system (3.7a) can represent *any* descent flow of the form (2.38a) defined in terms of an *arbitrary* nonnegative symmetric mapping $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$. Specifically, given such a mapping Ω, let the mappings $\widetilde{\alpha}, \widetilde{\Theta} \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ be defined by

$$\widetilde{\Theta}(x,y) = \begin{cases} \Omega(x,y) & \text{if } y \in \mathcal{N}(x), \\ 0 & \text{else,} \end{cases} \qquad \widetilde{\alpha}^2(x,y) = 1, \qquad x,y \in \mathcal{V}. \tag{3.8}$$

Further, denote by $\Theta, \alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ the extensions of $\widetilde{\alpha}, \widetilde{\Theta}$ to $\overline{\mathcal{V}} \times \overline{\mathcal{V}}$ by 0, that is,

$$\Theta(x,y) = \Big(\delta_{\mathcal{V}\times\mathcal{V}}(\widetilde{\Theta})\Big)(x,y), \quad \alpha(x,y) := \big(\delta_{\mathcal{V}\times\mathcal{V}}(\widetilde{\alpha})\big)(x,y), \quad x,y \in \overline{\mathcal{V}}, \tag{3.9}$$

where $\delta_{\mathcal{V}\times\mathcal{V}} \colon \mathbb{Z}^d \times \mathbb{Z}^d \to \{0,1\}$ is the indicator function of the set $\mathcal{V} \times \mathcal{V} \subset \mathbb{Z}^d \times \mathbb{Z}^d$. Then the potential flow (2.38a) with Ω satisfying $\Omega(x,y) = \Omega(y,x)$ is equivalently represented by the system (3.7) with an empty interaction domain (2.8). This shows how Proposition 3.3 generalizes the assignment flow introduced in section 2.2 by ignoring the constraint (2.34) imposed on Ω, and thus enables use of a broader class of parameter matrices Ω controlling the labeling process; see also Remark 3.2.

3.2. Tangent-space parametrization of the S-flow G-PDE. Because $S(x,t)$ solving (3.7) evolves on the non-Euclidean space $\mathcal{S}$, applying some standard discretization in order to evaluate (3.7) numerically will not work. Therefore, motivated by the work [20] on the geometric numerical integration of the original assignment flow system (2.35), we devise a parametrization of (3.7) on the *flat* tangent space (2.24) by means of the equation

$$S(t) = \exp_{S^0}(V(t)) \in \mathcal{W}, \qquad V(t) \in \mathcal{T}_0, \qquad S^0 = S(0) \in \mathcal{W}, \tag{3.10}$$

where analogous to (2.40)

$$\exp_{S^0}(V(t)) = \Big(\ldots, \exp_{S^0(x)}(-V(x,t)), \ldots\Big)^\top \in \mathcal{W} \tag{3.11}$$

with $\exp_{S^0(x)}$ given by (2.41). Applying $\frac{d}{dt}$ to both sides and using the expression of the differential of the mapping $\exp_{S^0}$ due to [29, Lemma 3.1], we get

$$\dot{S}(t) = R_{\exp_{S^0}(V(t))}\dot{V}(t) \overset{(3.10)}{=} R_{S(t)}\dot{V}(t). \tag{3.12}$$

Comparing this equation and (2.38a), and taking into account $R_S = R_S\Pi_0$, shows that $V(t)$ solving the nonlinear ODE

$$\dot{V}(t) = \Pi_0 \Omega \exp_{S^0}(V(t)), \qquad V(0) = 0, \tag{3.13}$$

determines $S(t)$ by (3.10) solving (2.38a). Hence it suffices to focus on (3.13), which evolves on the flat space $\mathcal{T}_0$. Repeating the derivation above that resulted in the G-PDE representation

(3.7) of the S-flow (2.38a) yields the nonlinear PDE representation of (3.13)

$$\partial_t V(x,t) = \left(\frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(\exp_{S^0}(V))\right) + \lambda \exp_{S^0}(V)\right)(x,t) \qquad \text{on } \mathcal{V}\times\mathbb{R}_+, \tag{3.14a}$$

$$\overline{V}(x,t) = 0 \qquad \text{on } \mathcal{V}_{\mathcal{I}}^\alpha \times \mathbb{R}_+, \tag{3.14b}$$

$$V(x,0) = \overline{V}(x)(0) \qquad \text{on } \overline{\mathcal{V}}\times\mathbb{R}_+, \tag{3.14c}$$

where $\overline{V} \in \mathcal{F}_{\overline{\mathcal{V}},\mathcal{T}_0}$ denotes the zero extension of the $\mathcal{T}_0$-valued vector field to the interaction domain $\mathcal{V}_{\mathcal{I}}^\alpha$. From the numerical point of view, this new formulation (3.10), (3.14) has the following expedient properties. First, using a parameter matrix as specified by (3.3) and (3.9) enables us to define the entire system (3.14) on $\mathcal{V}$. Second, since $V(x,t)$ evolves on the flat space T_0, numerical techniques of geometric integration as studied by [20] can here be applied as well. We utilize this fact in sections 3.4.1 and 5.

3.3. Nonlocal balance law. A key property of PDE-based models are balance laws implied by the model; see [28, section 7] for a discussion of various scenarios. The following proposition reveals a *nonlocal* balance law of the assignment flow based on the novel G-PDE-based parametrization (3.14), which we express for this purpose in the form

$$\partial_t V(x,t) + \mathcal{D}^\alpha(F(V))(x,t) = b(x,t), \qquad b(x,t) = \lambda(x)S(x,t), \qquad x\in\mathcal{V}, \tag{3.15a}$$

$$F(V(t))(x,y) = -\frac{1}{2}\left(\Theta\mathcal{G}^\alpha\left(\exp_{S^0}(V(t))\right)\right)(x,y), \tag{3.15b}$$

where $S(x,t) = \exp_{S^0}(V(x,t))$ is given by (3.10) and $\lambda(x)$ is given by (3.5).

Proposition 3.4 (nonlocal balance law of assignment flows). *Under the assumptions of Lemma 3.1, let $V(t)$ solve (3.14). Then, for each component $S_j(t) = \{S_j(x,t)\colon x\in\mathcal{V}\}$, $j\in[c]$, of $S(t) = \exp_{S^0}(V(t))$, the identity*

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\langle S_j, \mathbb{1}\rangle_{\mathcal{V}} + \frac{1}{2}\langle \mathcal{G}^\alpha(S_j), \Theta\mathcal{G}^\alpha(S_j)\rangle_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}} + \langle S_j, \phi_S - \lambda S_j\rangle_{\mathcal{V}} \\ + \langle S_j, \mathcal{N}^\alpha(\Theta\mathcal{G}^\alpha(S_j))\rangle_{\mathcal{V}_{\mathcal{I}^\alpha}} = 0\end{aligned} \tag{3.16}$$

holds, where the inner products are given by (2.5) and (2.6), and $\phi_S(\cdot)\in\mathcal{F}_{\mathcal{V}}$ is defined in terms of $S(t)\in\mathcal{W}$ by

$$\phi_S : \mathcal{V}\to\mathbb{R}, \qquad x\mapsto \left\langle S(x), \Pi_0\left(\Omega S\right)(x)\right\rangle. \tag{3.17}$$

Proof. See section A.2 for the proof.

The nonlocal balance law (3.16) comprises four terms. Since $\sum_{j\in[c]} S_j(x) = 1$ at each vertex $x\in\mathcal{V}$, the first term of (3.16) measures the *rate of "mass"* assigned to label j over the entire image. This rate is governed by two interacting processes corresponding to the three remaining terms:

(i) *spatial propagation of assignment mass* through the nonlocal diffusion process including nonlocal boundary conditions (second and fourth terms);
(ii) *exchange of assignment mass* with the remaining labels $\{l\in[c]\colon l\neq j\}$ (third term comprising the function ϕ_S (3.17)).

We point out that other approaches to image labeling, including Markov random fields and deep networks, do *not* reveal the flow of information during inference in such an explicit manner.

3.4. Illustration: Parametrization and nonlocal boundary conditions. In this section, we illustrate two aspects of the mathematical results presented above by numerical results:

(1) The use of *geometric integration* for numerically solving the nonlocal G-PDE (3.7). Here we exploit a basic numerical scheme established for the assignment flow (2.38a) and the one-to-one correspondence to the nonlocal G-PDE (3.7), due to Proposition 3.3.

(2) The effect of zero versus nonzero nonlocal Dirichlet boundary conditions and uniform versus nonuniform parametrizations (3.3). Using nonzero boundary conditions refers to the observation stated above in connection with (3.8), (3.9): the nonlocal G-PDE (3.7) generalizes the assignment flow when constraints are dropped. Here specifically, the homogeneous Dirichlet boundary condition may be nonhomogeneous, and the constraint (2.34) is ignored; see also Remark 3.2.

Topic (1) is addressed here to explain how the results illustrating topic (2) were computed, and to set the stage for section 5, which presents an advanced numerical scheme. Item (2) merely illustrates basic choices of the parametrization and boundary conditions. More advanced generalizations of the assignment flow are conceivable but are beyond the scope of this paper; see section 8.

3.4.1. Numerically solving the nonlocal G-PDE by geometric integration. According to section 3.2, imposing the homogeneous Dirichlet condition via the interaction domain (2.8) makes the right-hand side of (3.14a) equivalent to (3.13). Applying to (3.14a) a simple explicit time discretization with step size h results in the iterative update formula

$$V(x,t+h) \approx V(x,t) + h\Pi_0 \exp_{S^0(x)}(\Omega V(x,t)), \qquad h>0. \tag{3.18}$$

By virtue of the parametrization (3.10), one recovers with any nonnegative symmetric mapping Ω as in Lemma 3.1 the *explicit geometric Euler* scheme on $\mathcal{W}$

$$S(t+h) \approx \exp_{S^0}\left(V(t)+h\dot V(t)\right) \overset{\substack{(2.42a)\\(3.10)}}{=} \exp_{S(t)}\left(h\dot V(t)\right) \tag{3.19a}$$

$$\overset{\substack{(2.42b)\\(3.13)}}{=} \exp_{S(t)}\left(h\Omega S(t)\right). \tag{3.19b}$$

Higher-order geometric integration methods [20] generalizing (3.19) can be applied in a similar way. This provides a new perspective on solving a certain class of nonlocal G-PDEs numerically, conforming to the underlying geometry, as we demonstrate in section 5.2.

3.4.2. Basic parametrizations, effect of nonlocal Dirichlet boundary conditions. We consider two different parametrizations as well as zero and nonzero nonlocal Dirichlet boundary conditions.

Uniform parametrization: Mappings $\Theta, \alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ are given by

$$|\mathcal{N}(x)| = \mathcal{N}\ \forall x, \qquad |\mathcal{N}| = (2k+1)\times(2k+1),\ k\in\mathbb{N}, \tag{3.20a}$$

$$\alpha^2(x,y) = \begin{cases} \frac{1}{(2k+1)^2} & \text{if } y\in\mathcal{N}(x), \\ 0 & \text{otherwise,} \end{cases} \qquad \Theta(x,y) = \begin{cases} \frac{1}{(2k+1)^2} & \text{if } x = y, \\ 1 & \text{otherwise.} \end{cases} \tag{3.20b}$$

Nonuniform parametrization: Uniform neighborhoods (3.20a) and mappings $\Theta, \alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ are given by

$$\begin{aligned} \alpha^2(x,y) &= \begin{cases} e^{-\frac{\|x-y\|^2}{2\sigma_s^2}} & \text{if } y\in\mathcal{N}(x) \\ 0 & \text{otherwise,} \end{cases}, \qquad \sigma_s > 0, \\ \Theta(x,y) &= \begin{cases} e^{-G_{\sigma_p} * \|s(x)-s(y)\|^2} & \text{if } y\in\mathcal{N}(x), \\ 0 & \text{otherwise,} \end{cases} \qquad \sigma_p > 0, \end{aligned} \tag{3.21}$$

where the nonlocal function Θ is designed using a patchwise similarity measure analogous to the basic nonlocal means approach [4]: $s(x) = \{s(x,z)\colon z\in\overline{\mathcal{V}}, s(x,z) = \overline{X}(z)\}$ with $\overline{X}\in\mathcal{F}_{\overline{\mathcal{V}},\mathbb{R}^c}$ denoting the zero extension of data $X\in\mathcal{F}_{\mathcal{V},\mathbb{R}^c}$ to $\mathcal{V}_{\mathcal{I}}^{\alpha}$. G_{σ_p} is the Gaussian kernel at scale σ_p and $*$ denotes spatial convolution.

We iterated (3.19) with step size $h = 1$ until assignment states (2.38b) of low average entropy 10^{-3} were reached. To ensure a fair comparison and to assess solely the effects of the boundary conditions through nonlocal regularization, we initialized (3.7) in the same way as (2.38a) and adopted a uniform encoding of the 31 labels as described by [17, Figure 6].

Figure 5 depicts labelings computed using the uniform parametrization with zero and nonzero nonlocal Dirichlet boundary conditions, respectively. Inspecting panels (c) (zero boundary condition) and (d) (nonzero boundary condition) shows that using the latter may improve labeling near the boundary (cf. close-up views), whereas the labelings almost agree in the interior of the domain.

Figure 6 shows how the average entropy values of label assignments decrease as the iteration proceeds (left panel) and the number of iterations required to converge (right panel), for different neighborhood sizes. Moreover, a closer look at the right panel of Figure 6 reveals besides a slightly slower convergence of the scheme (3.18) applied to the nonlocal G-PDE (3.14) (red curve), the dependence of number of iterations required until convergence is comparable to the S-flow (green curve). Consequently, generalizing the S-flow by the nonlocal model (3.7) does not have a detrimental effect on the overall numerical behavior. We observe, in particular, that integral label assignments corresponding to zero entropy are achieved no matter which boundary condition is used, at comparable computational costs.

Iterating (3.19) with step size $h = 0.1$ and $\sigma_s = 1, \sigma_p = 5$ in (3.21) yields labeling results for different patch sizes as depicted by Figure 7. As opposed to segmentation results obtained with uniform parametrization (3.20b) for $\mathcal{N} = 7$ depicted in Figure 5(d), a direct comparison with Figure 7 (close-up views) indicates more accurate labelings when using regularization as given by the nonuniform parametrization (3.21).

4. Related work.

In this section, we discuss how the system (3.7) relates to approaches based on PDEs and variational models in the literature. Specifically, we conduct an *analytical ablation* study of the nonlocal model (3.7) in order to clarify the impact of omitting operators

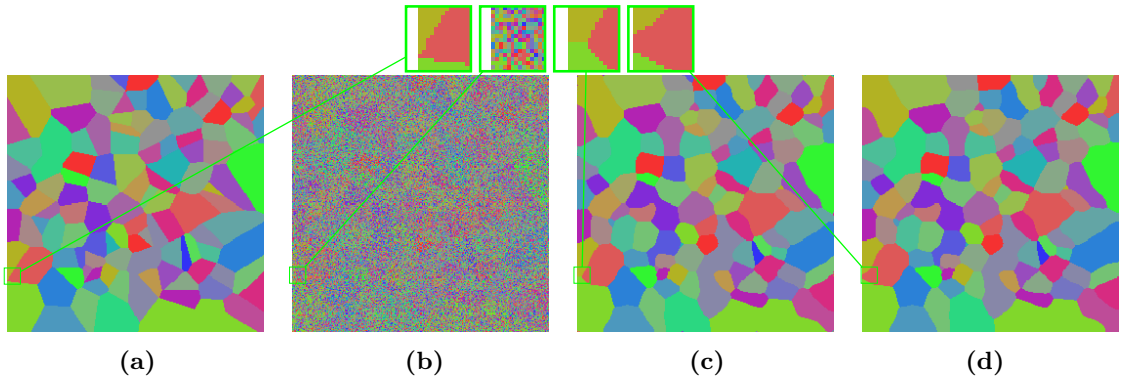


Figure 5. *Labeling through the nonlocal geometric assignment flow with uniform parametrization* (3.20b) *and neighborhood size* $|\mathcal{N}| = 7$. (a) *Ground truth with* 31 *labels.* (b) *Noisy input data used to evaluate* (2.38a) *and* (3.7)*, respectively.* (c) *Labeling returned when using the zero nonlocal Dirichlet boundary condition.* (d) *Labeling returned when using the nonzero nonlocal Dirichlet boundary condition (uniform extension to the interaction domain). The close-up views show differences close to the boundary, whereas the results in the interior domain are almost equal.*

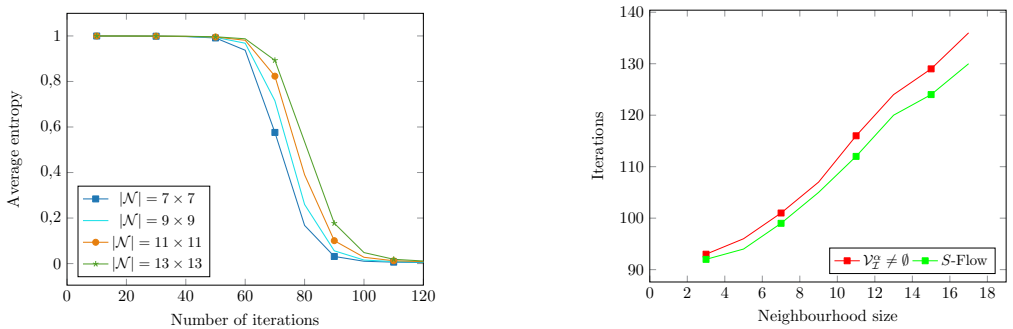


Figure 6. *Left: Convergence rates of the scheme* (3.19) *solving* (3.7) *with nonzero nonlocal Dirichlet boundary condition. The convergence behavior is rather insensitive with respect to the neighborhood size* $|\mathcal{N}|$*. Right: Number of iterations until convergence for* (3.7) (●) *and* (2.38a) (●)*, with zero nonlocal boundary condition in the latter case. The result shows that different nonlocal boundary conditions have only a minor influence on the required number of geometric integration steps.*

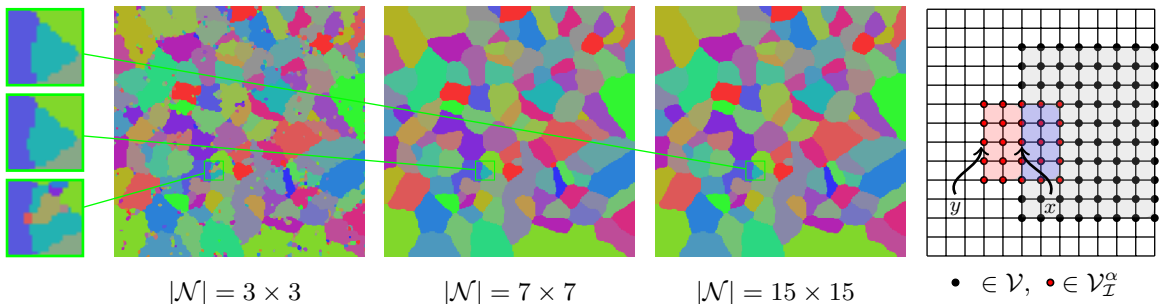


Figure 7. *From left to right: Labeling results using* (3.7) *with the nonuniform parametrization* (3.21)*, zero nonlocal Dirichlet boundary conditions, and neighborhood sizes* $|\mathcal{N}| \in \{3\times3, 7\times7, 15\times15\}$*. Schematic illustration of the nonlocal interaction domain* $y \in \mathcal{V}_{\mathcal{I}}^{\alpha}$ *(red area) induced by nodes (blue area) in* $\mathcal{N}(x)$ *with* $|\mathcal{N}| = 5 \times 5$*. Using nonuniform weights* (3.21) *improves labeling accuracy.*

of the nonlocal model and the connection to existing methods. We exhibit both structural similarities from the viewpoint of diffusion processes and differences that account for the *different scope* of our approach: *labeling* metric data on graphs.

4.1. General nonlocal processes on graphs. We consider again the identity (3.4) that defines the nonlocal G-PDE (3.7) in terms of *symmetric* parameter mapping (3.3) and show next how (3.4) is generalized when a *nonsymmetric* parameter matrix $\Omega \in \mathcal{F}_{\mathbb{Z}^d \times \mathbb{Z}^d}$ is used. Specifically, suppose a kernel $k \in \mathcal{F}_{\mathbb{Z}^d \times \mathbb{Z}^d}$ is given and the induced nonlocal functional

$$\mathcal{L}_k f(x) = \sum_{y \in \mathbb{Z}^d} \left(f(y) k(y,x) - f(x) k(x,y) \right). \tag{4.1}$$

Then, for a mapping α that satisfies $\alpha^2(x,y) = 1$ whenever $k(x,y) \neq 0$, the decomposition

$$k = k^s + k^a \quad \text{with} \quad k^s = \frac{k + k'}{2}, \quad k^a = \frac{k - k'}{2}, \qquad k'(x,y) := k(y,x), \quad x, y \in \mathbb{Z}^d, \tag{4.2}$$

results in the representation

$$k(x,y) = \begin{cases} 2\Theta(x,y)\alpha^2(x,y) + \alpha(x,y)\nu(x,y), & x \neq y, \\ 2\Theta(x,x), & x = y, \end{cases} \tag{4.3}$$

of the kernel k in terms of $\alpha, \Theta \in \mathcal{F}_{\mathbb{Z}^d \times \mathbb{Z}^d}$ and $\nu \in \mathcal{F}_{\mathbb{Z}^d \times \mathbb{Z}^d}$ given by

$$\Theta(x,y) := \frac{1}{2} k^s(x,y), \qquad \nu(x,y) := k^a(x,y)\alpha(x,y), \tag{4.4}$$

where the mapping ν is symmetric due to the antisymmetry of α. Inserting (4.3) into (4.1) yields

$$\mathcal{L}_k f(x) = 2 \sum_{y \in \mathbb{Z}^d} \Theta(x,y)\alpha^2(x,y) \left(f(y) - f(x) \right) - \sum_{y \in \mathbb{Z}^d} \alpha(x,y)\nu(x,y) \left(f(y) - f(x) \right) \tag{4.5}$$

and applying nonlocal calculus of section 2.1 along with Lemma 3.1, we arrive at an equivalent representation of $\mathcal{L}_k$ through nonlocal divergence and gradient operators

$$\mathcal{L}_k f(x) \overset{(4.3)}{=} \underbrace{\mathcal{D}^\alpha \left(\Theta \mathcal{G}^\alpha(f) \right)(x)}_{\text{diffusion}} - \underbrace{\mathcal{D}^\alpha(\nu f)(x)}_{\text{convection}} + \underbrace{\lambda(x) f(x)}_{\text{fidelity}}, \tag{4.6}$$

where ν plays the role of the convection parameter. Consequently, on a grid graph $\mathcal{G}$ with $\mathcal{V} \subset \mathbb{Z}^d$ and setting Ω by (4.3), we get

$$\partial_t S(x,t) = R_{S(x,t)} \left(\mathcal{D}^\alpha \left(\Theta \mathcal{G}^\alpha(S) \right) - \mathcal{D}^\alpha(\nu S) \right)(x,t) + \lambda(x) S(x,t) \quad \text{on} \quad \mathcal{V} \times \mathbb{R}_+, \tag{4.7a}$$

$$\overline{S}(x,t) = 0 \quad \text{on} \quad \mathcal{V}_{\mathcal{I}}^\alpha \times \mathbb{R}_+, \tag{4.7b}$$

$$\overline{S}(x,0) = S(x)(0) \quad \text{on} \quad \overline{\mathcal{V}} \times \mathbb{R}_+, \tag{4.7c}$$

with the interaction domain (2.8) directly expressed through the connectivity of kernel k by

$$\mathcal{V}_{\mathcal{I}}^\alpha = \{ x \in \mathbb{Z}^d \setminus \mathcal{V} \colon k(x,y) \neq 0 \text{ for some } y \in \mathcal{V} \}. \tag{4.8}$$

Table 1

Summary of the analytical ablation study. *Key differences of our approach to existing nonlocal diffusion models are inclusion of the replicator operator R_S and a nonzero fidelity term λS that results in nontrivial solution at the steady state $S^* = S(t=\infty)$.*

	Labeling		Denoising and inpainting	
Parameters	**G-PDE (3.7)**	**Local PDE [29]**	**Nonlocal Laplacian [5]**	**Descent flow [3]**
$\Theta \geq 0$	✓	✘	✘	✘
λ	$\lambda > 0$	$\lambda = 1$	$\lambda = 0$	$\lambda = 0$
R_S	✓	✓	✘	✘
ν	✘	✘	✘	✘
$\mathcal{V}_{\mathcal{I}}^{\alpha}$	$\subseteq \mathbb{Z}^d \setminus \mathcal{V}$	$\partial \mathcal{V}^h$	$\partial \mathcal{A} \subset \mathcal{V}$	$\emptyset$
$S^*(t \to \infty)$	✓	✓	✘	✘

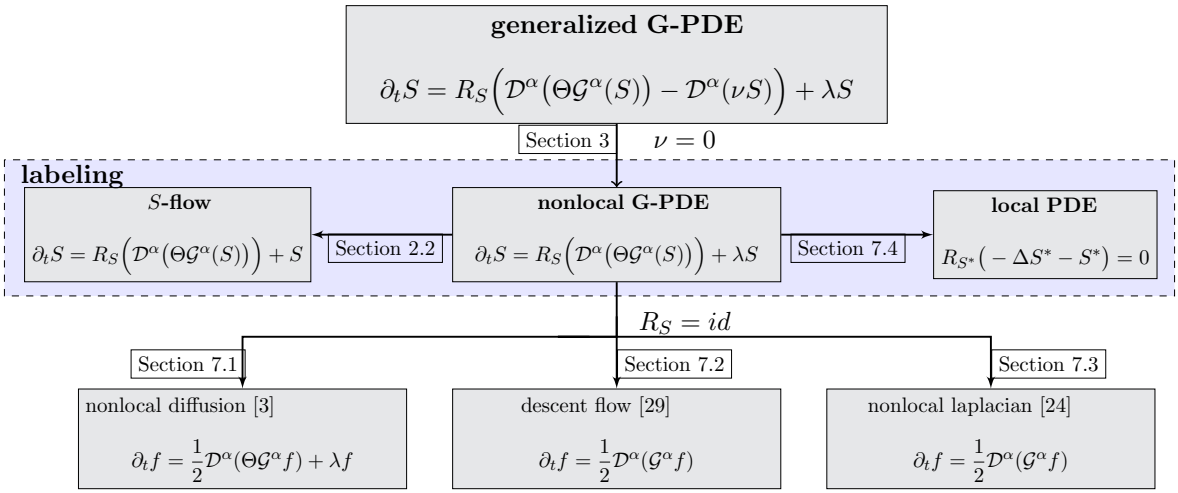


Figure 8. *Overview of nonlocal diffusion processes proposed in related work [3, 5, 37] and their interrelations to the nonlocal G-PDE (4.7). The approaches highlighted by the blue region only model the image labeling problem. Edge labels refer to the corresponding sections of the analytical ablation study.*

In view of (4.7), we therefore recognize the system (3.7) as a *specific* nonlocal process that is induced by *nonnegative symmetric* kernels k with nonzero fidelity parameter λ, which account for nontrivial steady state solutions and zero convection ($\nu(x,y) = 0$).

In the following sections, we relate different established nonlocal models to the proposed G-PDE (3.7) by adapting the parameter mappings $\Theta, \alpha \in \mathcal{F}_{\overline{\mathcal{V}} \times \overline{\mathcal{V}}}$ that parametrize the G-PDE and determine the interaction domain (2.8). Figure 8 provides an overview of the analytical ablation study by specifying the model and the corresponding section where it is derived from the generalized G-PDE (4.7). Table 1 lists the involved parameters for each model.

4.2. Relation to a local PDE that characterizes labelings. We focus on the connection of the system (3.7) and the *continuous-domain local* formulation of (2.38a) on an open simply connected bounded domain $\mathcal{D} \subset \mathbb{R}^2$, as introduced by [29]. The *variational* formulation has been rigorously derived in [29] along with a PDE that formally characterizes solutions

$S^* = \lim_{t\to\infty} S(t) \in \overline{\mathcal{W}}$ only under strong regularity assumptions. This nonlinear PDE reads

$$R_{S^*(x)}\left(-\Delta S^*(x) - S^*(x)\right) = 0, \quad x \in \mathcal{D}. \tag{4.9}$$

We next show that our novel approach (3.7) includes, as a special case, a natural discretization of (4.9) on the spatial discrete grid $\mathcal{V}^h = h\mathbb{Z}^d \cap \mathcal{D}$ with boundary $\partial\mathcal{V}^h$ specified by a small spatial scale parameter $h > 0$. (4.9) is complemented by *local zero Dirichlet* boundary conditions imposed on S^* on $\partial\mathcal{V}^h$. Adopting the sign convention $L^h_\vartheta = -\Delta^h_\vartheta$ for different discretizations of the *continuous negative Laplacian* on $\mathcal{V}^h$, by a nine-point stencil [45] parametrized by $\vartheta \in [0,1]$, leads to strictly positive entries $L^h_\vartheta(x,x) > 0$ on the diagonal.

We introduce the weighted undirected graph $(\mathcal{V}^h, \Omega^h)$ and identify nodes $x = (k,l) \in \mathcal{V}^h$ with interior grid points $(hk, hl) \in \mathcal{V}^h$ (grid graph). Let the parameter matrix Ω^h be given by (3.3) and the mappings $\alpha, \Theta \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ be defined by

$$\alpha^2(x,y) = \begin{cases} 1, & y \in \widetilde{\mathcal{N}}(x), \\ 0 & \text{else}, \end{cases} \qquad \Theta(x,y) = \begin{cases} -L^h_\vartheta(x,y), & y \in \widetilde{\mathcal{N}}(x), \\ 1 - L^h_\vartheta(x,x), & x = y, \\ 0 & \text{else}\ , \end{cases} \tag{4.10}$$

where the neighborhoods $\widetilde{\mathcal{N}}(x) = \mathcal{N}(x) \setminus \{x\}$ represent the connectivity of the stencil of the discrete Laplacian L^h_ϑ on the mesh $\mathcal{V}^h \dot\cup \partial\mathcal{V}^h$. Recalling the definitions from section 2.1 with respect to undirected graphs and setting α by (4.10), the interaction domain (2.8) agrees for parameter choices $\vartheta \neq 0$ with the discrete local boundary, i.e., $\mathcal{V}^\alpha_{\mathcal{I}} = \partial\mathcal{V}^h$; see Figure 9 and the caption for further explanation. Then, for each $x \in \mathcal{V}^h$, the action of Ω^h on S reads

$$(\Omega^h S)(x) = \sum_{y\in\widetilde{\mathcal{N}}(x)} -L^h_\vartheta(x,y)S(y) + \left(1 - L^h_\vartheta(x,x)\right) S(x) = -\left(-\Delta^h_\vartheta(S) - S\right)(x), \tag{4.11}$$

which is the discretization of (4.9) by L^h_ϑ multiplied by the minus sign. In particular, due to the relation $R_S(-W) = -R_S(W)$ for $W \in \mathcal{W}$, we conclude that the novel approach (3.7) includes the *local* PDE (4.9) as a special case and hence provides a *natural nonlocal extension*.

4.3. Continuous-domain nonlocal diffusion processes. We follow [37]. Consider a bounded domain $\mathcal{D} \subset \mathbb{R}^d$ and let $J : \mathbb{R}^d \to \mathbb{R}_+$ be a radial continuous function satisfying

$$\int_{\mathbb{R}^d} J(x-y)\mathrm{d}y = 1, \quad J(0) > 0, \qquad \forall x \in \mathbb{R}^d. \tag{4.12}$$

The term $J(x-y)$ in (4.12) may be interpreted as a probability density governing jumps from position $y \in \mathbb{R}^d$ to $x \in \mathbb{R}^d$. The authors of [37] introduced the integral operator

$$\mathcal{L}f(x) = \int_{\mathbb{R}^d} J(x-y)f(y,t)\mathrm{d}y - f(x,t), \quad x \in \mathbb{R}^d, \tag{4.13}$$

acting on $f \in C(\mathbb{R}^d, \mathbb{R}_+)$ and studied *nonlocal linear* diffusion processes of the form

$$\partial_t f(x,t) = \mathcal{L}f(x,t) \quad \text{on} \quad \mathcal{D}\times\mathbb{R}_+, \tag{4.14a}$$

$$f(x,t) = g(x) \quad \text{on} \quad \mathbb{R}^d \setminus \mathcal{D} \times \mathbb{R}_+, \tag{4.14b}$$

$$f(x,0) = \overline{f}_0 \quad \text{on} \quad \mathbb{R}^d \times \mathbb{R}_+, \tag{4.14c}$$

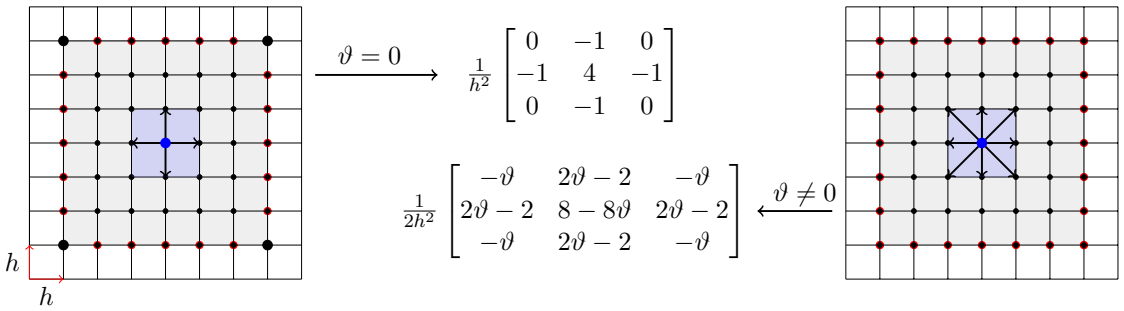


Figure 9. *Illustration of the rectangular grid $\mathcal{V}^h$ and the interaction domain $\mathcal{V}_{\mathcal{I}}^{\alpha}$ represented by • and •, respectively, with $\alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ given by (4.10) for a family of discrete Laplacians $-\Delta_{\vartheta}^{h}$ proposed in [45]. Left: Neighborhood $\widetilde{\mathcal{N}}(x)$ specified in terms of the connectivity of the standard five-point stencil ($\vartheta = 0$). The corresponding interaction domain is part of the local boundary $\mathcal{V}_{\mathcal{I}}^{\alpha} \subset \partial\mathcal{V}^h$. Right: Analogous construction with the nine-point stencil ($\vartheta \neq 0$). The interaction domain coincides with the discrete local boundary configuration, i.e., $\mathcal{V}_{\mathcal{I}}^{\alpha} = \partial\mathcal{V}^h$.*

where $f_0 \in C(\mathcal{D}, \mathbb{R}_+)$ and $g \in C(\mathbb{R}^d \setminus \mathcal{D}, \mathbb{R}_+)$ specify the initial state and the nonlocal boundary condition of the system (4.14), respectively. We compare this system with our model (3.7) and introduce, as in section 4.3, the weighted undirected graph $(\mathcal{V}^h, \Omega^h)$ with a Cartesian mesh $\mathcal{V}^h$, with boundary $\partial\mathcal{V}^h$ and neighborhoods (2.3), and with Ω^h defined by (3.8) through

$$\Theta(x,y) = \begin{cases} 0 & \text{for } x,y \notin \mathcal{V}^h, \\ J(0)-1 & \text{for } x = y, \\ 1 & \text{else,} \end{cases} \qquad \alpha^2(x,y) = J(x-y). \tag{4.15}$$

Then, for the particular case $g = 0$ in (4.14a) and using (3.4) with $\lambda(x)$ defined by (3.5), the spatially discrete counterpart of (4.14) is the *linear nonlocal scalar-valued* diffusion process

$$\partial_t f(x,t) = \frac{1}{2}\mathcal{D}^{\alpha}(\Theta\mathcal{G}^{\alpha}f)(x,t) + \lambda(x)f(x,t) \qquad \text{on} \quad \mathcal{V}\times\mathbb{R}_+, \tag{4.16a}$$

$$f(x,t) = 0 \qquad \text{on} \quad \mathcal{V}_{\mathcal{I}}^{\alpha}\times\mathbb{R}_+, \tag{4.16b}$$

$$f(x,0) = \overline{f}_0 \qquad \text{on} \quad \overline{\mathcal{V}}\times\mathbb{R}_+. \tag{4.16c}$$

System (4.16) possesses a structure which resembles the structure of nonlinear system (3.7) after dropping the replicator mapping R_S and assuming $S(x) \in \mathbb{R}$ to be a scalar-valued rather than simplex-valued $S(x) \in \mathcal{S}$, as in our approach.

This comparison shows by virtue of the structural similarity that assignment flows may be characterized as genuine nonlocal diffusion processes. Essential differences, i.e., simplex-valued variables and the underlying geometry, reflect the entirely different scope of this process, however: labeling metric data on graphs.

4.4. Nonlocal variational models in image analysis. We relate the system (4.16) to variational approaches presented in [3] and to graph-based nonlocal PDEs proposed by [2, 5].

Based on a scalar-valued positive function $\phi(t)$ which is convex in $\sqrt{t}$ with $\phi(0) = 0$, Gilboa and Osher [3] studied *isotropic* and *anisotropic nonlocal regularization functionals* on

a continuous spatial domain $\mathcal{D} \subset \mathbb{R}^d$ defined in terms of a nonnegative symmetric mapping $\omega : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_+$:

$$J_i^\phi(f) = \int_{\mathcal{D}} \phi(|\nabla_\omega(f)(x)|^2)\mathrm{d}x \qquad \textbf{(isotropic)}, \tag{4.17a}$$

$$J_a^\phi(f) = \int_{\mathcal{D}} \int_{\mathcal{D}} \phi(f(y) - f(x))^2 \omega(x,y)\mathrm{d}y\mathrm{d}x \qquad \textbf{(anisotropic)}. \tag{4.17b}$$

(4.17a) involves the nonlocal graph-based gradient operator which for given neighborhoods $\mathcal{N}(x)$ reads

$$\nabla_\omega f(x) = \Big(\dots, (f(y) - f(x))\sqrt{\omega(x,y)}, \dots\Big)^T, \qquad y \in \mathcal{N}(x). \tag{4.18}$$

Given an initial real-valued function $f_0(x)$ on Ω, the variational models of (4.17) define dynamics in terms of the steepest descent flows

$$\partial_t f(x,t) = -\partial_f J_i^\phi(f)(x,t), \qquad \partial_t f(x,t) = -\partial_f J_a^\phi(f)(x,t), \qquad f(x,0) = f_0(x), \tag{4.19}$$

where the variation with respect to f on right-hand side of (4.19) is expressed in terms of (4.18) via

(4.20)

$$\partial_f J_i^\phi(f)(x,t) = -2 \int_{\mathcal{D}} (f(y,t) - f(x,t))\omega(x,y) \left(\phi'(|\nabla_\omega f(y,t)|^2)(y) + \phi'(|\nabla_\omega f(x,t)|^2)(x)\right) \mathrm{d}y,$$

$$\partial_f J_a^\phi(f)(x,t) = -4 \int_{\mathcal{D}} (f(y,t) - f(x,t))\, \omega(x,y)\phi'\left((f(y,t) - f(x,t))^2 \omega(x,y)\right) \mathrm{d}y.$$

Then, given a graph $(\mathcal{V}, \mathcal{E}, \omega)$ with neighborhoods as in section 2.1, the discrete counterparts of the dynamical systems (4.19) on $\mathcal{V}$ read

$$\dot{f}(x,t) = \sum_{y \in \mathcal{N}(x)} A_{\omega,f}^\phi(x,y) f(y), \qquad \dot{f}(x,t) = \sum_{y \in \mathcal{N}(x)} B_{\omega,f}^\phi(x,y) f(y), \tag{4.21}$$

where the mappings $A_{\omega,f}^\phi, B_{\omega,f}^\phi \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ represent explicit expressions of the right-hand sides of (4.19) on $\mathcal{V}$,

$$A_{\omega,f}^\phi(x,y) = \begin{cases} 2\omega(x,y)\left(\phi'(|\nabla_\omega f(y,t)|^2)(y) + \phi'(|\nabla_\omega f(x,t)|^2)(x)\right), & x \neq y, \\ -2 \sum\limits_{\substack{z \in \mathcal{N}(x) \\ z \neq x}} \omega(x,z)\left(\phi'(|\nabla_\omega f(z,t)|^2)(z) + \phi'(|\nabla_\omega f(x,t)|^2)(x)\right), & x = y, \end{cases} \tag{4.22a}$$

$$B_{\omega,f}^\phi(x,y) = \begin{cases} 4\omega(x,y)\phi'\left((f(z,t) - f(x,t))^2\omega(x,y)\right), & x \neq y, \\ -4 \sum\limits_{\substack{z \in \mathcal{N}(x) \\ z \neq x}} \omega(x,z)\phi'\left((f(z,t) - f(x,t))^2\omega(x,y)\right), & x = y. \end{cases} \tag{4.22b}$$

Depending on the specification of $\phi(t)$, the dynamics governed by the systems (4.21) define nonlinear nonlocal diffusion processes with various smoothing properties according to

the mappings (4.22). Specifically, for $\phi(t) = t$, the functionals (4.17) coincide, as do the systems (4.21), since the mappings (4.22) do not depend on $f(x,t)$, but only on ω, which is symmetric and nonnegative, and hence agree. Invoking Lemma 3.1 with $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ defined through (4.22), setting $\Theta, \alpha \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ by $\Theta(x,y) = 1, \alpha^2(x,y) = 4\omega(x,y), x \neq y$, and $\Theta(x,x) = -4\sum_{y\in\mathcal{N}(x)} \omega(x,y), x \in \mathcal{V}$, yields the decomposition (3.3), which characterizes (4.18) in terms of the nonlocal operators from section 2.1 if $f_{|\mathcal{V}_I^\alpha} = 0$ holds, by means of relation (3.4). Consequently, (4.21) admits the representation by (4.16) for the particular case of zero nonlocal Dirichlet conditions.

While the above approaches are well suited for image denoising and inpainting, our *geometric* approach performs *labeling* of arbitrary metric data on arbitrary graphs.

4.5. Nonlocal graph Laplacians. Elmoataz, Toutain, and Tenbrinck [5] studied discrete nonlocal differential operators on weighted graphs $(\mathcal{V}, \mathcal{E}, \omega)$. Specifically, based on the nonlocal gradient operator (4.18), a class of Laplacian operators acting on functions $f \in \mathcal{F}_{\mathcal{V}}$ was defined by

(4.23a)
$$\mathcal{L}_{\omega,p} f(x) = \begin{cases} \beta^+(x) \sum\limits_{y\in\mathcal{N}^+(x)} (\nabla_\omega f(x,y))^{p-1} + \beta^-(x) \sum\limits_{y\in\mathcal{N}^-(x)} (-1)^p (\nabla_\omega f(x,y))^{p-1}, & p \in [2,\infty), \\ \beta^+(x) \max\limits_{y\in\mathcal{N}^+(x)} (\nabla_\omega f(x,y)) + \beta^-(x) \max\limits_{y\in\mathcal{N}^-(x)} (-1)^p (\nabla_\omega f(x,y)), & p = \infty, \end{cases}$$

where

(4.23b)
$$\mathcal{N}^+(x) = \{y \in \mathcal{N}(x) : f(y) - f(x) > 0\}, \quad \mathcal{N}^-(x) = \{y \in \mathcal{N}(x) : f(y) - f(x) < 0\}.$$

As detailed in [5, section 4] depending on the weighting function $\omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ and on the positive functions $\beta^+, \beta^- \in \mathcal{F}_{\mathcal{V}}$ satisfying $\beta^+(x) + \beta^-(x) = 1, x \in \mathcal{V}$, the Laplacians (4.23) enable generalizing a broad class of variational approaches including [2], whose Euler Lagrange equations involve graph Laplacians.

In the following, we focus on undirected graphs $(\mathcal{V}, \mathcal{E}, \omega)$ with $\omega(x,y) = \omega(y,x)$. Then, for the purpose of data inpainting and following [5], given a vertex set $\mathcal{A} \subset \mathcal{V}$ together with a function $g \in \mathcal{F}_{\partial\mathcal{A},\mathbb{R}^c}$ specifying the boundary condition imposed on

(4.24)
$$\partial\mathcal{A} = \{x \in \mathcal{V} \setminus \mathcal{A} : \exists y \in \mathcal{A} \text{ with } y \in \mathcal{N}(x)\},$$

the nonlocal Laplacian (4.23) generates a family of nonlocal discrete diffusion processes of the form

(4.25a)
$$\partial_t f(x,t) = \mathcal{L}_{\omega,p} f(x,t) \quad \text{on } \mathcal{A}\times\mathbb{R}_+,$$

(4.25b)
$$f(x,t) = g(x,t) \quad \text{on } \partial\mathcal{A}\times\mathbb{R}_+,$$

(4.25c)
$$f(x,0) = f_0(x) \quad \text{on } \mathcal{A}.$$

To establish a comparison with the proposed nonlocal formulation (3.7), we represent the model (4.25) with $g = 0$ on $\partial\mathcal{A}$ in terms of the operators introduced in section 2.1. Following [5, section 5] and setting the weighting function

(4.26)
$$\alpha^f(x,y) = \begin{cases} \beta^+(x)\sqrt{\omega(x,y)}^{p-1} (\nabla_\omega f(x,y))^{p-2} & \text{if } f(y) > f(x), \\ \beta^-(x)\sqrt{\omega(x,y)}^{p-1} (\nabla_\omega f(y,x))^{p-2} & \text{if } f(y) < f(x), \end{cases}$$

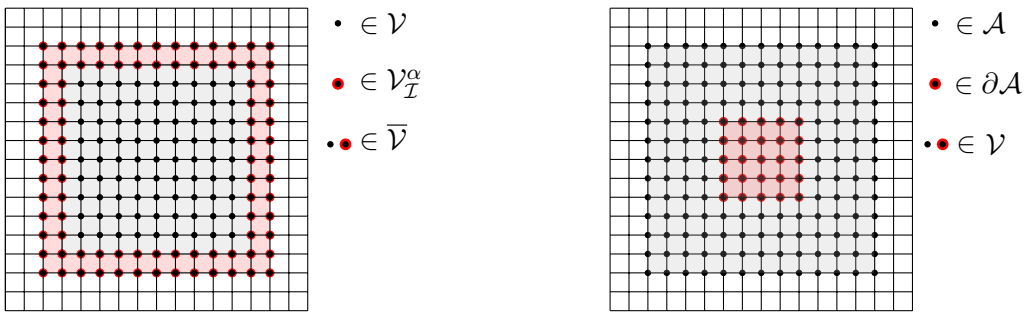
the particular case $p=2$ simplifies to a linear diffusion process (2.18) with (4.26) directly given in terms of weights $\omega(x,y)$ prescribed by the adjacency relation of the graph $\mathcal{V}$. Moreover, if at each vertex $x \in \mathcal{V}$ the equation $\beta^+(x) = \beta^-(x) = \frac{1}{2}$ holds, then for any $p \in [2,\infty)$ the mapping (4.26) is nonnegative and symmetric. As a consequence, α^f from (4.26) can substitute $\omega(x,y)$ in (2.18) and hence specifies a representation of the form (2.17) when choosing the antisymmetric mapping $\alpha \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ to satisfy $2\alpha^2(x,y) = \alpha^f(x,y)$. Finally, specifying the symmetric mapping $\Theta \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ as $\Theta(x,y) = 1$ if $x \neq y$ and $\Theta(x,x) = -\sum_{y\in\mathcal{N}(x)} \alpha^2(x,y)$ expresses the system (4.25) through (4.16) with $\mathcal{V}$ and $\mathcal{V}_{\mathcal{I}}^{\alpha}$ given by $\mathcal{A}$ and $\partial\mathcal{A}$, respectively.

We conclude with a comment similar to the previous sections. While the similarity of the above mathematical structures to our approach is evident from the viewpoint of diffusion processes, the scope of our approach, data labeling, differs and is not directly addressed by established diffusion-based approaches. We further point out the different role of interaction domain (2.8). While for model (4.25) we set α through (4.26) to satisfy $\mathcal{V}_{\mathcal{I}}^{\alpha} = \partial\mathcal{A}$ which is subset of given set of vertices $\mathcal{V}$, i.e., $\overline{\mathcal{V}} = \mathcal{V}$ as illustrated by the right panel of Figure 10, we focus in our work on mappings α that lead to an *extension* of $\mathcal{V}$ by vertices in $\mathbb{Z}^d \setminus \mathcal{V}$, as presented by the left panel of Figure 10.

5. Nonconvex optimization by geometric integration. We show in section 5.1 how geometric integration provides a numerical scheme for solving the nonlocal partial difference equation (3.7) on a regular discrete grid $\mathcal{V}$ by generating a sequence of states on $\mathcal{W}$ that monotonically decrease the energy objective (2.46). In particular, we show that the geometric Euler scheme is equivalent to the basic two-step iterative approach provided by [33] for solving nonconvex optimization problems in DC format.

In section 5.2, we prove the monotonic decrease property for a *novel* class of geometric *multistage* integration schemes that speed up convergence and show the relation of this class to the nonconvex optimization framework presented in [46, 47]

Figure 11 provides a schematic overview over key components of the two proposed al-



nonlocal G-PDE (3.7) **nonlocal approach** [24]

Figure 10. *Schematic illustration of two different instances of $\mathcal{V}_{\mathcal{I}}^{\alpha}$. Nodes $\bullet$ and $\bullet$ represent points of the interaction domain $\mathcal{V}_{\mathcal{I}}^{\alpha}$ and the vertex set $\mathcal{V}$, respectively, in terms of the mapping $\alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$. Left: Boundary configuration for the nonlocal G-PDE (3.7) introduced in this paper. Nonzero interaction of nodes in $\mathcal{V}$ with nodes outside the graph $\mathbb{Z}^d\setminus\mathcal{V}$ results in an extended domain $\overline{\mathcal{V}}$ according to (2.6). Right: Boundary configuration for the task of inpainting as proposed in [5]. The parameter α is specified entirely on $\mathcal{V}$ resulting in a disjoint decomposition $\mathcal{V} = \mathcal{A}\dot{\cup}\partial\mathcal{A}$ where now $\mathcal{V}_{\mathcal{I}}^{\alpha}$ satisfies $\mathcal{V}_{\mathcal{I}}^{\alpha} = \partial\mathcal{A}$ to represent the set of all nodes with missing information $\mathcal{V}\setminus\mathcal{A}$.*

gorithms, including references to the corresponding subsections. Proofs are provided in section A.4 to enable efficient reading.

5.1. First-order geometric integration and DC programming. We focus on a one-stage iterative numerical scheme derived by discretizing the explicit geometric Euler integration (3.19) in time with a fixed time step size $h > 0$. In this specific case, (3.19) generates the sequence of iterates for approximately solving (2.38a) given by

(5.1)
$$(S^k)_{k\geq 1} \subset \mathcal{F}_{\mathcal{V},\mathcal{W}}, \quad S^{k+1}(x) = \exp_{S^k(x)}\left(h(\Omega S)(x)\right), \quad S^0(x) = \exp_{\mathbb{1}_c}\left(-\frac{D\mathcal{X}(x)}{\rho}\right), \quad x \in \mathcal{V},$$

where the index k represents the point in time kh. We next show that the sequence (5.1) locally minimizes the potential (2.46) and hence, based on the formulation derived in Proposition 3.3, how geometric integration provides a finite difference scheme for numerically solving the nonlocal G-PDE (3.7) for the particular case of zero nonlocal boundary conditions.

Proposition 5.1. *Let $\alpha, \Theta \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}, \lambda \in \mathcal{F}_{\mathcal{V}}$, and $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ be given as in Lemma 3.1. Then the sequence (5.1) satisfies*

$$S^{k+1}(x) = \exp_{S^k(x)}\left(h\left(\frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(h\overline{S}^k)\right) + \lambda\overline{S}^k\right)(x)\right), \qquad x \in \mathcal{V}, \tag{5.2}$$

where the zero extension $\overline{S}^k$ of S^k to $\overline{\mathcal{V}}$ is a discrete approximation $S(hk)$ of the continuous time solution to the system (3.7), initialized by $S^0(x)$ (5.1) with imposed zero nonlocal boundary

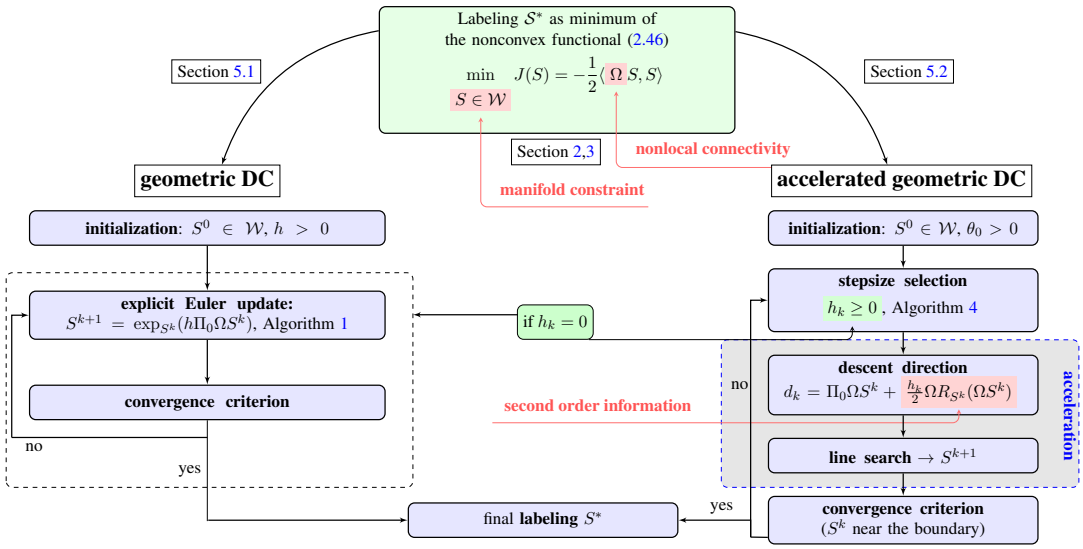


Figure 11. *Sketch of the two algorithmic schemes, Algorithms 1 and 4, developed in section 5. Common basic components as well as essential differences are highlighted. The major difference corresponds to the acceleration of the basic numerical scheme by geometric integration for solving the nonconvex DC program displayed in the top box.*

Algorithm 1 Geometric DC programming scheme.

Initialization: $\gamma > |\lambda_{\min}(\Omega)|$ (DC-decomposition parameter, see proof of Proposition 5.1)
$S^0 = S(0) \in \mathcal{W}$ (initial point by (2.38a))
$\epsilon > 0$ (termination threshold)
$\epsilon_0 = \| \operatorname{grad}_g J(S^0) \| \quad (\operatorname{grad}_g J(S) = R_S(\partial_S J(S)))$
$k = 0$
while $\epsilon_k > \epsilon$ **do**
  $\widetilde{S}^k = \Omega S^k + \gamma \log S^k$
  compute: $S^{k+1} = \operatorname{argmin}_{S \in \overline{\mathcal{W}}} \{\gamma S \log S - \langle \widetilde{S}^k, S \rangle\}$ given by (5.1), resp., (5.2) with $h = \frac{1}{\gamma}$
  $\epsilon_k = \| \operatorname{grad}_g J(S^{k+1}) \|$
  $k \leftarrow k + 1$

conditions. In addition, if

$$h \leq \frac{1}{|\lambda_{\min}(\Omega)|}, \tag{5.3}$$

where $\lambda_{\min}(\Omega)$ *denotes the smallest eigenvalue of* Ω*, then the sequence* (S^k) *achieves the monotone decrease property*

$$J(S^{k+1}) \leq J(S^k), \qquad k \in \mathbb{N}, \tag{5.4}$$

for the potential function (2.46).

Proof. See section A.3 for the proof.

Recent work [19] on the convergence of (2.38a) showed that, up to negligible situations that cannot occur when working with real data, limit points $S^* = \lim_{t\to\infty} S(t)$ of (2.38a) are *integral* assignments $S^* \in \overline{\mathcal{W}}$. Proposition 5.1 says that for step sizes $h < 1$ the geometric integration step (5.1) yields a *descent* direction for moving $S(t) \in \mathcal{W}$ to $S(t+h) \in \mathcal{W}$ and therefore sufficiently approximates the integral curve corresponding to (2.38a) at time $t + h$. We conclude that the fixed point determined by Algorithm 1 solves the nonlocal G-PDE (3.7).

5.2. Higher-order geometric integration. In this section we show how higher-order geometric integration schemes can be used and we enhance the first-order method of the previous section.

We continue the discussion of the numerical integration of the assignment flow (2.38a) by employing the tangent space parameterization (3.10). For a discussion of relations to the geometry of $\mathcal{W}$, we refer to [20]. In what follows, we drop the argument $x \in \mathcal{V}$ and just work with matrix products (cf. (2.48)), besides the lifting map $\exp_S$ that acts rowwise as defined by (2.40).

Our starting point is the explicit geometric Euler scheme (3.19) and (5.1), respectively,

$$S(t+h) \approx \exp_{S^0}\left(V(t) + h\dot{V}(t)\right) = \exp_{S(t)}\left(h(\Omega S)(t)\right). \tag{5.5}$$

Now compute the second-order derivative of all component functions on $\mathcal{T}_0$,

$$\ddot{V}(t) \overset{(3.13)}{=} \Pi_0 \Omega \frac{\mathrm{d}}{\mathrm{d}t} \exp_{S^0}(V(t)) \overset{(3.12)}{\overset{(3.10)}{=}} \Pi_0 \Omega R_{\exp_{S^0}(V(t))} \dot{V}(t) \overset{(3.10)}{=} \Pi_0 \Omega R_{S(t)} (\Omega S(t)) . \tag{5.6}$$

Then the second-order expansion $V(t+h) = V(t) + h\dot{V}(t) + \frac{h^2}{2}\ddot{V}(t) + \mathcal{O}(h^3)$ in $\mathcal{T}_0$ leads to the second-order geometric integration scheme

$$S(t+h) \approx \exp_{S(t)} \left(h\dot{V}(t) + \frac{h^2}{2} \ddot{V}(t) \right) \tag{5.7a}$$

$$= \exp_{S(t)} \left(h\Omega S(t) + \frac{h^2}{2} \Omega R_{S(t)}(\Omega S(t)) \right), \tag{5.7b}$$

which may be read due to (2.44a) as the *two-stage iterative algorithm*

$$\widetilde{S}(t) = \exp_{S(t)} (h\Omega S(t)) , \tag{5.8a}$$

$$S(t+h) = \exp_{\widetilde{S}(t)} \left(\frac{h^2}{2} \Omega R_{S(t)}(\Omega S(t)) \right) . \tag{5.8b}$$

Below, we set in view of (3.10)

$$J(V) := J(S)|_{S=\exp_{S^0}(V)} = J(\exp_{S^0}(V)) \tag{5.9}$$

to simplify the notation. The following lemma prepares our main result.

Lemma 5.2. *Based on the parametrization* (3.10)*, the Euclidean gradient of the function* $V \mapsto J(V)$ *is given by*

$$\partial J(V) = -R_{\exp_{S^0}(V)} (\Omega \exp_{S^0}(V)) = \mathrm{grad}_g J(S), \tag{5.10}$$

that is, by the Riemannian gradient of the potential (2.46)*.*

Proof. See section A.4 for the proof.

The next proposition asserts that applying the second-order geometric integration scheme (5.8) leads to a sufficient decrease of the sequence of values $(J(S^k))_{k\in\mathbb{N}}$ if at each iteration the step sizes are chosen according to a *Wolfe rule* like the line search procedure [48, 49]. Specifically, the step sizes h and h^2 in (5.8a) and (5.8b), respectively, are replaced by step size sequences $(\theta_k)_{k\geq 0}$ and $(h_k\theta_k)_{k\geq 0}$. In addition, the proposition reveals that, under mild assumptions on the sequence $(h_k)_{k\geq 0}$, the norm of the Riemannian gradient (5.10) becomes arbitrarily small. The proposition is proved in section A.4.

Proposition 5.3. *Let* $\Omega(x,y)$ *be as in Lemma* 3.1 *and let* $d : \mathcal{W} \times \mathbb{R}_+ \to \mathcal{T}_0$ *be a mapping given by*

$$d(S,h) = \Pi_0 \left(\Omega S + \frac{h}{2} \Omega R_S(\Omega S) \right), \qquad S \in \mathcal{W}, \quad h \in \mathbb{R}_+. \tag{5.11}$$

Then the following holds:

(i) *There exist sequences* $(h_k)_{k\geq 0}, (\theta_k)_{k\geq 0}$ *and constants* $0 < c_1 < c_2 < 1$ *such that setting*

$$S^{k+\frac{1}{2}} = \exp_{S^k}(\theta_k \Omega S^k), \tag{5.12a}$$

$$S^{k+1} = \exp_{S^{k+\frac{1}{2}}}\left(\frac{h_k \theta_k}{2} \Omega R_{S^k}(\Omega S^k)\right), \tag{5.12b}$$

and

$$d^k := d(S^k, h_k) \in \mathcal{T}_0 \tag{5.13}$$

yields iterates

$$S^{k+1} = \exp_{S^k}(\theta_k d^k), \qquad k \in \mathbb{N}, \tag{5.14}$$

satisfying

$$J(S^{k+1}) - J(S^k) \leq c_1 \theta_k \langle \mathrm{grad}_g J(S^k), R_{S^k}(d^k)\rangle_{S^k} \quad \text{(Armijo condition)}, \tag{5.15a}$$

$$|\langle \mathrm{grad}_g J(S^{k+1}), R_{S^k}(d^k)\rangle_{S^k}| \leq c_2 |\langle \mathrm{grad}_g J(S^k), R_{S^k}(d^k)\rangle_{S^k} \quad \text{(curvature condition)} \tag{5.15b}$$

and (recall (2.22)*),*

$$\langle U, V\rangle_S = \sum_{x\in\mathcal{V}} g_{S(x)}\left(U(x), V(x)\right), \qquad U, V \in \mathcal{T}_0, \quad S \in \mathcal{W}. \tag{5.16}$$

(ii) *Suppose the limit point* γ_* *of* $(\theta_k)_{k\geq 0}$ *is bounded away from zero, i.e.,* $\gamma_* = \lim_{k\to\infty} \theta_k > 0$*. Then any limit point* $S^* \in \overline{\mathcal{W}}$ *of the sequence* (5.12) *is an equilibrium of the flow* (2.38a)*.*

(iii) *If* S^* *is a limit point of* (5.12) *which locally minimizes* $J(S)$*, with sequences* $(\theta_k)_{k\geq 0}, (h_k)_{k\geq 0}$ *as in* (ii)*, then* $S^* \in \overline{\mathcal{W}} \setminus \mathcal{W}$*.*

(iv) *If additionally* $\sum_{k\geq 0} h_k = 0$ *holds in* (ii)*, then the sequence* $(\epsilon_k)_{k\geq 0}$ *with* $\epsilon_k := \|\mathrm{grad}_g J(S^k)\|$ *is a zero sequence.*

Proof. See section A.4 for the proof.

Given a state $S^k \in \mathcal{W}$, Proposition 5.3 asserts the existence of step size sequences $(h_k)_{k\geq 0}, (\theta_k)_{k\geq 0} \subset \mathbb{R}_+$ that guarantee a sufficient decrease of the objective (2.46) through (5.14) while still remaining numerically efficient by avoiding too small step sizes through (5.15). A corresponding proper step size selection procedure is summarized as Algorithm 3 that calls Algorithm 2 as a subroutine. Based on Algorithm 3, the two-stage geometric integration scheme (5.8) that *accelerates* Algorithm 1 is given as Algorithm 4. Acceleration is accomplished by utilizing at each S^k descent directions d_k given by (5.13), based on second-order information provided by the vector field (5.6).

In section 6, we show that Algorithm 4 converges. This implies, in particular, that Algorithms 1 and 4 terminate after a finite number of steps for any termination parameter ε with respect to the entropy of the assignment vectors, which measures closeness to an integral

Algorithm 2 Search $(S^k, \theta_k, d_k, c_1, c_2, a, b)$.

Input: current iterate: $S^k \in \mathcal{W}$, initial step size $\theta_k > 0$,
descent direction d_k with $\langle \mathrm{grad}_g J(S^k), R_{S^k} d_k \rangle_{S^k} < 0$,
$k = 1$.
repeat
  $S^{k+1} = \exp_{S^k}(\theta_k d^k)$
  if $J(S^{k+1}) - J(S^k) > \theta_k c_1 \langle \mathrm{grad}_g J(S^k), R_{S^k} d_k \rangle_{S^k}$ **then**
    $a = a, b = \theta_k$
  else
    if $|\langle \mathrm{grad}_g J(S^{k+1}), R_{S^k} d_k \rangle_{S^k}| \leq |c_2 \langle \mathrm{grad}_g J(S^k), R_{S^k} d_k \rangle_{S^k}|$ **then**
      stop
    $a = \theta_k, b = b, \theta_{k+1} = \frac{a+b}{2}$.
  $k \leftarrow k + 1$.
until θ_k *satisfies* (5.15);
Return: S^k, θ_k

Algorithm 3 Step $(S^k, \theta_k, d_k, c_1, c_2, \lambda_{\min}(\Omega))$.

Input: current iterate: $S^k \in \mathcal{W}$, initial step size $\theta_k > 0$,
descent direction d_k with $\langle \mathrm{grad}_g J(S^k), R_{S^k} d_k \rangle_{S^k} < 0$,
smallest eigenvalue of Ω, $\lambda_{\min}(\Omega)$ $c_1, c_2 \in (0, 1)$ with $c_2 \in (c_1, 1)$,
initial search interval: $a_1 = \theta_k, b_1 = \frac{1}{|\lambda_{\min}(\Omega)|}$ with $a_1 < b_1$,
$k = 1$.
repeat
  $\theta_k = \frac{a_k + b_k}{2}, S^{k+1} = \exp_{S^k}(\theta_k d_k)$,
  if $J(S^{k+1}) - J(S^k) > \theta_k c_1 \langle \mathrm{grad}_g J(S^k), R_{S^k} d_k \rangle_{S^k}$ **then**
    $S^{k+1}, \theta_{k+1} \leftarrow \mathrm{Search}(S^k, \theta_k, c_1, c_2, a_k, b_k)$ (Algorithm 2), **stop**
  else
    if $|\langle \mathrm{grad}_g J(S^{k+1}), R_{S^k} d_k \rangle_{S^k}| \leq |c_2 \langle \mathrm{grad}_g J(S^k), R_{S^k} d_k \rangle_{S^k}|$ **then**
      stop
    else
      $a_{k+1} = \theta_{k+1}, b_{k+1} = b_k$.
  $k \leftarrow k + 1$.
until θ_k *satisfies* (5.15a);
Return: S^k

solution. Theorem 6.6 asserts the existence of basins of attraction around integral solutions from which the sequence (S^k) can never escape once it has reached such a region.

We elaborate in terms of Theorem 6.4 a theoretical guideline for choosing a sequence $(h_k)_{k\geq 0}$ which meets the condition of Proposition 5.3(iv). In practice, to achieve an acceleration by Algorithm 4 in comparison with Algorithm 1, we choose a large value of the step size parameter h_k in the beginning and monotonically decrease h_k to zero after a fixed number of iterations. One particular step size selection strategy that we used for the numerical experiments will be highlighted in section 7.

The following remark clarifies how the line search procedure formulated as Algorithm 3, which is used in Algorithm 4, differs from the common line search accelerated DC programming schemes proposed by [46] and [47].

Remark 5.4 (directly related work). Using the notation of Proposition 5.1 and its proof, the step iterated by Algorithm 1 at $S^k \in \mathcal{W}$ reads

$$\widetilde{S}^k = \operatorname{argmin}_{S\in\mathbb{R}^n} \left\{ h^*(S) - \langle S^k, S\rangle \right\} \qquad \text{with} \qquad h(S) = \langle S, \Omega S\rangle + \gamma S\log S, \tag{5.17a}$$

$$S^{k+1} = \operatorname{argmin}_{S\in\mathbb{R}^n} \left\{ g(S) - \langle S, \widetilde{S}^k\rangle \right\} \qquad \text{with} \qquad g(S) = \delta_{\overline{\mathcal{W}}}(S) + \gamma S\log S, \tag{5.17b}$$

where h^* is the conjugate of the convex function h. Motivated by the work [46], Aragón Artacho, Fleming, and Vuong [47] proposed an accelerated version of the above scheme by performing an additional line search step along the descent direction

$$\widetilde{d}^k = S^{k+1} - S^k \tag{5.18}$$

in (5.17b) for scenarios where the primary variable S to be determined is *not* manifold-valued.

The direct comparison with Algorithm 1 reveals that for the specific choice $h_k = 0, k \in \mathbb{N}$, in (5.13), (5.11), line search is performed along the descent direction

$$d^k = \Pi_0 \Omega S^k = V^{k+1} - V^k \in \mathcal{T}_0, \tag{5.19}$$

where the last equation follows from applying the parametrization (3.10) to (5.12) while taking into account (2.41) and $R_S = R_S\Pi_0$ for $S \in \mathcal{W}$.

Comparing $\widetilde{d}^k$ and d^k shows the geometric nature of our algorithm in order to handle properly the manifold-valued variable S and the more general descent directions d^k with step sizes $h_k > 0$ in Algorithm 4.

5.3. Influence of nonlocal boundary conditions.

We conclude this section by explaining in more detail the effect of imposing in (3.7) the zero nonlocal boundary condition on the nonempty interaction domain, on the step size selection procedure presented as Algorithm 3. This explanation is formulated as Remark 5.6 below after the following proposition, which states a result analogous to [37, Proposition 2.3]. The proposition is proved in section A.5.

Proposition 5.5. *For mappings $\Theta, \alpha \in \mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$, let $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ and $\lambda \in \mathcal{F}_{\mathcal{V}}$ be given as in Lemma 3.1 such that property (2.34) holds and $\lambda = 1, x \in \mathcal{V}$ in (3.5) is achieved. Assume further that the weighted graph $(\mathcal{V}, \mathcal{E}, \Omega)$ in (2.1) is connected. Then the following holds:*

Algorithm 4 Accelerated geometric DC-optimization

Initialization: (DC-decomposition parameter, see the proof of Proposition 5.1),
$S^0 = S(0) \in \mathcal{W}$, (initial iterate (2.38a)),
$\epsilon > 0$, (termination threshold),
$\lambda_{\min}(\Omega)$, (smallest eigenvalue of Ω),
$c_1, c_2 \in (0,1)$ (cf. Proposition 5.3),
$\epsilon_0 = \|\mathrm{grad}_g J(S^0)\|$, $\theta_0 = \frac{1}{\gamma}$ (cf. (A.10))
$k = 0$.
while $\epsilon_k > \epsilon$ **do**
  Choose: $h_k \in \left(0, \frac{\|R_{S^k}(\Omega S^k)\|^2_{S^k}}{|\langle R_{S^k}(\Omega S^k), \Omega R_{S^k}(\Omega S^k)\rangle|}\right)$
  $d_k = \Pi_0 \Omega S^k + \frac{h_k}{2} \Omega R_{S^k}(\Omega S^k)$ (descent direction by (5.13), (5.11))
  if θ_k *satisfies* (5.15) **then**
    Set: $\widetilde{S}^k = \frac{1}{\theta_k} \log(\frac{S^k}{\mathbb{1}_c}) + d_k$
    Compute: $S^{k+1} = \mathrm{argmin}_{S \in \overline{\mathcal{W}}} \{\frac{1}{\theta_k} S \log S - \langle \widetilde{S}^k, S\rangle\}$, by
    $S^{k+1} = \exp_{S^k}(\theta_k d^k)$
  else
    $S^{k+1} \leftarrow \mathrm{Step}\big(S^k, \theta_k, d_k, c_1, c_2, \lambda_{\min}(\Omega)\big)$ by Algorithm 3.
  $\epsilon_{k+1} = \|\mathrm{grad}_g J(S^{k+1})\|$,
  $k \leftarrow k+1$.
Returns: $S^k \approx S^*$

(i) *The smallest Dirichlet eigenvalue of the nonlocal operator* (2.17)

$$\lambda_1^D = \inf_{f \neq 0} -\frac{\frac{1}{2}\langle f, \mathcal{D}^\alpha(\Theta \mathcal{G}^\alpha f)\rangle_{\overline{\mathcal{V}}}}{\langle f, f\rangle_{\overline{\mathcal{V}}}}, \qquad f \in \mathcal{F}_{\overline{\mathcal{V}}}, \quad f_{|\mathcal{V}_{\mathcal{I}}^\alpha} = 0, \tag{5.20}$$

is bounded away from zero and admits the equivalent expression

$$0 < \lambda_1^D = \inf_{f \neq 0} \frac{\langle f, (\Lambda - \Omega) f\rangle_{\mathcal{V}}}{\langle f, f\rangle_{\mathcal{V}}}, \tag{5.21}$$

where

$$\Lambda = \mathrm{Diag}(\lambda), \qquad \lambda = (\ldots, \lambda(x), \ldots)^\top \tag{5.22}$$

with $\lambda(x)$ *given by* (3.5).

(ii) *One has* $\lambda_{\min}(\Omega) > -1$.

Proof. See section A.5 for the proof.

We are now in position to characterize the effect of imposing the zero nonlocal boundary condition on the step size selection procedure (Algorithm 3).

Remark 5.6 (parameter selection). Recalling the proof of Proposition 5.1, the update (5.2) amounts to performing at each step $k \in \mathbb{N}$ one iteration of a basic DC programming scheme

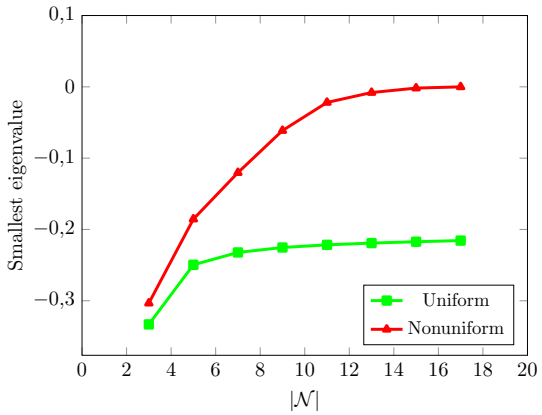


Figure 12. *Effect of imposing nonlocal boundary conditions. The green (●) and red (●) curves plot the smallest eigenvalues* $\lambda_{\min}(\Omega)$ *of the parameter matrix* (3.3) *for uniform and nonuniform averaging, respectively, and for different neighborhood sizes* $|\mathcal{N}|$*. Choosing larger neighborhoods* (2.3) *increases the smallest eigenvalue and consequently, by* (5.3)*, enables us to choose bigger step sizes in Algorithm* 1 *that achieve the monotone decrease property* (5.4)*.*

[33] with respect to the suitable DC-decomposition (A.10) of (2.46), with Ω satisfying (2.2), (2.34) by choosing parameter $\gamma > 0$ such that $\lambda_{\min}(\Omega + \gamma \mathrm{Diag}(\frac{\mathbb{1}}{S})) > 0$. In the case of a *nonzero* interaction domain (2.8) with Ω, α, Θ as in Lemma 5.5, Proposition 5.5(ii) and estimate (A.13) yield for $S \in \mathcal{W}$

$$\lambda_{\min}\left(\Omega + \gamma \mathrm{Diag}\left(\frac{\mathbb{1}}{S}\right)\right) > -1 + \beta + \gamma > 0 \qquad \text{for} \qquad \gamma > 1 - \beta, \tag{5.23a}$$

$$\beta = \sum_{x \in \mathcal{V}_b} \sum_{y \in \mathcal{V}_{\mathcal{I}}^{\alpha}} \Theta(x,y) \alpha^2(x,y) f^2(x). \tag{5.23b}$$

In particular, following the steps in the proof of Lemma 5.1, relation $h = \frac{1}{\gamma}$ in connection with (5.23) accounts for bigger step sizes in Algorithm 1 for integrating (3.7) with nonzero interaction domain (2.8). This will be numerically validated in section 7 (see Figure 12).

We conclude this section with a final comment on the lower bound of the objective (2.46).

Remark 5.7 (global minimizer of (2.46)). Recalling the terms involved in the objective (2.46), the lower bound is attained precisely when the first term $\sum_{x \in \mathcal{V}} \sum_{y \in \mathcal{N}(x)} \Omega(x,y) \|S(x) - S(y)\|^2$ is minimal and the last term $-\frac{1}{2}\|S\|_F^2$ is maximal. Therefore the global minimizers of $J(S)$ are given by the set of spatially constant assignments, where to each node in graph $\mathcal{V}$ the same prototype $X_j^* \in \mathcal{X}$ is assigned.

6. Convergence analysis.

This section is devoted to the convergence analysis of Algorithm 4 that performs accelerated geometric integration of the Riemannian descent flow (2.38a). The main results are stated as Theorems 6.4 and 6.6 in section 6.2. The lengthy proofs have been relegated to section A.6.

6.1. Preparatory lemmata.

Lemma 6.1. *For a nonnegative, symmetric mapping $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$, let the sequences $(S^k)_{k\geq 0}$, $(\theta_k)_{k\geq 0}, (h_k)_{k\geq 0}$ be recursively defined by Algorithm 4 and let Λ denote the set of all limit points of the sequence $(S^k)_{k\geq 0}$,*

$$\Lambda = \{S \in \overline{\mathcal{W}} : \exists (S^{k_l})_{l\geq 0} \text{ with } S^{k_l} \to S \text{ for } l \to \infty\}. \tag{6.1}$$

Then there exists $J^* \in \mathbb{R}$ with $\lim_{k\to\infty} J(S^k) = J^*$, i.e., $J(S)$ is constant on Λ.

Proof. See section A.6 for the proof.

Next, we inspect the behavior of the iterates generated by Algorithm 4 near a limit point $S^* \in \overline{W}$. To this end, the following index sets are considered at each node $x \in \mathcal{V}$:

$$J_+(S^*(x)) = \{j \in [c] : (\Omega S^*)_j(x) - \langle S^*(x), (\Omega S^*)(x)\rangle < 0\}, \tag{6.2a}$$

$$J_-(S^*(x)) = \{j \in [c] : (\Omega S^*)_j(x) - \langle S^*(x), (\Omega S^*)(x)\rangle > 0\}, \tag{6.2b}$$

$$J_0(S^*(x)) = \{j \in [c] : (\Omega S^*)_j(x) - \langle S^*(x), (\Omega S^*)(x)\rangle = 0\}. \tag{6.2c}$$

Lemma 6.2. *Let $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ and $(S^k)_{k\geq 0}, (\theta_k)_{k\geq 0}, (h_k)_{k\geq 0}$ be as in Proposition 5.3(iv) with a sequence $(\theta_k)_{k\geq 0}$ bounded by $\theta_k \in [\theta_{\min}, \theta_{\max}]$. Let $S^* \in \overline{\mathcal{W}}$ be a limit point of $(S^k)_{k\geq 0}$. Then, for the positive function $Q(S) = \sum_{x\in\mathcal{V}} \sum_{j\in J_+(S^*(x))} S_j(x)$, there are constants $\varepsilon > 0$, $M^* > 1$ and an index k_0 such that $\forall\, k \geq k_0$ with $\|S^* - S^k\| < \varepsilon$ the inequality*

$$Q(S^{k+1}) - Q(S^k) < \frac{\theta_k}{M^*} \sum_{x\in\mathcal{V}} \sum_{j\in J_+(S^*(x))} S_j^k(x)((\Omega S^*)_j(x) - \langle \Omega S^*(x), S^*(x)\rangle) < 0 \tag{6.3}$$

is satisfied.

Proof. See section A.6 for the proof.

6.2. Main results. This section provides the main results of our convergence analysis: convergence of the accelerated Algorithm 4 (Theorem 6.4) and an estimate of the basins of attraction around equilibria that enable early stopping of Algorithm 4 (Theorem 6.6).

Definition 6.3 (convex functions of Legendre type [50, Chapter 26]). *Let $f : X \to (-\infty, \infty]$ be a lower-semicontinuous proper convex function with nonempty open domain $C = \operatorname{int}(\operatorname{dom} f) \neq \emptyset$. Then f is called*

(i) *essentially smooth if f is differentiable on C and for every sequence $(x_k)_{k\in\mathbb{N}} \subset C$ with $x_k \to x^* \in \overline{C} \setminus C$ converging to a boundary point for $k \to \infty$, it follows that $\|\nabla f(x_k)\| \to \infty$;*

(ii) *a Legendre type function if h is essentially smooth and strictly convex on C.*

Convex functions f of Legendre type yield a class of *Bregman divergence functions* D_f through

$$\begin{aligned} D_f : \overline{C} \times C &\to \mathbb{R}_+, \\ (x, y) &\mapsto f(x) - f(y) - \langle \nabla f(y), x - y\rangle; \end{aligned} \tag{6.4}$$

see, e.g., [51, 52] for a detailed exposition. Strict convexity of f and Jensen's inequality imply

$$\forall (x,y) \in \overline{C} \times C\colon \qquad D_f(x,y) \geq 0 \quad \text{and} \quad (D_f(x,y) = 0) \Leftrightarrow (x = y). \tag{6.5}$$

In the following, we will use the *Kullback–Leibler (KL) divergence* (a.k.a. *relative entropy, information divergence*) $D_{\mathrm{KL}} = D_f$,

$$D_{\mathrm{KL}}\colon \overline{\mathcal{S}} \times \mathcal{S} \to \mathbb{R}_+, \qquad D_{\mathrm{KL}}(s,p) = \left\langle s, \log \frac{s}{p} \right\rangle, \tag{6.6}$$

induced by the negative discrete entropy function

$$f = \langle s, \log s \rangle + \delta_{\overline{S}}(s) \tag{6.7}$$

(with the convention $0 \cdot \log 0 = 0$). Accordingly, we define with abuse of notation

$$D_{\mathrm{KL}}\colon \overline{\mathcal{W}} \times \mathcal{W} \to \mathbb{R}_+, \qquad D_{\mathrm{KL}}(S,P) = \sum_{x \in \mathcal{V}} D_{\mathrm{KL}}\left(S(x), P(x)\right). \tag{6.8}$$

Theorem 6.4 (convergence of Algorithm 4). *Let $(S^k)_{k\geq 0}$ be a sequence generated by Algorithm 4, where the sequences of step sizes $(\theta_k)_{k\geq 0}, (h_k)_{k\geq 0}$ additionally satisfy the assumptions of Lemma 6.2 and Proposition 5.3, respectively. If there exists an index $K \in \mathbb{N}$ such that the sequence $(h_k)_{k\geq K}$ satisfies*

$$h_k \leq C(\Omega) \frac{\|\mathrm{grad}_g J(S^k)\|^2_{S^k}}{n} \tag{6.9a}$$

$$\text{with} \qquad C(\Omega) := 2\frac{\theta_{\min} c_1}{\lambda^2(\Omega)}, \qquad \lambda(\Omega) = \max\{|\lambda_{\min}(\Omega)|, |\lambda_{\max}|(\Omega)\}, \tag{6.9b}$$

then the set $\Lambda = \{S^*\}$ defined by (6.1) is a singleton and $\lim_{k\to\infty} D_{\mathrm{KL}}(S^*, S^k) = 0$ holds, i.e., the sequence $(S^k)_{k\geq 0}$ converges to a unique $S^* \in \overline{\mathcal{W}}$ which is an equilibrium of (2.38a).

Proof. See section A.7 for the proof.

According to Proposition 5.3(iii), (iv), the sequence $(S^k)_{k\geq 0}$ converges to a critical point $S^* \in \overline{\mathcal{W}} \setminus \mathcal{W}$ on the boundary of convex set $\overline{\mathcal{W}}$. Since both functions g, h of the DC-decomposition (A.10) have been regularized by the negative entropy, global Lipschitz continuity of the derivatives does *not* hold and hence does not allow us to study the convergence rate of Algorithm 4 along the lines pursued in [47], [53], [54]. Therefore, we confine ourselves to establishing a *local linear* rate of convergence $S^k \to S^*$ within a suitably defined basin of attraction in $\mathcal{W}$ around S^*. To this end, we adopt the following basic assumption.

Assumption. Any stationary point $S^* \in \overline{\mathcal{W}}$ of the sequence (S^k) generated by Algorithm 4 is a stable equilibrium of the flow (2.38a):

$$(\Omega S^*)_j(x) - (\Omega S^*)_{j^*(x)}(x) < 0, \qquad j \in [c] \setminus j^*(x) = \arg\max_{l\in[c]} S_l^*(x), \qquad \forall x \in \mathcal{V}. \tag{6.10}$$

Remark 6.5. As worked out in [19, section 2.3.2], the set of initial points $S(0)$ of the flow (2.38a) for which assumption (6.10) is not satisfied has measure zero. Hence assumption (6.10) holds in all practically relevant cases.

Based on assumption (6.10), we adopt the results reported in [19, section 2.3.3] by defining the open convex polytope for each integral equilibrium $S^* \in \mathcal{W}^*$ as

$$A(S^*) := \bigcap_{x\in\mathcal{V}} \bigcap_{j\neq j^*(x)} \{S \in \mathcal{F}_{\mathbb{R}^{n\times c}} \colon (\Omega S)_j(x) < (\Omega S)_{j^*(x)}(x)\} \tag{6.11}$$

and by introducing the *basins of attraction*

$$B_\varepsilon(S^*) := \{S \in \overline{W} \colon \max_{x\in\mathcal{V}} \|S(x) - S^*(x)\|_1 < \varepsilon\} \subset A(S^*) \cap \overline{W}, \tag{6.12}$$

where $\varepsilon > 0$ is small enough such that the inclusion in (6.12) holds. Due to [19, Proposition 2.3.13] a sufficient upper bound $\varepsilon \leq \varepsilon^*$ for the inclusion (6.12) to hold is

$$\varepsilon^* = \min_{x\in\mathcal{V}} \min_{j\in[c]\setminus j^*(x)} \frac{2\left((\Omega S^*)_{j^*(x)} - (\Omega S^*)_j\right)(x)}{\sum_{y\in\mathcal{N}(x)} \Omega(x,y) + \left((\Omega S^*)_{j^*(x)} - (\Omega S^*)_j\right)(x)} > 0. \tag{6.13}$$

The following theorem asserts that a modified criterion applies to the sequence generated by Algorithm 4, together with a linear convergence rate $S^k \to S^*$, whenever the sequence (S^k) enters a basin of attraction $B_\varepsilon(S^*)$.

Theorem 6.6 (basins of attraction). *For $\Omega \in \mathcal{F}_{\mathcal{V}\times\mathcal{V}}$ as in Lemma 3.1, let $(S^k)_{k\geq 0}$ be a sequence generated by Algorithm 4. Let $S^* \in \overline{\mathcal{W}}$ be a limiting point $(S^k)_{k\geq 0}$ that fulfills assumption (6.10) and let $\varepsilon^* > 0$ be as in (6.13). Then, introducing the positive constants*

$$\overline{h} = \max_{k\in\mathbb{N}} h_k, \quad \rho^* = \max_{S\in\overline{\mathcal{W}}} \Big(\max_{\substack{x\in\mathcal{V},\\ j\in[c]\setminus j^*(x)}} \left((\Omega S)_{j^*(x)} - (\Omega S)_j\right)(x) \Big), \quad N = \max_{y\in\mathcal{V}} |\mathcal{N}(y)|, \tag{6.14}$$

$\forall\, \varepsilon > 0$ small enough such that

$$\varepsilon \leq \min_{x\in\mathcal{V}} \min_{j\in[c]\setminus j^*(x)} \frac{2\cdot\left((\Omega S^*)_{j^*(x)} - (\Omega S^*)_j\right)(x)}{1 + C\cdot\rho^* + \left((\Omega S^*)_{j^*(x)} - (\Omega S^*)_j\right)(x)}, \quad C = \overline{h}\cdot c\cdot N, \tag{6.15}$$

the following applies: If for some index $k_0 \in \mathbb{N}$ it holds that $S^{k_0} \in B_\varepsilon(S^*) \subset B_{\varepsilon^*}(S^*)$, then $\forall\, k \geq k_0$ there exists a mapping $\xi \in \mathcal{F}_\mathcal{V}$ with $\xi(x) \in (0,1), \forall x \in \mathcal{V}$, such that

$$\|S^k(x) - S^*(x)\|_1 < \xi^{k-k_0}(x)\|S^{k_0}(x) - S^*(x)\|_1 \qquad \forall x \in \mathcal{V}. \tag{6.16}$$

Proof. See section A.7 for the proof.

7. Experiments and discussion. In this section, we report numerical results obtained with the algorithms introduced in section 5. Details of the implementation and parameters

settings are provided in section 7.1. Section 7.2 deals with the impact of the nonlocal boundary conditions of system (3.14) on properties of averaging matrices Ω (see section 3) and how this affects the selection of the step size parameter $h > 0$ in Algorithm 1. Section 7.3 reports results obtained by computing the assignment flow with Algorithm 1 and different constant step sizes $h > 0$ using the nonlocal G-PDE parametrization (3.14). In addition, we studied numerical consequences of nonlocal boundary conditions (3.7b), (3.7c) using the maximal allowable step size (5.3) according to Proposition 5.1. Finally, in section 7.4, we compare Algorithm 1 and the accelerated Algorithm 4 by evaluating their respective convergence rates to an integral solution of the assignment flow corresponding to a stationary point of the potential (2.46) for various nonlocal connectivities.

7.1. Implementation details. All evaluations were performed using the noisy image data depicted by Figure 5(b). System (3.7) was initialized by $S^0 = L(\mathbb{1}_{\mathcal{W}}) \in \mathcal{W}$ with $\rho = 1$, as specified by (2.29). Since the iterates (S^k) converge in all cases to integral solutions which are located at vertices on the boundary $\partial\mathcal{W}$ of $\mathcal{W}$, whereas the numerics is designed for evolutions *on* $\mathcal{W}$, we applied the renormalization routine adopted in [17, section 3.3.1] with $\varepsilon = 10^{-10}$ whenever the sequence $(S^k)_{k\geq 0}$ came that close to $\partial\mathcal{W}$ on its path to the vertex.

The averaging matrix Ω was assembled in two ways as specified in section 3.4.2 as items (i) and (ii), called *uniform* and *nonuniform* averaging in this section. In the latter case, the parameter values $\sigma_s = 1, \sigma_p = 5$ were chosen in (3.21), as for the experiments reported in section 3.4.2. The iterative algorithms were terminated at step k when the averaged gradient norm

$$\epsilon_k = \frac{1}{n} \sum_{x \in \mathcal{V}} \| R_{S^k(x)}(\Omega S^k(x)) \| \leq \epsilon \tag{7.1}$$

reached a threshold ϵ which when chosen sufficiently small to satisfy bound (6.15) guarantees a linear convergence rate as specified in Theorem 6.6.

We point out that during the evaluation and discussion of realized experiments our focus was *not* on assessing a comparison of computational speed in terms of absolute runtimes, but on the numerical behavior of the proposed schemes with regard to number of iterations required to solve system (3.14) and in terms of the labeling performance. Thus, we did not confine ourselves to impose any restriction on the minimum time step size and the maximum number of iterations and instead appropriately adjusted the parameter (7.1) to stop the algorithm when a stationary point at the boundary of $\mathcal{W}$ was reached.

Since S^* is unknown, we cannot directly access the exact bound in (6.15) beforehand and therefore it is not evident how to set ϵ in practice. However, based on experimental evidence, setting the termination threshold by $\epsilon = 10^{-7}$ in (7.1) serves as a good estimate; see Figures 16 and 18. Algorithm 3 requires specifying two parameters c_1, c_2 (see line 3). We empirically found that using $c_1 = 0.4, c_2 = 0.95$ is a good choice that we used in all experiments.

7.2. Step size selection. This section reports results of several experiments that highlight aspects of imposing nonlocal boundary conditions (3.7b), (3.7c) and their influence on the selection of step sizes in Algorithms 1 and 4.

To demonstrate these effects we used two different parameter matrices Ω defined in accordance with Lemma 3.1, with Θ, α given as in section 3.4.2, called *uniform* and *nonuniform*

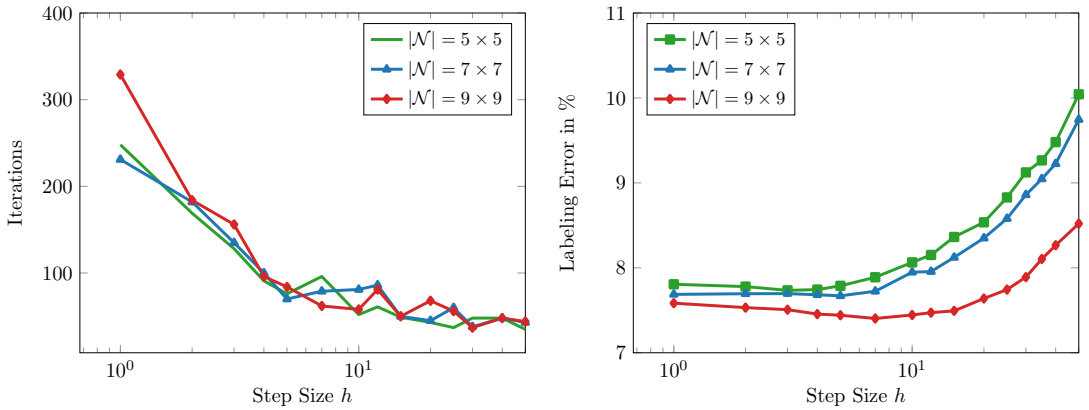


Figure 13. *Effects of selecting the step size h in Algorithm 1 for various neighborhood sizes $|\mathcal{N}|$. Dashed vertical lines indicate the step size upper bound $\frac{1}{|\lambda_{\min}(\Omega)|}$ that guarantees the monotone decrease property (Proposition 5.1). Left: Number of iterations required to satisfy the termination criterion (7.1). Larger step sizes decrease the number of iterations but yield unreliable numerical computation when h exceeds the upper bound (see text). Right: Pixelwise labeling error compared to ground truth. Labeling accuracy quickly deteriorates when h exceeds the upper bound.*

averaging, respectively. To access the maximal bound (5.3) for the step size $h > 0$, as derived in Proposition 5.1 in order to achieve the monotone decrease property (5.4), we directly approximated the exact smallest eigenvalue $\lambda_{\min}(\Omega)$ using available software [55].

Figure 12 displays values of the smallest eigenvalue for uniform and nonuniform averaging, respectively, and different sizes of the nonlocal neighborhoods (2.3): Increasing the size $|\mathcal{N}|$ decreases the value of $\lambda_{\min}(\Omega)$ and consequently, by virtue of relation $h \geq \frac{1}{|\lambda_{\min}(\Omega)|}$ in Proposition 5.1, to a larger upper bound for setting the step size h in Algorithm 1. This confirms our observation and statement formulated as Remark 5.6.

In practice, however, it is too expensive to compute $\lambda_{\min}$ numerically for choosing the step size h. Figure 13 shows the following for three sizes of neighborhoods $|\mathcal{N}|$ and for step sizes h *smaller and larger* than the upper bound (5.3) indicated by dashed vertical lines:

(i) the number of iterations required to reach the termination criterion (7.1) (Figure 13, left panel);

(ii) the labeling accuracy compared to ground truth (Figure 13, right panel).

The results show that the bound (5.3) should be considered as a hard constraint indeed: Increasing the step size h up to this bound (cf. Figure 13, left panel) decreases the required number of iterations, as to be expected. But exceeding the bound yields unreliable computation, possibly caused by a too small DC-decomposition parameter $\gamma < |\lambda_{\min}(\Omega)|$ which compromises the convexity and hence convergence of the auxiliary optimization problems in Algorithm 1, line 8. Likewise, Figure 13, right panel, shows that labelings quickly become inaccurate once the step size exceeds the upper bound. Figure 14 visualizes examples.

Overall, these results show that a wide range of save choices of the step size parameter h exists and that choosing the "best" value depends on how accurate $\lambda_{\min}(\Omega)$ is known beforehand.

7.3. First-order optimization. This section is devoted to the evaluation of Algorithm 1. We examine how effectively this algorithm converges to an integral solution (labeling) for

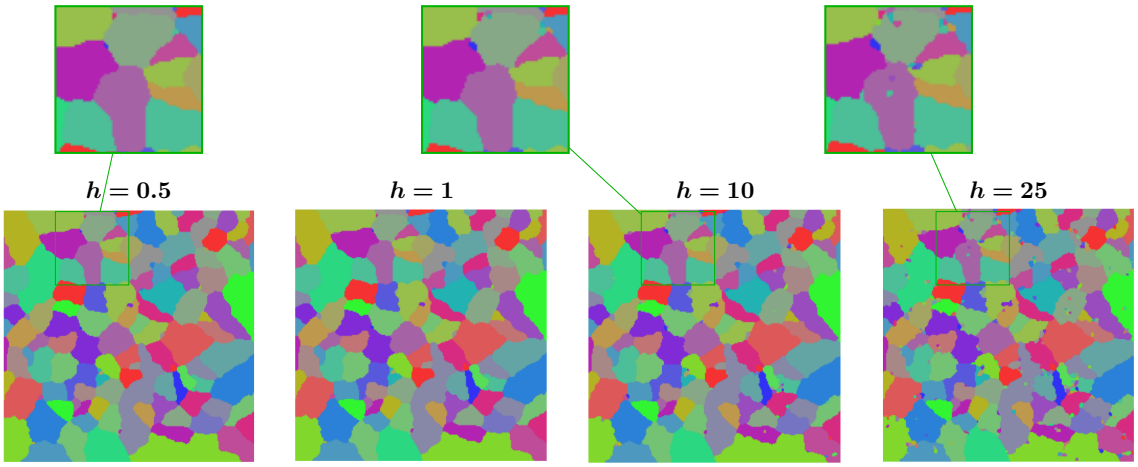


Figure 14. *Visualization of regularization impacts when increasing the step size h corresponding to the results in Figure 13. Labeling results for various step sizes and the neighborhood size $|\mathcal{N}| = 9 \times 9$. Conforming to Figure 13, right panel, labeling accuracy quickly deteriorates once h exceeds the upper bound (5.3) (rightmost panel).*

both uniform and nonuniform averaging, for different sizes of nonlocal neighborhoods $|\mathcal{N}|$, and for different admissible step sizes h based on the insights gained in section 7.2: the largest admissible step size increases with the neighborhood size $|\mathcal{N}|$ and when using nonuniform, rather than uniform, averaging.

Figure 15 displays the corresponding values of the objective function (2.46) as a function of the iteration counter. We observe that this first-order algorithm minimizes quite effectively the nonconvex objective function during the first few dozen iterations.

Figure 16 displays the same information, this time in terms of the function $k \mapsto \frac{1}{n}\|S^k - S^*\|_1$, however. We observe two basic facts: (i) Due to using admissible step sizes, the sequences $(S^k)_{k\geq 0}$ always converge to the integral solution S^*. (ii) In agreement with Theorem 6.6, the order of convergence increases whenever the sequence $(S^k)_{k\geq 0}$ reaches the basin of attraction.

7.4. Accelerated geometric optimization. In this section, we report the evaluation of Algorithm 4 using Algorithm 1 as the baseline. The main ingredients of Algorithm 4 are as follows:

(i) The descent direction d^k given by (5.11) exploits the second-order term $\frac{1}{2}\Omega R_{S_k}(\Omega S_k)$ weighted by parameter h_k which, according to line 9 of Algorithm 4, is determined with negligible additional computational cost by

$$h_k = \tau \cdot \left(\frac{\|R_{S^k}(\Omega S^k)\|_{S^k}^2}{|\langle R_{S^k}(\Omega S^k), \Omega R_{S^k}(\Omega S^k)\rangle|} \right), \qquad \tau \in (0,1). \tag{7.2}$$

Choosing the parameter τ is a compromise between making larger steps (large value of τ) and accuracy of labeling (small value of τ). According to our experience, $\tau = 0.1$ is a reasonable choice that never compromised labeling accuracy. This value was chosen for all experiments discussed in this section.

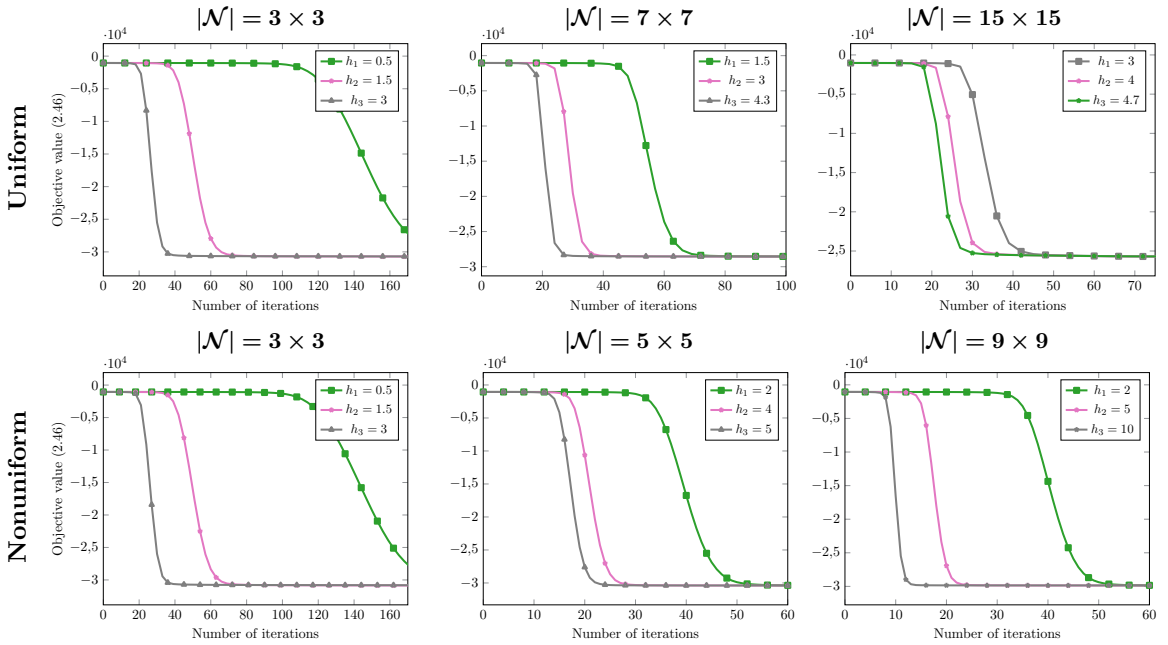


Figure 15. *Minimization of the nonconvex potential (2.46) by Algorithm 1 for various neigborhood sizes $|\mathcal{N}|$, for uniform averaging (top row) and nonuniform averaging (bottom row), and for three constant step sizes $0 < h_1 < h_2 < h_3$, where in each experiment h_3 was chosen smaller than the upper bound discussed in section 7.2 that guarantees a monotonously decreasing sequence of potential values (Proposition 5.1). All experiments illustrate this property and that the largest admissible step size h_3 is most effective. The curves show that the objective function values $J(S)$ stop to decrease at a* local minimizer *S^*. The corresponding objective function value $J(S^*)$ does* not *equal the global lower bound $-\frac{|\mathcal{V}|}{2}$ which is attained at constant labelings as global minimizers that are of no interest, see Remark 5.7.*

(ii) Algorithm 4 calls Algorithm 3, which in turn calls Algorithm 2 in order to satisfy both conditions (5.15) for sufficient decrease. In order to reduce the computational costs of the inner loop started in line 16 of Algorithm 4, we only checked the conditions (5.15a) and (5.15b) at each iteration up to $K_{\max} = 100$ iterations. Figure 17 illustrates that while condition (5.15a) is satisfied throughout all outer loop iterations, condition (5.15b) is satisfied too except for a tiny fraction of inner loops, and therefore the validity of (5.15) is still guaranteed up to a negligible part of iteration steps.

Parameter θ_k of Algorithm 4 corresponds to the step size parameter h_k of Algorithm 4. According to the discussion of proper choices of h_k in section 7.2, parameter θ_k was initialized by values $\theta_0 \in \{\frac{1}{2}, 2\}$ and the adaptive search of θ_k was not allowed to exceed the upper bound $\theta_{\max} = 10$.

Like Algorithm 1, Algorithm 4 terminated when condition (7.1) was satisfied with $\epsilon = 10^{-7}$.

Figure 18 illustrates the convergence of Algorithms 1 and 4 toward labelings for the two initial step sizes $\theta_0 \in \{\frac{1}{2}, 2\}$ corresponding to the fixed step size $h \in \{\frac{1}{2}, 2\}$ of Algorithm 1, and for different sizes $|\mathcal{N}|$ of neighborhoods with nonuniform averaging. Throughout all experiments, we observed that due to using adaptive step sizes θ_k and second-order information for determining the search direction, Algorithm 4 terminates after a smaller num-

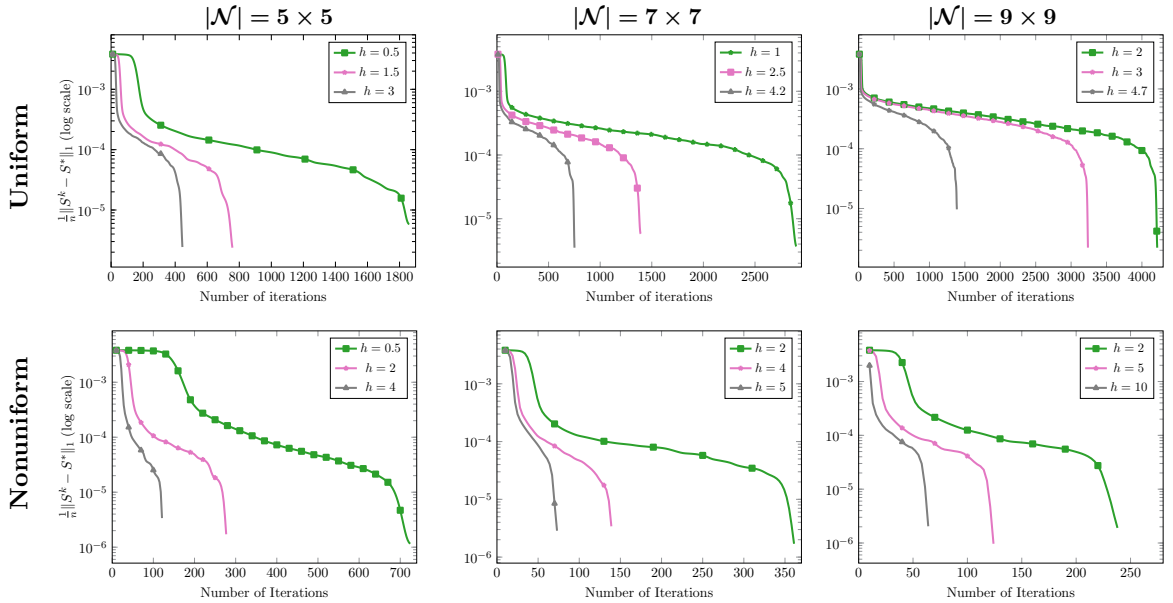


Figure 16. *Norm convergence of the sequence generated by Algorithm 1 toward an integral solution (labeling). Once the basin of attraction of the integral solution has been reached (Theorem 6.6), the convergence rate increases considerably.*

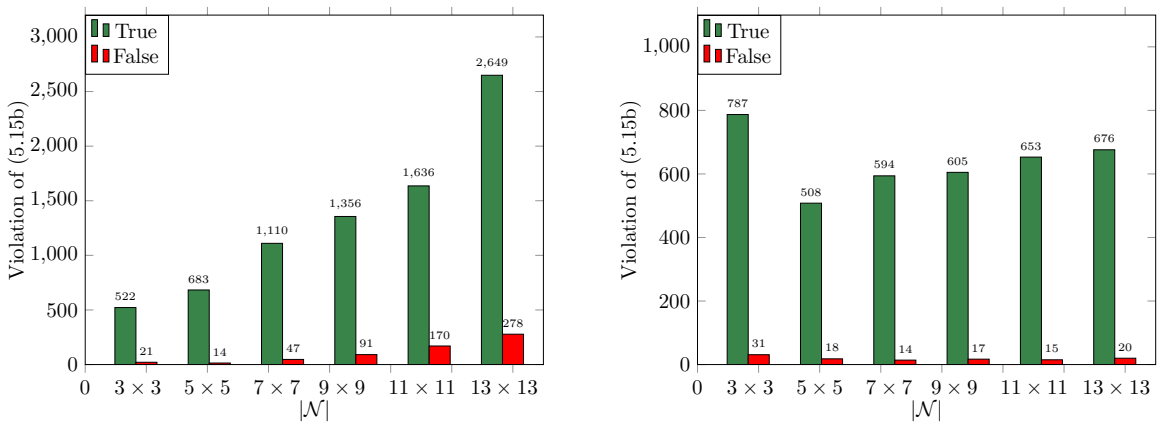


Figure 17. *Fraction of inner loops of Algorithm 4 based on condition (5.15a) that also satisfied condition (5.15b) ({ ●}= True) or not ({ ●}= False), with initialization $\theta_0 = 0.5$ and uniform averaging (left panel) or nonuniform averaging (right panel). Up to a tiny fraction, condition (5.15b) is satisfied, which justifies reducing the computational costs of the inner loop by only checking condition (5.15a) and dispensing with condition (5.15b) after $K_{\max}$ iterations.*

ber of iterations. In particular, the fast convergence of Algorithm 1 within the basins of attraction is preserved.

Table 2 compares Algorithms 1 and 4 in terms of factors of additional iterations required by Algorithm 1 to terminate. We observe that the efficiency of Algorithm 4 is more pronounced when larger neighborhood sizes $|\mathcal{N}|$ or uniform averaging are used.

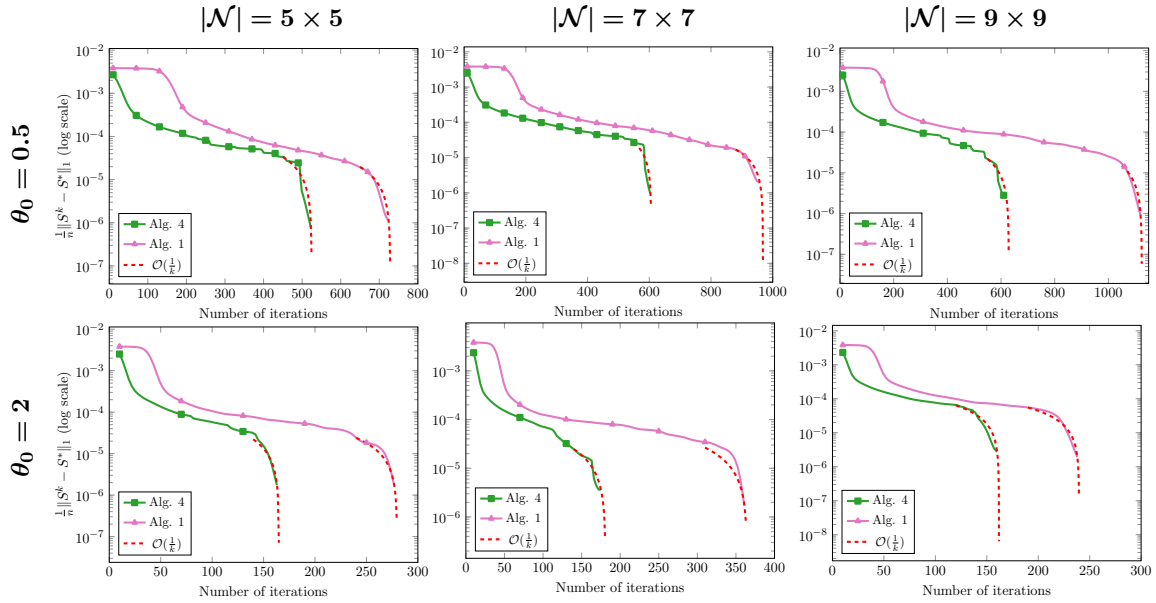


Figure 18. *Comparison of the convergence of Algorithm 1 (●) and Algorithm 4 (●) toward integral solutions (labelings) for various sizes $|\mathcal{N}|$ of neighborhoods and nonuniform averaging. For all parameter settings, Algorithm 4 terminates after a smaller number of iterations.*

Table 2

Number of iterations required by Algorithms 1 and 4 until convergence to a solution of the nonlocal PDE (3.7), for uniform and nonuniform averaging and various neighborhood sizes $|\mathcal{N}|$. The Acc. *columns list the additional factor of iterations required by Algorithm 1 relative to Algorithm 4.*

	Uniform			Nonuniform		
$\lvert\mathcal{N}\rvert$	Alg. 1	Alg. 4	Acc.	Alg. 1	Alg. 4	Acc.
3×3	828	**543**	1.52	760	**557**	1.36
5×5	1860	**697**	2.66	726	**526**	1.38
7×7	3465	**1158**	3	961	**608**	1.58
9×9	4707	**1447**	3.25	1123	**622**	1.81
11×11	9216	**1806**	5.10	1402	**668**	2.1
13×13	9957	**2927**	3.40	1510	**696**	2.17

8. Conclusion and future work.

Conclusion. Using established nonlocal calculus, we devised a novel nonlocal PDE with nonlocal boundary conditions on weighted graphs. Our work adds a novel approach to the literature on PDE-based image analysis that extends the scope from denoising and inpainting to image labeling. An in-depth discussion (section 4) clarified common aspects and differences to related nonlocal approaches from the mathematical viewpoint. Our work has been motivated by the assignment flow approach [17, 18] to metric data labeling, which was shown to constitute a special instance of our general approach introduced in this paper. In particular, our PDE contains the local PDE derived in [29] as a special case and thus provides a natural nonlocal generalization.

The second major contribution of our work rests upon the reparametrization introduced in [29] that turns the assignment flow into a Riemannian descent flow with respect to a nonconvex potential. We established in the present paper two relations to numerical schemes [20] for the geometric integration of the assignment flow: (i) Geometric integration can be applied to solve the novel nonlocal PDE. (ii) We showed that the basic geometric Euler integration scheme corresponds to the basic DC-algorithm of DC programming [56]. Moreover, the geometric viewpoint reveals how second-order information can be used in connection with line search in order to accelerate the basic DC-algorithm for nonconvex optimization.

A range of numerical results were reported in order to illustrate properties of the approach and the theoretical convergence results. This includes, in particular, a linear convergence rate whenever a basin of attraction corresponding to an integral labeling solution is reached, whose existence was established in [19].

Future work. The assignment flow approach (2.35) may be considered as a particular "neural ODE" from the viewpoint of machine learning that generates layers of a deep network by geometric integration of the flow at discrete points of time. For recent work on learning the parameters from data and on quantifying the uncertainty of label assignments, respectively, we refer to [23, 24, 25] and [57]. In the present paper, Lemma 3.1 characterizes parametrizations for which the theoretical results hold. Uniform and data-driven nonuniform parametrizations were used in the experiments to demonstrate broad applicability. Learning these parameters from data is conceivable but beyond the scope of this paper and hence left for future work. Generalizations of the scalar-valued mappings Θ, α to tensor-valued mappings are conceivable as well in order to model not only the interaction across the graph but also the interaction between labels. For the specific case of classification of entire data sets, rather than labeling individual data points, a first step has been done recently using deep linearized assignment flows [26].

Finally, we point out recent work [58, 59] on characterizing assignment flows as critical points of an action functional, provided the nonlocal mapping which specifies the interaction of label assignments across the graph satisfies a certain condition. Reconsidering the PDE (1.1) from this viewpoint defines another problem to be addressed by future work.

Appendix A. Proofs.

A.1. Proofs of section 3.1.

Proof of Lemma 3.1. In order to show (3.4), we directly compute using assumption (3.2) and the parametrization (3.3), for any $x \in \mathcal{V}$,

$$\text{(A.1a)} \qquad \sum_{y\in\mathcal{V}} \Omega(x,y) f(y) \overset{(3.3)}{=} \sum_{y\in\mathcal{N}(x)} \Theta(x,y)\alpha^2(x,y) f(y) + \Theta(x,x) f(x)$$

$$\text{(A.1b)} \qquad = \sum_{y\in\mathcal{N}(x)} \Theta(x,y)\alpha^2(x,y) f(y) + \Theta(x,x) f(x) + (\lambda(x) - \lambda(x))\, f(x)$$

$$\text{(A.1c)} \qquad \overset{(3.5)}{=} \sum_{y\in\mathcal{N}(x)} \Theta(x,y)\alpha^2(x,y)\,(f(y) - f(x)) + \lambda(x) f(x)$$

$$\overset{f|_{\mathcal{V}_I^\alpha}=0}{=} -\sum_{y\in\overline{\mathcal{V}}}\Theta(x,y)\alpha^2(x,y)\left(-\left(f(y)-f(x)\right)\right)+\lambda(x)f(x) \tag{A.1d}$$

$$\overset{(2.13)}{=} -\sum_{y\in\overline{\mathcal{V}}}\Theta(x,y)\left((\mathcal{D}^\alpha)^*(f)(x,y)\right)\alpha(x,y)+\lambda(x)f(x) \tag{A.1e}$$

$$= \sum_{y\in\overline{\mathcal{V}}}\frac{1}{2}\Theta(x,y)\left(-2(\mathcal{D}^\alpha)^*(f)(x,y)\alpha(x,y)\right)+\lambda(x)f(x) \tag{A.1f}$$

$$\overset{(2.14)}{=} \sum_{y\in\overline{\mathcal{V}}}\frac{1}{2}\Theta(x,y)\left(2\mathcal{G}^\alpha(f)(x,y)\alpha(x,y)\right)+\lambda(x)f(x) \tag{A.1g}$$

$$\overset{(2.17)}{=} \frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(f)\right)(x)+\lambda(x)f(x), \tag{A.1h}$$

which proves (3.4).

Assume that $\lambda(x)\le 1\ \forall\, x\in\mathcal{V}$. Then, properties (2.2) easily follow from the nonnegativity of $\Theta\in\mathcal{F}_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}$ and definition (3.5). In addition, if Ω is given by (3.3) and also satisfies (2.34), then equality in (3.5) is achieved:

$$1=\sum_{y\in\mathcal{V}}\Omega(x,y)=\sum_{y\in\mathcal{V}}\Theta(x,y)\alpha^2(x,y)+\Theta(x,x) \tag{A.2a}$$

$$=\lambda(x)-\underbrace{\sum_{y\in\mathcal{V}_I^\alpha}\Theta(x,y)\alpha^2(x,y)}_{\ge 0}\le\lambda(x)\overset{(3.5)}{\le}1. \tag{A.2b}$$

■

Proof of Proposition 3.3. Recalling definition (2.39), we directly compute

$$R_{S(x,t)}\left((\Omega S)(x,t)\right) = R_{S(x,t)}\left(\sum_{y\in\mathcal{V}}\Omega(x,y)S(y,t)\right) \tag{A.3a}$$

$$\overset{(3.4)}{=} R_{S(x,t)}\left(\frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(S)\right)(x)+\lambda(x)S(x)\right). \tag{A.3b}$$

■

A.2. Proof of section 3.3.

Proof of Proposition 3.4. For brevity, we omit the argument t and simply write $S=S(t), V=V(t)$. Recall the componentwise operation $\odot$ defined by (2.30), e.g., $(S\odot V)_j(x)=S_j(x)V_j(x)$ for $j\in[c]$, and $S^2(x)=(S\odot S)(x)$.

Multiplying both sides of (3.15a) with $S(x)=\exp_{S^0}(V(x))$ and summing over $x\in\mathcal{V}$ yields

$$\sum_{x\in\mathcal{V}}\left(S\odot\dot{V}\right)_j(x)-\sum_{x\in\mathcal{V}}\frac{1}{2}\left(S\odot\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(S)\right)\right)_j(x)=\sum_{x\in\mathcal{V}}\left(\lambda S^2\right)_j(x). \tag{A.4}$$

Applying Green's nonlocal first identity (2.15) with $u(x)=S_j(x)$ to the second term on the left-hand side yields with (2.6)

$$\sum_{x\in\mathcal{V}}\left(S\odot\dot{V}\right)_j(x)+\frac{1}{2}\sum_{x\in\overline{\mathcal{V}}}\sum_{y\in\overline{\mathcal{V}}}\left(\mathcal{G}^\alpha(S)\odot(\Theta\mathcal{G}^\alpha(S))\right)_j(x,y) \tag{A.5a}$$

$$+\sum_{y\in\mathcal{V}_I^\alpha}S_j(y)\mathcal{N}^\alpha\left(\Theta\mathcal{G}^\alpha(S_j)\right)(y)=\sum_{x\in\mathcal{V}}\left(\lambda S^2\right)_j(x). \tag{A.5b}$$

Now, using the parametrization (3.10) of S, we compute at each $x\in\mathcal{V}$,

$$\dot{S}(x) = \frac{d}{dt}\exp_{S^0(x)}(V(x)) \tag{A.6a}$$

$$\overset{(3.11)}{=}\frac{\left(\frac{d}{dt}\left(S^0(x)\odot e^{V(x)}\right)\right)\langle S^0(x),e^{V(x)}\rangle-\left(\frac{d}{dt}\langle S^0(x),e^{V(x)}\rangle\right)S^0(x)\odot e^{V(x)}}{\langle S^0(x),e^{V(x)}\rangle^2} \tag{A.6b}$$

$$=\frac{\langle S^0(x),e^{V(x)}\rangle(S^0\odot e^V)(x)\odot\dot{V}(x)-\langle S^0(x)\odot e^{V(x)},\dot{V}(x)\rangle(S^0\odot e^V)(x)}{\langle S^0(x),e^{V(x)}\rangle^2} \tag{A.6c}$$

$$=(S\odot\dot{V})(x)-\langle S(x),\dot{V}(x)\rangle S(x) \tag{A.6d}$$

$$\overset{(3.13)}{=}(S\odot\dot{V})(x)-\langle S(x),(\Pi_0\Omega\exp_{S^0}(V))(x)\rangle S(x) \tag{A.6e}$$

$$\overset{(3.17)}{=}(S\odot\dot{V})(x)-\phi_S(x)S(x). \tag{A.6f}$$

Solving the last equation for $(S\odot\dot{V})(x)$ and substitution into (A.5) yields after taking the sum over $x\in\mathcal{V}$, for each $S_j=\{S_j(x)\colon x\in\mathcal{V}\}$, $j\in[c]$,

$$\frac{1}{2}\frac{d}{dt}\left(\sum_{x\in\mathcal{V}}S_j(x)\right)+\frac{1}{2}\langle\mathcal{G}^\alpha(S_j),\Theta\mathcal{G}^\alpha(S_j)\rangle_{\overline{\mathcal{V}}\times\overline{\mathcal{V}}}+\sum_{x\in\mathcal{V}}\phi_S(x)S_j(x) \tag{A.7a}$$

$$+\sum_{y\in\mathcal{V}_{\mathcal{I}^\alpha}}S_j\mathcal{N}^\alpha\left(\Theta\mathcal{G}^\alpha(S_j)\right)(y)=\sum_{x\in\mathcal{V}}\left(\lambda S_j^2\right)(x), \tag{A.7b}$$

which after rearranging the terms is equal to (3.16). ∎

A.3. Proof of section 5.1.

Proof of Proposition 5.1. Equation (5.2) directly follows from Proposition 3.3, from the specification (2.32) of the similarity mapping and from the relation $\exp_p=\mathrm{Exp}_p\circ R_p$ for $p\in\mathcal{S}$ (cf. (2.40), (2.41)). Leveraging the parametrization (3.14) of system (3.7) and discretization of (3.14) by forward finite differences with step size parameter $h>0$ yields for $x\in\mathcal{V}$

$$\frac{V^{k+1}(x)-V^k(x)}{h}=\left(\frac{1}{2}\mathcal{D}^\alpha\left(\Theta\mathcal{G}^\alpha(\exp_{S^0}(V^k))\right)+\lambda\exp_{S^0}(V^k)\right)(x), \tag{A.8}$$

which is (5.2) after applying the lifting map (2.41) to V^{k+1}. Consequently, in view of zero nonlocal boundary conditions, the zero extension of (5.2) to $\overline{\mathcal{V}}$ verifies that $\overline{S}^k$ is indeed a first-order approximation of solution $S(kh)$ to (3.7).

It remains to show that (5.1) implies (5.4). Adding and subtracting a convex negative entropy term

$$\langle S,\log S\rangle=\sum_{x\in\mathcal{V}}\langle S(x),\log S(x)\rangle,\qquad \log S(x)=(\log S_1(x),\dots,\log S_c(x))^\top \tag{A.9}$$

to the potential (2.46), we write with the convex constraint $S \in \overline{\mathcal{W}}$ represented by the delta-function $\delta_{\overline{\mathcal{W}}}$,

$$J(S) = \underbrace{\gamma\langle S, \log S\rangle + \delta_{\overline{\mathcal{W}}}(S)}_{g(S)} - \Big(\underbrace{\frac{1}{2}\langle S, \Omega S\rangle + \gamma\langle S, \log S\rangle}_{h(S)}\Big), \quad \gamma > |\lambda_{\min}(\Omega)|, \tag{A.10}$$

which is a DC-function [60] if $\gamma > |\lambda_{\min}(\Omega)|$, i.e., both $g(S)$ and $h(S)$ are convex. Indeed, while the convexity of g is obvious, the convexity of h becomes apparent when inspecting its Hessian. Writing

$$s = \mathrm{vec}_r(S) \tag{A.11}$$

with the row-stacking mapping vec_r, we have ($\otimes$ denotes the Kronecker matrix product)

$$\langle S, \Omega S\rangle = \langle s, (\Omega \otimes I_c)s\rangle, \tag{A.12a}$$

$$\langle S, \log S\rangle = \langle s, \log s\rangle, \qquad \log s = (\ldots, \log s_i, \ldots)^\top \tag{A.12b}$$

and hence for any $v \in \mathbb{R}^{nc}$ with $\|v\| = 1$

$$d^2h(S)(v,v) = \left\langle v, \left((\Omega \otimes I_c) + \gamma \mathrm{Diag}\left(\frac{\mathbb{1}}{s}\right)\right) v \right\rangle > \lambda_{\min}(\Omega) + \gamma, \tag{A.13}$$

where the last inequality follows from $\lambda \geq \lambda_{\min}(\Omega)$ for any eigenvalue λ of symmetric matrix Ω (recall (2.2), (2.34)), $\lambda(A \otimes B) = \lambda_i(A)\lambda_j(B)$ for some i, j [61], and $\lambda_{\min}(\mathrm{Diag}(\frac{\mathbb{1}}{s})) > 1$ if $S \in \overline{\mathcal{W}}$.

Thus, if $\gamma > |\lambda_{\min}(\Omega)|$, then h is convex and minimizing (A.10) is a DC-programming problem [32, 33]. Using Fenchel's inequality $-h(S^k) \leq h^*(\widetilde{S}) - \langle S^k, \widetilde{S}\rangle \, \forall \widetilde{S}$, let $\widetilde{S}^k$ minimize at the current iterate S^k the upper bound

$$J(S^k) = g(S^k) - h(S^k) \leq g(S^k) + h^*(\widetilde{S}) - \langle S^k, \widetilde{S}\rangle \qquad \forall \widetilde{S} \tag{A.14a}$$

with respect to $\widetilde{S}$, i.e.,

$$0 = \partial h^*(\widetilde{S}^k) - S^k \qquad \Leftrightarrow \qquad \widetilde{S}^k \in \partial h(S^k) = \nabla h(S^k). \tag{A.14b}$$

In particular, $-h(S^k) = h^*(\widetilde{S}^k) - \langle S^k, \widetilde{S}^k\rangle$ and hence

$$J(S^k) = g(S^k) + h^*(\widetilde{S}^k) - \langle S^k, \widetilde{S}^k\rangle. \tag{A.15}$$

Minimizing in turn the right-hand side with respect to S^k guarantees (5.4) and defines the update S^{k+1} by

$$S^{k+1} = \arg\min_S \{g(S) - \langle S, \widetilde{S}^k\rangle\} \qquad \Leftrightarrow \qquad 0 = \partial g(S^{k+1}) - \widetilde{S}^k \tag{A.16a}$$

$$\Leftrightarrow \quad \gamma(\log S^{k+1}(x) + \mathbb{1}) + \partial\delta_{\overline{\mathcal{S}}}\left(S^{k+1}(x)\right) \overset{(A.14b)}{=} \nabla h(S^k)(x) \tag{A.16b}$$

$$= \quad (\Omega S^k)(x) + \gamma(\log S^k(x) + \mathbb{1}). \tag{A.16c}$$

Solving for $S^{k+1}(x)$ yields (5.1), respectively, (5.2), with step size $h = \frac{1}{\gamma} < 1$ due to $\gamma > |\lambda_{\min}(\Omega)|$. ∎

A.4. Proofs of section 5.2.

Proof of Lemma 5.2. Taking into account the parametrization (3.10), we compute the partial derivative of (2.46) (recall the operation $\odot$ defined by (2.30))

$$\partial_i J(V) = -\langle \Omega \exp_{S^0}(V), \partial_i \exp_{S^0}(V)\rangle \tag{A.17a}$$
$$= -\langle \Omega \exp_{S^0}(V), \exp_{S^0}(V) \odot e_i + \exp_{S^0}(V)_i \exp_{S^0}(V)\rangle \tag{A.17b}$$
$$= -\left(\Omega \exp_{S^0}(V) \odot \exp_{S^0}(V)\right)_i + \langle \Omega \exp_S(V), \exp_{S^0}(V)\rangle \exp_{S^0}(V)_i \tag{A.17c}$$
$$= -\left(R_{\exp_{S^0}(V)}(\Omega \exp_{S^0}(V))\right)_i \tag{A.17d}$$

and consequently $\partial J(V) = \partial_V J(V) = -R_{\exp_{S^0}(V)}(\Omega \exp_{S^0}(V)) = R_S \partial_S J(S) = \operatorname{grad}_g J(S)$. ∎

Proof of Proposition 5.3.

(i) Using $S^k = \exp_{S^0}(V^k)$ and

$$\partial J(V^k) = -R_{S^k}(\Omega S^k) = \operatorname{grad}_g J(S^k), \tag{A.18}$$

by Lemma 5.2 along with the identities (recall that both R_S and the orthogonal projection Π_0 act rowwise)

$$R_S = \Pi_0 R_S = R_S \Pi_0 = \Pi_0 R_S \Pi_0 = R_S|_{\mathcal{T}_0}, \quad S \in \mathcal{W}, \qquad \Pi_0^2 = \Pi_0, \tag{A.19}$$

and

$$(R_{S^k}|_{\mathcal{T}_0})^{-1} V = \left(\ldots, \Pi_0 \frac{V(x)}{S^k(x)}, \ldots\right)^\top, \quad x \in V, \quad V \in \mathcal{T}_0, \quad S^k \in \mathcal{W}, \tag{A.20}$$

by [29, Lemma 3.1], we have

$$\langle \partial J(V^k), d^k\rangle \overset{(5.3)}{=} \langle \partial J(V^k), d(S^k, h_k)\rangle \tag{A.21a}$$
$$= -\langle R_{S^k}(\Omega S^k), \Pi_0 \Omega S^k\rangle - \frac{h_k}{2}\langle \partial J(V^k), \Pi_0 \Omega \partial J(V^k)\rangle \tag{A.21b}$$
$$= -\langle R_{S^k}(\Omega S^k), \left((R_{S^k}|_{\mathcal{T}_0})^{-1} R_{S^k}|_{\mathcal{T}_0}\right) \Pi_0 \Omega S^k\rangle - \frac{h_k}{2}\langle \partial J(V^k), \Pi_0 \Omega \partial J(V^k)\rangle \tag{A.21c}$$
$$\overset{(5.16),(A.19),(A.20)}{=} -\langle R_{S^k}(\Omega S^k), R_{S^k}(\Omega S^k)\rangle_{S^k} - \frac{h_k}{2}\langle \partial J(V^k), \Pi_0 \Omega \partial J(V^k)\rangle. \tag{A.21d}$$

Since the first term on the right-hand side of (A.21d) is negative on $\mathcal{T}_0$, setting

$$h_k \in \left(0, \frac{\|R_{S^k}(\Omega S^k)\|_{S^k}^2}{|\langle \partial J(V^k), \Pi_0 \Omega \partial J(V^k)\rangle|}\right) \tag{A.22}$$

yields a sequence $(d^k)_{k\geq 1}$ satisfying

$$\langle \partial J(V^k), d^k\rangle < 0, \qquad k \geq 1. \tag{A.23}$$

Consider $c_1, c_2 \in (0,1)$ with $c_1 < c_2$ and set

$$G(\gamma) = J(V^k + \gamma d^k), \tag{A.24a}$$
$$L(\gamma) = J(V^k) + c_1 \gamma \langle \partial J(V^k), d^k\rangle \quad \text{for} \quad \gamma \geq 0. \tag{A.24b}$$

Due to $c_1 < 1$ and (A.23), the inequality

$$G'(0) = \langle \partial J(V^k), d^k \rangle < c_1 \langle \partial J(V^k), d^k \rangle = L'(0) < 0 \tag{A.25}$$

holds. Hence there is a constant $t_k > 0$ such that

$$G(\gamma) < L(\gamma), \quad \gamma \in (0, t_k), \tag{A.26a}$$
$$G(t_k) = L(t_k). \tag{A.26b}$$

Substituting the first-order Taylor expansion

$$G(t_k) = J(V^k + t_k d^k) = G(0) + t_k G'(\widetilde{\gamma}_k) \tag{A.27a}$$
$$= J(V^k) + t_k \langle \partial J(V^k + \widetilde{\gamma}_k d^k), d^k \rangle, \qquad \widetilde{\gamma}_k \in (0, t_k), \tag{A.27b}$$

into (A.26b) yields with (A.24b), (A.23) and $0 < c_1 < c_2 < 1$

$$\langle \partial J(V^k + \widetilde{\gamma}_k d^k), d^k \rangle = c_1 \langle \partial J(V^k), d^k \rangle \geq c_2 \langle \partial J(V^k), d^k \rangle. \tag{A.28a}$$

Therefore, with $\partial J(V^k), d^k \in \mathcal{T}_0$, and using that the restriction $R_{S^k}|_{\mathcal{T}_0}$ of the map R_{S^k} to $\mathcal{T}_0$ is invertible with the inverse $(R_{S^k})^{-1}_{|\mathcal{T}_0}$ acting rowwise as specified by (A.20), the right-hand side of (A.28) becomes

$$c_2 \langle \partial J(V^k), d^k \rangle = c_2 \left\langle \partial J(V^k), (R_{S^k}|_{\mathcal{T}_0})^{-1} (R_{S^k}(d^k)) \right\rangle \tag{A.28b}$$
$$\overset{(5.16),(A.20)}{=} c_2 \left\langle \Pi_0 \partial J(V^k), R_{S^k}(d^k) \right\rangle_{S^k}. \tag{A.28c}$$

By virtue of (A.18) and $\Pi_0 \partial J(V^k) = \partial J(V^k)$, both sides of (A.28) correspond to the expressions of (5.15b) between the bars $|\cdots|$. Since the above derivation shows that both sides of (A.28) are negative, taking the magnitude on both sides proves (5.15b). Recalling the shorthand (5.9) and inequality (A.27) and setting θ_k small enough with $\theta_k \leq \widetilde{\gamma}_k$, the iterates $S^{k+1} = \exp_{S^0}(V^k + \theta_k d_k)$ satisfy

$$J(S^{k+1}) - J(S^k) \overset{(A.27)}{=} t_k \langle \partial J(V^k + \widetilde{\gamma}_k d^k), d^k \rangle \tag{A.29a}$$
$$\leq \theta_k \langle \partial J(V^k + \widetilde{\gamma}_k d^k), d^k \rangle \tag{A.29b}$$
$$\overset{(A.28)}{\leq} \theta_k c_2 \langle \partial J(V^k), d^k \rangle \tag{A.29c}$$
$$\overset{(A.18)}{\overset{(A.28)}{=}} \theta_k c_2 \langle \mathrm{grad}_g J(S^k), R_{S^k}(d^k) \rangle_{S^k} \tag{A.29d}$$

which proves inequality (5.15a) since both sides are nonpositive and $c_1 < c_2$.

(ii) We prove by contradiction. Assume, on the contrary, that there exists a sequence $(S^k)_{k \geq 0} \subset \overline{\mathcal{W}}$ in the compact set $\overline{\mathcal{W}}$ and a convergent subsequence $(S^{k_l})_{l \geq 0}$ with limit point $\lim_{l \to \infty} S^{k_l} = S^*$ which is *not* an equilibrium of (2.38a). Then, since

the functional (2.46) is bounded from below on $\overline{\mathcal{W}}$, taking the sum in (5.15a) yields

(A.30)
$$\sum_{l=0}^{\infty} c_1 \gamma_{k_l} \langle \mathrm{grad}_g J(S^{k_l}), R_{S^{k_l}}(d^{k_l}) \rangle_{S^{k_l}} > \sum_{l=0}^{\infty} \left(J(S^{k_{l+1}}) - J(S^{k_l}) \right) = \underbrace{J(S^*) - J(S^0)}_{>-\infty},$$

and consequently

$$c_1 \gamma_* \langle \mathrm{grad}_g J(S^*), R_{S^*}(d^*) \rangle_{S^*} = 0. \tag{A.31}$$

Using $d^* = d(S^*, h_*)$ given by (5.11) along with $c_1 > 0$ and the assumption $\gamma_* > 0$, we evaluate this equation similarly to (A.21),

$$\begin{aligned}
0 &= \langle \mathrm{grad}_g J(S^*), R_{S^*}(d^*) \rangle_{S^*} && \text{(A.32a)}\\
&\overset{(A.19)}{=} \left\langle -R_{S^*}(\Omega S^*), R_{S^*}\left(\Omega S^* + \frac{h_*}{2} \Omega R_{S^*}(\Omega S^*) \right) \right\rangle_{S^*} && \text{(A.32b)}\\
&\overset{(5.16),(A.19)}{=} -\sum_{x \in \mathcal{V}} \left\langle \Pi_0 R_{S^*(x)}(\Omega S^*)(x), \frac{R_{S^*(x)}\left(\Omega S^* + \frac{h_*}{2} \Omega R_{S^*}(\Omega S^*) \right)(x)}{S^*(x)} \right\rangle && \text{(A.32c)}\\
&\overset{(A.20)}{=} -\sum_{x \in \mathcal{V}} \left\langle R_{S^*(x)}(\Omega S^*)(x), (R_{S^*(x)}|_{T_0})^{-1} R_{S^*(x)}\left(\Omega S^* + \frac{h_*}{2} \Omega R_{S^*}(\Omega S^*) \right)(x) \right\rangle && \text{(A.32d)}\\
&\overset{(A.19)}{=} -\langle \Omega S^*, R_{S^*}(\Omega S^*) \rangle - \frac{h_*}{2} \langle \Omega S^*, R_{S^*}\left(\Omega R_{S^*}(\Omega S^*) \right) \rangle . && \text{(A.32e)}
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{h_*}{2} \langle \Omega S^*, R_{S^*}\left(\Omega R_{S^*}(\Omega S^*) \right) \rangle &= -\langle \Omega S^*, R_{S^*}(\Omega S^*) \rangle && \text{(A.33a)}\\
&= -\sum_{x \in \mathcal{V}} \left\langle (\Omega S^*)(x), R_{S^*(x)}(\Omega S^*)(x) \right\rangle && \text{(A.33b)}
\end{aligned}$$

using $R_p \mathbb{1}_c = 0$, $p \in \mathcal{S}$,

$$\begin{aligned}
&= -\sum_{x \in \mathcal{V}} \left\langle (\Omega S^*)(x) - \langle (\Omega S^*)(x), S^*(x) \rangle \mathbb{1}_c, R_{S^*(x)}(\Omega S^*)(x) \right\rangle && \text{(A.33c)}\\
&\overset{(2.36)}{=} -\sum_{x \in \mathcal{V}} \Big\langle (\Omega S^*)(x) - \langle (\Omega S^*)(x), S^*(x) \rangle \mathbb{1}_c, && \text{(A.33d)}\\
&\qquad S^*(x) \odot \left((\Omega S^*)(x) - \langle S^*(x), (\Omega S^*)(x) \rangle \mathbb{1}_c \right) \Big\rangle && \text{(A.33e)}\\
&= -\sum_{x \in \mathcal{V}} \sum_{j \in [c]} S_j^*(x) \left((\Omega S^*)_j(x) - \langle (\Omega S^*)(x), S^*(x) \rangle \right)^2 . && \text{(A.33f)}
\end{aligned}$$

By [19, Proposition 5], S^* is an equilibrium of the flow (2.38a) if and only if

$$(\Omega S^*)_j(x) = \langle (\Omega S)^*(x), S^*(x)\rangle \qquad \forall x \in \mathcal{V}, \qquad \forall j \in \operatorname{supp}(S^*(x)). \tag{A.33g}$$

Therefore, by assumption, there exists $\widetilde{x} \in \mathcal{V}$ and $l \in \operatorname{supp}(S^*(\widetilde{x}))$ with $(\Omega S^*)_l(\widetilde{x}) \neq \langle \Omega S^*(\widetilde{x}), S^*(\widetilde{x})\rangle$ and consequently

$$\begin{aligned}
\frac{h_*}{2}\langle \Omega S^*, R_{S^*}(\Omega R_{S^*}(\Omega S^*))\rangle &= -\langle \Omega S^*, R_{S^*}(\Omega S^*)\rangle && \text{(A.33h)}\\
&\leq -S_l^*(\widetilde{x})\left((\Omega S^*)_l(\widetilde{x}) - \langle (\Omega S^*)(\widetilde{x}), S^*(\widetilde{x})\rangle\right)^2 && \text{(A.33i)}\\
&< 0. && \text{(A.33j)}
\end{aligned}$$

Since the first two expressions are strictly negative, this yields the contradiction

$$\begin{aligned}
&-\frac{1}{2}\langle \Omega S^*, R_{S^*}(\Omega S^*)\rangle && \text{(A.34a)}\\
&\quad = -\frac{1}{2}\frac{\langle \Omega S^*, R_{S^*}(\Omega S^*)\rangle}{|\langle \Omega S^*, R_{S^*}(\Omega R_{S^*}(\Omega S^*))\rangle|}|\langle \Omega S^*, R_{S^*}(\Omega R_{S^*}(\Omega S^*))\rangle| && \\
&\quad \overset{(A.19),(5.9)}{=} -\frac{1}{2}\frac{\langle \Omega S^*, R_{S^*}(\Omega S^*)\rangle}{|\langle \operatorname{grad}_g J(S^*), \Pi_0 \Omega \operatorname{grad}_g J(S^*)\rangle|}|\langle \Omega S^*, R_{S^*}(\Omega R_{S^*}(\Omega S^*))\rangle| && \text{(A.34b)}\\
&\quad \overset{(A.22),(5.9)}{\leq} -\frac{h_*}{2}|\langle \Omega S^*, R_{S^*}(\Omega R_{S^*}(\Omega S^*))\rangle| && \text{(A.34c)}\\
&\quad \overset{(A.33h)}{=} -\langle \Omega S^*, R_{S^*}(\Omega S^*)\rangle, && \text{(A.34d)}
\end{aligned}$$

which proves (ii).

(iii) We prove by contraposition and show that a limit point $S^* \in \mathcal{W}$ cannot locally minimize $J(S)$. Let $\overline{S}_{(l)} \in \overline{\mathcal{W}}$ be a constant vector field given for each $x \in \mathcal{V}$ by

$$\overline{S}_{(l)}(x) = e_l = (0, \ldots, 0, 1, 0 \ldots, 0)^\top \in \mathbb{R}^c \tag{A.35}$$

for arbitrary $l \in [c]$. Then, for any $S \in \overline{\mathcal{W}}$ with $S(x) \in \Delta_c$ for each $x \in \mathcal{V}$, and with $\Omega(x,y) \geq 0$,

$$\begin{aligned}
\langle S, \Omega S\rangle &= \sum_{x\in\mathcal{V}}\sum_{j\in[c]}\sum_{y\in\mathcal{N}(x)} \Omega(x,y) S_j(x) S_j(y) \leq \sum_{x\in\mathcal{V}}\left(\sum_{y\in\mathcal{N}(x)}\Omega(x,y)\right)\underbrace{\sum_{j\in[c]} S_j(x)}_{=1} && \text{(A.36a)}\\
&= \sum_{x\in\mathcal{V}}\sum_{j\in[c]}\sum_{y\in\mathcal{N}(x)} \Omega(x,y)\overline{S}_{(l)j}(x)\overline{S}_{(l)j}(y) && \text{(A.36b)}\\
&= \langle \overline{S}_{(l)}, \Omega \overline{S}_{(l)}\rangle, && \text{(A.36c)}
\end{aligned}$$

where the inequality is strict if $S \in \mathcal{W}$. Consequently, the constant vector $\overline{S}_{(l)}$ is a global minimizer of the objective function $J(S)$ (2.46) with minimal value

$J(\overline{S}_{(l)}) = -\frac{1}{2}\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{N}(x)}\Omega(x,y)$. Let $B_\delta(S^*)\subset\mathcal{W}$ be the open ball with radius $\delta>0$ containing S^*. By assumption, $S^*_j(x)>0\;\forall x\in\mathcal{V},\forall j\in[c]$ and there exists an $\epsilon>0$ small enough such that

$$S^*_\epsilon := S^* + \epsilon(\overline{S}_{(l)} - S^*) \in B_\delta(S^*)\subset\mathcal{W}. \tag{A.37}$$

Evaluating $J(S)$ at S^*_ϵ yields

$$J(S^*_\epsilon) \overset{(A.37)}{=} -\frac{1}{2}\left\langle S^* + \epsilon(\overline{S}_{(l)} - S^*), \Omega(S^* + \epsilon(\overline{S}_{(l)} - S^*))\right\rangle \tag{A.38a}$$

$$= J(S^*) - \epsilon\langle S^*, \Omega(\overline{S}_{(l)} - S^*)\rangle - \frac{\epsilon^2}{2}\langle \overline{S}_{(l)} - S^*, \Omega(\overline{S}_{(l)} - S^*)\rangle \tag{A.38b}$$

$$\overset{(ii),(2.2)}{=} J(S^*) - \epsilon\left\langle\langle S^*, \Omega S^*\rangle\mathbb{1}, \overline{S}_l - S^*\right\rangle + \frac{\epsilon^2}{2}\left\langle\langle S^*, \Omega S^*\rangle\mathbb{1}, \overline{S}_{(l)} - S^*\right\rangle \tag{A.38c}$$

$$+\epsilon^2\left(J(\overline{S}_{(l)}) + \frac{1}{2}\langle\overline{S}_{(l)}, \Omega S^*\rangle\right), \tag{A.38d}$$

and since $\langle\mathbb{1}, \overline{S}_{(l)} - S^*\rangle = \sum_{x\in\mathcal{V}}\sum_{j\in[c]}(\overline{S}_{(l)j}(x) - S^*_j(x)) \overset{(A.35)}{=} \sum_{x\in\mathcal{V}}(1-\sum_{j\in[c]}S^*_j(x)) = 0$,

$$= J(S^*) + \epsilon^2\left(J(\overline{S}_{(l)}) + \frac{1}{2}\langle\overline{S}_{(l)}, \Omega S^*\rangle\right). \tag{A.38e}$$

It follows from (ii) that S^* is an equilibrium point. Hence we can invoke condition (A.33g) to obtain the identity

$$\frac{1}{2}\langle\overline{S}_{(l)}, \Omega S^*\rangle = \frac{1}{2}\sum_{x\in\mathcal{V}}\sum_{j\in[c]}(\Omega S^*)_j(x)\overline{S}_{(l)j}(x) = \frac{1}{2}\sum_{x\in\mathcal{V}}(\Omega S^*)_l(x) \tag{A.38f}$$

$$\overset{(A.33g)}{=} \frac{1}{2}\sum_{x\in\mathcal{V}}\langle S^*(x), \Omega S^*(x)\rangle = -J(S^*) \tag{A.38g}$$

and consequently, since $\overline{S}_{(l)}$ was shown above to be a global minimizer of J,

$$J(S^*_\epsilon) = J(S^*) + \epsilon^2\left(J(\overline{S}_{(l)}) - J(S^*)\right) < J(S^*). \tag{A.38h}$$

By assumption we have $S^*\in\mathcal{W}$ and using (A.36) it holds that $J(S^*_\epsilon) < J(S^*)$. As $\delta>0$ was chosen arbitrarily subject to the constraint (A.37), this shows that S^* cannot be a local minimizer, which proves (iii).

(iv) Analogous to (A.33) we compute

$$
\begin{aligned}
&-\frac{h_k}{2}\left\langle \Omega S^k, R_{S^k}\left(\Omega R_{S^k}(\Omega S^k)\right)\right\rangle - \langle \Omega S^k, R_{S^k}(\Omega S^k)\rangle \\
&\quad = -\frac{h_k}{2}\left\langle \Omega S^k, R_{S^k}\left(\Omega R_{S^k}(\Omega S^k)\right)\right\rangle \\
&\qquad - \sum_{x\in\mathcal{V}}\sum_{j\in[c]} S_j^k(x)\left((\Omega S^k)_j(x) - \left\langle (\Omega S^k)(x), S^k(x)\right\rangle\right)^2 \\
&\quad = -\frac{h_k}{2}\left\langle \Omega S^k, R_{S^k}\left(\Omega R_{S^k}(\Omega S^k)\right)\right\rangle \\
&\qquad - \sum_{x\in\mathcal{V}}\sum_{j\in[c]} \frac{1}{S_j^k(x)}\left(S_j^k(x)\left((\Omega S^k)_j(x) - \left\langle (\Omega S^k)(x), S^k(x)\right\rangle\right)\right)^2 \\
&\quad = -\frac{h_k}{2}\left\langle \Omega S^k, R_{S^k}\left(\Omega R_{S^k}(\Omega S^k)\right)\right\rangle \\
&\qquad - \sum_{x\in\mathcal{V}}\left\langle \frac{\mathbb{1}}{S^k(x)}, \operatorname{grad}_g(J(S^k))(x)\odot \operatorname{grad}_g(J(S^k))(x)\right\rangle.
\end{aligned}
\tag{A.39}
$$

Since this expression converges to 0 for $k\to\infty$, the additional assumption $\sum_{k=0}^{\infty} h_k < \infty$ implies that the second term on the right-hand side is a zero sequence which shows (iv). ∎

A.5. Proof of section 5.3.

Proof of Proposition 5.5.

(i) Let D be the diagonal degree matrix

$$
D(x,x) = \sum_{y\in\mathcal{V}} \Omega(x,y), \tag{A.40}
$$

and let $f\in\mathcal{F}_{\mathcal{V}}$. Then, using $\sum_{x,y\in\mathcal{V}} f^2(x) = \sum_{x,y\in\mathcal{V}} f^2(y)$, one has

$$
\begin{aligned}
\langle f, (D-\Omega)f\rangle_{\mathcal{V}} &= \sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}} \Omega(x,y)\left((f^2(x) - f(x)f(y)\right) && \text{(A.41a)}\\
&\overset{\Omega(x,y)=\Omega(y,x)}{=} \sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}} \Omega(x,y)\left(\left(\frac{1}{2}f^2(x) - f(x)f(y) + \frac{1}{2}f^2(y)\right)\right. && \text{(A.41b)}\\
&= \frac{1}{2}\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}} \Omega(x,y)(f(x)-f(y))^2. && \text{(A.41c)}
\end{aligned}
$$

Now we directly derive the right-hand side of (5.21) from (5.20).

$$
\begin{aligned}
&-\frac{\langle f, \mathcal{D}^\alpha(\Theta\mathcal{G}^\alpha f)\rangle_{\overline{\mathcal{V}}}}{\langle f, f\rangle_{\overline{\mathcal{V}}}} \\
&\overset{(2.17),(2.14)}{=} \frac{\sum_{x\in\overline{\mathcal{V}}} f(x) 2\left(\sum_{y\in\overline{\mathcal{V}}} \Theta(x,y)\alpha^2(x,y)(f(x)-f(y))\right)}{\sum_{x\in\overline{\mathcal{V}}} f^2(x)}
\end{aligned}
\tag{A.42a}
$$

$$\overset{(2.6),f|_{\mathcal{V}_{\mathcal{I}}^{\alpha}}=0}{=} \frac{\sum_{x\in\mathcal{V}} f(x) 2\left(\sum_{y\in\mathcal{V}\dot{\cup}\mathcal{V}_{\mathcal{I}}^{\alpha}} \Theta(x,y)\alpha^2(x,y)(f(x)-f(y))\right)}{\sum_{x\in\mathcal{V}} f^2(x)} \tag{A.42b}$$

$$= \frac{\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}} \left(\Theta(x,y)\alpha^2(x,y)(f^2(x)-2f(x)f(y)+f^2(x))\right)}{\sum_{x\in\mathcal{V}} f^2(x)} \tag{A.42c}$$

$$+ \frac{2\sum_{x\in\mathcal{V}}\left(\sum_{y\in\mathcal{V}_{\mathcal{I}}^{\alpha}} \Theta(x,y)\alpha^2(x,y)\right) f^2(x)}{\sum_{x\in\mathcal{V}} f^2(x)} \tag{A.42d}$$

and analogous to (A.41)

$$= \frac{\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}} \Theta(x,y)\alpha^2(x,y)(f(x)-f(y))^2 + 2\sum_{x\in\mathcal{V}}\left(\sum_{y\in\mathcal{V}_{\mathcal{I}}^{\alpha}} \Theta(x,y)\alpha^2(x,y)\right) f^2(x)}{\sum_{x\in\mathcal{V}} f^2(x)} \tag{A.42e}$$

$$\overset{\substack{(2.6)\\(3.5)\\(3.3)}}{=} \frac{\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}} \Omega(x,y)(f(x)-f(y))^2 + 2\sum_{x\in\mathcal{V}}\left(\lambda(x)-\sum_{y\in\mathcal{V}}\Omega(x,y)\right) f^2(x)}{\sum_{x\in\mathcal{V}} f^2(x)} \tag{A.42f}$$

$$\overset{(A.41)}{=} 2\frac{\langle f,(D-\Omega)f\rangle_{\mathcal{V}} + \langle f,(\Lambda-D)f\rangle_{\mathcal{V}}}{\langle f,f\rangle_{\mathcal{V}}} \tag{A.42g}$$

$$= 2\frac{\langle f,(\Lambda-\Omega)f\rangle_{\mathcal{V}}}{\langle f,f\rangle_{\mathcal{V}}}, \tag{A.42h}$$

which proves that the right-hand sides of (5.20) and (5.21) are equal. By virtue of (3.5), which is an equation by assumption, the matrix $\Lambda-\Omega$ defined by (5.22) and (3.3) is diagonal dominant, i.e.,

$$\left|(\Lambda(x,x)-\Omega(x,x)) - \sum_{\substack{y\in\mathcal{V}\\ y\neq x}} \Omega(x,y)\right| = \sum_{y\in\mathcal{V}_{\mathcal{I}}^{\alpha}} \Theta(x,y)\alpha^2(x,y) \geq 0, \qquad x\in\mathcal{V}, \tag{A.43}$$

and therefore positive semidefinite, which shows $\lambda_1^D \geq 0$. In order to show that in fact the strict inequality $\lambda_1^D > 0$ holds, let $f\in\mathcal{F}_{\mathcal{V}}$ be such that equality is achieved in (5.20). We distinguish constant and nonconstant functions f. For constant $f = c\mathbb{1}$, $c\in\mathbb{R}$, since the set $\mathcal{V}_{\mathcal{I}}^{\alpha}$ given by (2.8) is nonempty, there exists an $\widetilde{x}\in\mathcal{V}$ with $\sum_{y\in\mathcal{V}_{\mathcal{I}^{\alpha}}} \Theta(\widetilde{x},y)\alpha^2(\widetilde{x},y) > 0$. Hence by (A.42e), (A.42h),

$$\lambda_1^D = \frac{\langle f,(\Lambda-\Omega)f\rangle_{\mathcal{V}}}{\langle f,f\rangle_{\mathcal{V}}} > \frac{\sum_{y\in\mathcal{V}_{\mathcal{I}^{\alpha}}} \Theta(\widetilde{x},y)\alpha^2(\widetilde{x},y)}{2n} > 0. \tag{A.44}$$

If f is nonconstant, then there exist $\widetilde{x},\widetilde{y}\in\mathcal{V}$ with $f(\widetilde{y})\neq f(\widetilde{x})$. Hence, since $\mathcal{V}$ is connected, (A.42e), (A.42h) yield

$$\lambda_1^D=\frac{\langle f,(\Lambda-\Omega)f\rangle_{\mathcal{V}}}{\langle f,f\rangle_{\mathcal{V}}}>\frac{\Omega(\widetilde{x},\widetilde{y})(f(\widetilde{x})-f(\widetilde{y}))^2}{2\max\limits_{x\in\mathcal{V}}f^2(x)}>0. \tag{A.45}$$

(ii) We perform similarly to (2.8) a disjoint decomposition of the vertex set $\mathcal{V}$ and introduce the sets

$$\mathcal{V}_i=\{x\in\mathcal{V}\colon \alpha(x,y)=0 \text{ for } y\in\mathcal{V}_{\mathcal{I}}^{\alpha}\},\qquad \mathcal{V}_b=\mathcal{V}\setminus\mathcal{V}_i. \tag{A.46}$$

Hence $\mathcal{V}_b\neq\emptyset$ if and only if $\mathcal{V}_{\mathcal{I}}^{\alpha}\neq\emptyset$ and (3.2), (3.3) yield

$$\forall x\in\mathcal{V}_i,\qquad \lambda(x)-\sum_{y\in\mathcal{V}}\Omega(x,y)=0. \tag{A.47}$$

Let f be a *normalized* eigenvector to the smallest eigenvalue $\lambda_{\min}(\Omega)$ of Ω. Then, using (A.47) and the inequality

$$(f(x)-f(y))^2\leq 2(f^2(x)+f^2(y)),\qquad x,y\in\mathcal{V},f\in\mathcal{F}_{\mathcal{V}}, \tag{A.48}$$

further yields

$$\begin{aligned}
-\lambda_{\min}(\Omega) &= -\langle f,\Omega f\rangle_{\mathcal{V}}=\langle f,(D-\Omega)f\rangle_{\mathcal{V}}-\langle f,Df\rangle_{\mathcal{V}} && \text{(A.49a)}\\
&\overset{(A.40),(A.41)}{=} \frac{1}{2}\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}}\Omega(x,y)(f(x)-f(y))^2-\sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}}\Omega(x,y)f^2(x) && \text{(A.49b)}\\
&\overset{(A.48)}{\leq} \sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}}\Omega(x,y)f^2(x) && \text{(A.49c)}\\
&\overset{(A.46)}{=} \sum_{x\in\mathcal{V}_i}\sum_{y\in\mathcal{V}}\Omega(x,y)f^2(x)+\sum_{x\in\mathcal{V}_b}\sum_{y\in\mathcal{V}}\Omega(x,y)f^2(x) && \text{(A.49d)}\\
&\overset{(2.34),(3.3)}{\leq} \sum_{x\in\mathcal{V}_i}f^2(x)+\sum_{x\in\mathcal{V}_b}\left(1-\sum_{y\in\mathcal{V}_{\mathcal{I}}^{\alpha}}\Theta(x,y)\alpha^2(x,y)\right)f^2(x) && \text{(A.49e)}\\
&= \sum_{x\in\mathcal{V}}f^2(x)-\sum_{x\in\mathcal{V}_b}\sum_{y\in\mathcal{V}_{\mathcal{I}}^{\alpha}}\Theta(x,y)\alpha^2(x,y)f^2(x) && \text{(A.49f)}\\
&\overset{(2.6)}{=} 1-\sum_{x\in\mathcal{V}_b}\left(1-\Theta(x,x)-\sum_{y\in\mathcal{V}}\Theta(x,y)\alpha^2(x,y)\right)f^2(x) && \text{(A.49g)}\\
&\overset{(3.5)}{<} 1. && \text{(A.49h)}
\end{aligned}$$

■

A.6. Proofs of section 6.1.

Proof of Lemma 6.1. Since $\overline{\mathcal{W}} \subset \mathbb{R}^{nc}$ is compact, $(S^k)_{k\geq 0} \subset \overline{\mathcal{W}}$ is bounded and there exists a convergent subsequence $(S^{k_l})_{l\geq 0}$ with $\lim_{l\to\infty} S^{k_l} = S^*$ and Λ nonempty and compact. Due to Proposition 5.3, the sequence $(J(S^k))_{k\geq 0}$ is nonincreasing and bounded from below with $\lim_{k\to\infty} J(S^k) = J^*$ for some $J^* > -\infty$.

In view of the definition (2.39) of the mapping $S \mapsto R_S(\Omega S)$, the right-hand side of (5.11) is bounded for any $S \in \mathcal{S}$. Hence the subsequence $(d^{k_l})_{l\geq 0}$ induced by $(S^{k_l})_{l\geq 0}$ through (5.11), (5.13) is convergent as well. Consequently, for any limit point $S^* \in \Lambda$, there exists a subsequence $(S^{k_l})_{l\geq 0}$ with

$$S^{k_l} \to S^* \quad \text{and} \quad d^{k_l} \to d^* \quad \text{as} \quad l \to \infty. \tag{A.50}$$

It remains to show that $\lim_{l\to\infty} J(S^{k_l}) = J(S^*) = J^*$.

Analogous to the proof of Proposition 5.1, we adopt the decomposition (A.10) of $J(S)$ by

$$J(S) = g(S) - h(S) \quad \text{with} \quad g(S) = \delta_{\overline{\mathcal{W}}}(S) + \gamma\langle S, \log S\rangle, \tag{A.51a}$$

$$h(S) = \frac{1}{2}\langle S, \Omega S\rangle + \gamma\langle S, \log S\rangle, \tag{A.51b}$$

with appropriately chosen initial decomposition parameter γ in Algorithm 4 such that g, h are strictly convex on $\mathcal{W}$. By the lower semicontinuity of $J(S)$, we have

$$\liminf_{l\to\infty} J(S^{k_l}) \geq J(S^*). \tag{A.52}$$

In addition, by invoking line 13 of Algorithm 4 defining the iterate S^{k_l} by the inclusion $\gamma\theta_{k_l-1}\widetilde{S}^{k_l-1} \in \partial g(S^{k_l})$ if θ_k satisfy the Wolfe conditions, and by line (16) otherwise, we have

$$g(S^{k_l}) - \gamma\theta_{k_l-1}\langle \widetilde{S}^{k_l-1}, S^{k_l} - S^{k_l-1}\rangle \leq g(S^*) - \gamma\theta_{k_l-1}\langle \widetilde{S}^{k_l-1}, S^* - S^{k_l-1}\rangle, \tag{A.53}$$

which after rearranging reads

$$g(S^{k_l}) \leq g(S^*) - \gamma\theta_{k_l-1}\langle d^{k_l-1}, S^* - S^{k_l}\rangle - \gamma\left\langle \log\left(\frac{S^{k_l-1}}{\mathbb{1}_c}\right), S^* - S^{k_l}\right\rangle. \tag{A.54}$$

Setting

$$\delta = \sum_{x\in\mathcal{V}} \sum_{j\in\operatorname{supp}(S^*(x))} \log(S_j^*(x)) \cdot S_j^*(x) \tag{A.55}$$

and using (A.50), we obtain for the last term

$$\lim_{l\to\infty} \left\langle \log\left(\frac{S^{k_l-1}}{\mathbb{1}_c}\right), S^* - S^{k_l}\right\rangle = \lim_{l\to\infty} \langle \log(S^{k_l-1}), S^* - S^{k_l}\rangle \tag{A.56a}$$

$$= \lim_{l\to\infty} \left(\langle \log(S^{k_l-1}) + \log(e^{\theta_{k_l-1}d^{k_l-1}}), S^* - S^{k_l}\rangle - \theta_{k_l-1}\langle d^{k_l-1}, S^* - S^{k_l}\rangle\right) \tag{A.56b}$$

$$= \lim_{l\to\infty} \left(\left\langle \log\left(\exp_{S^{k_l-1}}(\theta_{k_l-1}d^{k_l-1})\right) + \log\langle S^{k_l-1}, e^{\theta_{k_l-1}d^{k_l-1}}\rangle \mathbb{1}_c, S^* - S^{k_l}\right\rangle\right. \tag{A.56c}$$

$$\left. - \theta_{k_l-1}\langle d^{k_l-1}, S^* - S^{k_l}\rangle\right) \tag{A.56d}$$

using $\langle \mathbb{1}_c, S^* - S^{k_l}\rangle = 1 - 1 = 0$

$$\overset{(A.55)}{=} \underbrace{\lim_{l\to\infty} \langle \log(S^{k_l}), S^* - S^{k_l}\rangle}_{\to \delta - \delta = 0} - \underbrace{\lim_{l\to\infty} \langle \theta_{k_l - 1} d^{k_l - 1}, S^* - S^{k_l}\rangle}_{\to 0} \tag{A.56e}$$

$$= 0. \tag{A.56f}$$

Hence by noticing $\theta_k \in [\theta_0, \frac{1}{|\lambda_{\min}(\Omega)|}]$, the sequence (θ_{k_l}) is bounded and taking the limit in (A.54) yields

$$\limsup_{l\to\infty} g(S^{k_l}) \le g^*(S^*). \tag{A.57}$$

Now, turning to the function h of (A.51), lower semicontinuity yields $\liminf_{l\to\infty} h(S^{k_l}) \ge h(S^*)$ and hence

$$\limsup_{l\to\infty} J(S^{k_l}) = \limsup_{l\to\infty} \Big(g(S^{k_l}) - h(S^{k_l}) \Big) \le \limsup_{l\to\infty} g(S^{k_l}) - \liminf_{l\to\infty} h(S^{k_l}) \tag{A.58a}$$

$$\overset{(A.57)}{\le} g(S^*) - h(S^*). \tag{A.58b}$$

Finally, combining this with (A.52) and by uniqueness of the limit J^*, we have $J(S^*) = J^*$ for any $S^* \in \Lambda$, which completes the proof. ∎

Proof of Lemma 6.2. Throughout the proof we skip the action of projection operator Π_0 in $d^k(x)$ given by (5.11) and (5.14), due to the invariance of lifting map (2.41) by property (2.42b). By definition (5.14) of S^{k+1}, it follows for $x \in \mathcal{V}$ and $j \in J_+(S^*(x))$ that

$$\begin{aligned} \Big(S^{k+1}(x) - S^k(x)\Big)_j &= S_j^k(x) \left(\frac{e^{\theta_k d^k(x)}}{\langle S^k(x), e^{\theta_k d^k(x)}\rangle} - \mathbb{1} \right)_j \\ &= \frac{S_j^k(x)}{\langle S^k(x), e^{\theta_k d^k(x)}\rangle} \Big(e^{\theta_k d_j^k(x)} - \langle S^k(x), e^{\theta_k d^k(x)}\rangle \Big) \\ &= \frac{S_j^k(x)}{\langle S^k(x), e^{\theta_k d^k(x)}\rangle} \left(\sum_{l=0}^{\infty} \beta_{l,j}^k(x) \right) \qquad \forall J_+(S^*(x)), \end{aligned} \tag{A.59}$$

where we employed the power series of the exponential function and the shorthand $(\beta_{l,j}^k(x))_{l\ge 0}$

$$\beta_{l,j}^k(x) = \frac{\theta_k^l}{l!} \Big((d_j^k(x))^l - \langle S^k(x), (d^k(x))^l\rangle \Big) \tag{A.60a}$$

$$\overset{(5.11)}{=} \frac{\theta_k^l}{l!} \Big((\Omega S^k)_j^l(x) - \langle S^k(x), (\Omega S^k)^l(x)\rangle \Big) + \mathcal{O}(h_k). \tag{A.60b}$$

Let $M : \overline{\mathcal{W}} \times \mathbb{R}_+ \to \mathbb{R}_+$ denote the function

$$M(S,\gamma) = \max_{x\in\mathcal{V}} \max_{h\in[0,h_{\max}]} \langle S(x), e^{\gamma d(S,h)(x)}\rangle^2 \le M^*, \qquad S \in \mathcal{W}, \tag{A.61}$$

with $h_{\max} = \max_{k\geq 0} h_k$ and $d(S,h)$ as in (5.11). Since $M(S,\gamma)$ is a continuous mapping on a compact set $\mathcal{W} \times [\theta_{\min}, \theta_{\max}]$, it attains its maximum $M^* > 1$. Due to the equilibrium condition (A.33g) there exists an $\varepsilon_1 > 0$ such that, $\forall\, S \in \mathcal{W}$ with $\|S^* - S\| < \varepsilon_1$, the inequality

$$-\left((\Omega S)_j(x) - \langle \Omega S(x), S(x)\rangle\right) > -\frac{1}{\sqrt{M^*}}\left((\Omega S^*)_j(x) - \langle \Omega S^*(x), S^*(x)\rangle\right) > 0 \tag{A.62}$$

is satisfied for all indices $j \in J_+(S^*(x))$ given by (6.2) (i.e., the terms inside the brackets on either side are negative) and $x \in \mathcal{V}$. In particular, since $S^* \in \overline{\mathcal{W}}$ is a limit point of $(S^k)_{k\geq 0}$, there is a convergent subsequence $(S^{k_s})_{s\geq 0}$ with $S^{k_s} \to S^*$ and consequently $\|S^{k_{s_0}} - S^*\| < \varepsilon_1$ for some $k_{s_0} \in \mathbb{N}$. Now, using the componentwise inequality $p^l \leq p$ for $l \in \mathbb{N}$ and $p \in \mathcal{S}$, we have

$$0 \leq \left\langle \mathbb{1}, \left(S^k(x) \odot \Omega S^k(x)\right)^l \right\rangle \leq \left\langle S^k(x), (\Omega S^k(x))^l \right\rangle. \tag{A.63}$$

Employing (A.63) in (A.60) and using $h^{k_s} \to 0$ shows that there exists a smallest index $k_0 \geq k_{s_0}$ such that

$$\beta_{l,j}(x) \leq \frac{\theta_k^l}{l!}\left((\Omega S^{k_{s_0}})_j^l(x) - \langle S^{k_{s_0}}(x), (\Omega S^{k_{s_0}}(x))\rangle^l\right) + O(h^{k_{s_0}}) < 0 \quad \forall j \in J_+(S^*(x)),\, l \in \mathbb{N}. \tag{A.64}$$

Therefore, setting $\varepsilon_1 := \|S^* - S^{k_0}\|$ $\forall\, S^k$ satisfying $\|S^k - S^*\| < \varepsilon$ and $k \geq k_0$ with $\varepsilon := \min\{\varepsilon_0, \varepsilon_1\}$, the inequalities (A.62) and (A.64) are simultaneously satisfied and using

$$(\Omega S^{k_{s_0}})_j^l(x) \overset{(6.2)}{<} \langle (\Omega S^{k_{s_0}})(x), S^{k_{s_0}}(x)\rangle^l \quad \forall j \in J_+(S^*(x)), \quad l \in \mathbb{N}, \tag{A.65}$$

enables one to estimate (A.59) by

$$\left(S^{k+1}(x) - S^k(x)\right)_j = \frac{S_j^k(x)}{\langle S^k(x), e^{\theta_k d^k(x)}\rangle}\left(\sum_{l=1}^{\infty} \beta_{l,j}^k(x)\right) \tag{A.66a}$$

$$\overset{(A.64)}{\leq} \frac{S_j^k(x)}{\langle S^k(x), e^{\theta_k d^k(x)}\rangle}\Big(\theta_k\left((\Omega S^k)_j(x) - \langle S^k(x), \Omega S^k(x)\rangle\right) \tag{A.66b}$$

$$+ \sum_{l=2}^{\infty} \frac{\theta_k^l}{l!}\left((\Omega S^k)_j^l(x) - \langle S^k(x), \Omega S^k(x)\rangle^l\right) + \mathcal{O}(h^k)\Big) \tag{A.66c}$$

$$\overset{(A.62)}{\leq} \frac{-S_j^k(x)}{\langle S^k(x), e^{\theta_k d^k(x)}\rangle \cdot \sqrt{M^*}}\left(\theta_k\left(\langle \Omega S^*(x), S^*(x)\rangle - (\Omega S^*)_j(x)\right)\right) \tag{A.66d}$$

$$\overset{(A.61)}{\leq} -\theta_k \frac{S_j^k(x)}{M^*}\left(\langle \Omega S^*(x), S^*(x)\rangle - (\Omega S^*)_j(x)\right) \quad \forall J_+(S^+(x)). \tag{A.66e}$$

Taking the sum over $x \in \mathcal{V}$ shows (6.3). ∎

A.7. Proofs of section 6.2.

Proof of Theorem 6.4. Let $S^* \in \Lambda$ be a limiting point of $(S^k)_{k\geq 0}$ with $S^*(x) \in \overline{\mathcal{S}} \setminus \mathcal{S}$, $\forall x \in \mathcal{V}$, by Proposition 5.3(iii), and let $\theta_k \in \mathbb{R}_+$, $S^{k+1} \in \mathcal{W}$, and $\widetilde{S}^k$ be determined by Algorithm 4 (see

lines 13 and 14), respectively. Then, by the well-known *three-point identity* [62, Lemma 3.1] with respect to $S^{k+1}, S^k \in \mathcal{W}$, $S^* \in \overline{\mathcal{W}}$, one has

$$D_{\mathrm{KL}}(S^*, S^{k+1}) - D_{\mathrm{KL}}(S^*, S^k) = -D_{\mathrm{KL}}(S^{k+1}, S^k) - \langle \nabla f(S^{k+1}) - \nabla f(S^k), S^* - S^{k+1} \rangle. \tag{A.67}$$

Recalling step size selection 3 it holds that $\theta_k \in (\theta_0, \frac{1}{|\lambda_{\min}(\Omega)|})$ and leveraging the DC-decomposition (A.51) with $\gamma = \frac{1}{\theta_k}$, the inclusion $\Omega S^k + \frac{1}{\theta_k} \log(\frac{S^k}{\mathbb{1}_c}) \in \partial h(S^k)$ and the strict convexity of $h(S)$ on $\mathcal{W}$ imply by the gradient inequality

$$h(S^{k+1}) - h(S^k) - \left\langle \Omega S^k + \frac{1}{\theta_k} \log\left(\frac{S^k}{\mathbb{1}_c}\right), S^{k+1} - S^k \right\rangle > 0 \tag{A.68}$$

and hence

$$h(S^{k+1}) - h(S^k) - \left\langle \Omega S^k + \frac{1}{\theta_k} \log\left(\frac{S^k}{\mathbb{1}_c}\right), S^{k+1} - S^k \right\rangle \tag{A.69a}$$

$$\overset{(A.51b)}{=} \frac{1}{2}\langle S^{k+1}, \Omega S^{k+1} \rangle - \frac{1}{2}\langle S^k, \Omega S^k \rangle \tag{A.69b}$$

$$+ \frac{1}{\theta_k}\left(\langle S^{k+1}, \log(S^{k+1}) \rangle - \langle S^k, \log S^k \rangle - \left\langle \log\left(\frac{S^k}{\mathbb{1}_c}\right), S^{k+1} - S^k \right\rangle \right) \tag{A.69c}$$

$$- \langle \Omega S^k, S^{k+1} - S^k \rangle \tag{A.69d}$$

$$\overset{(2.46),(6.4)}{=} J(S^k) - J(S^{k+1}) + \frac{1}{\theta_k} D_{\mathrm{KL}}(S^{k+1}, S^k) - \langle \Omega S^k, S^{k+1} - S^k \rangle. \tag{A.69e}$$

Therefore inequality (A.68) is equivalent to

$$-D_{\mathrm{KL}}(S^{k+1}, S^k) \leq \theta_k \left(J(S^k) - J(S^{k+1}) - \langle \Omega S^k, S^{k+1} - S^k \rangle \right). \tag{A.70}$$

Combining (A.70) and (A.67) yields

$$\begin{aligned} D_{\mathrm{KL}}(S^*, S^{k+1}) - D_{\mathrm{KL}}(S^*, S^k) \leq{}& \theta_k \left(J(S^k) - J(S^{k+1}) - \langle \Omega S^k, S^{k+1} - S^k \rangle \right) \\ & - \langle \nabla f(S^{k+1}) - \nabla f(S^k), S^* - S^{k+1} \rangle. \end{aligned} \tag{A.71}$$

Next, in view of Algorithm 4, line 14, we rewrite the last term in (A.71) in the form

$$\langle \nabla f(S^{k+1}) - \nabla f(S^k), S^* - S^{k+1} \rangle \overset{S^k, S^{k+1} \in \mathcal{W}}{\overset{(6.7)}{=}} \langle \mathbb{1}_c + \log(S^{k+1}) - (\mathbb{1}_c + \log(S^k)), S^* - S^{k+1} \rangle \tag{A.72a}$$

$$\overset{\text{Algorithm 4 line 14}}{=} \langle \log(S^k) + \log(e^{\theta_k d^k}) - \log(S^k), S^* - S^{k+1} \rangle \tag{A.72b}$$

$$- \underbrace{\langle \log(\langle S^k, e^{\theta_k d^k} \rangle)\mathbb{1}_c, S^* - S^{k+1} \rangle}_{=0} \tag{A.72c}$$

$$= \theta_k \langle d^k, S^* - S^{k+1} \rangle. \tag{A.72d}$$

Consequently, (A.71) becomes

$$
\begin{aligned}
&\text{(A.73a)} \quad D_{\mathrm{KL}}(S^*, S^{k+1}) - D_{\mathrm{KL}}(S^*, S^k) \\
&\text{(A.73b)} \qquad \leq \ \theta_k \Big(J(S^k) - J(S^{k+1}) \Big) - \theta_k \langle \Omega S^k, S^* - S^k \rangle - \frac{\theta_k h_k}{2} \langle \Omega R_{S^k}(\Omega S^k), S^* - S^{k+1} \rangle \\
&\text{(A.73c)} \qquad \overset{(2.46)}{=} \theta_k \Big(2 \Big(J(S^*) - J(S^{k+1}) \Big) + J(S^{k+1}) - J(S^k) \\
&\text{(A.73d)} \qquad\quad - \frac{h_k}{2} \langle \Omega R_{S^k}(\Omega S^k), S^* - S^{k+1} \rangle - \langle S^k, \Omega S^* \rangle - 2 J(S^*) \Big).
\end{aligned}
$$

Using the inequality of Cauchy–Schwarz and taking into account $S^* \in \overline{W}$, $S \in \mathcal{W}$, we estimate with $\lambda(\Omega)$ defined by (6.9b)

$$
\text{(A.74)} \qquad |\langle \Omega R_S(\Omega S), S^* - S \rangle| \leq \| \Omega R_S(\Omega S) \| \cdot \| S^* - S \| \leq \frac{\lambda^2(\Omega)}{2} \| S \| \sqrt{n} \leq \frac{\lambda^2(\Omega) \cdot n}{2},
$$

where the factor $\frac{1}{2}$ is due to the fact that the matrices $R_{S(x)}$ given by (2.36) are positive semidefinite with $\lambda_{\max}(R_{S(x)}) \leq \frac{1}{2}$, which easily follows from Gershgorin's circle theorem. Using the descent step based on (5.11) and (A.23), we consider three further terms of (A.73):

$$
\begin{aligned}
&\text{(A.75a)} \quad J(S^{k+1}) - J(S^k) - \frac{h_k}{2} \langle \Omega R_{S^k}(\Omega S^k), S^* - S^{k+1} \rangle \\
&\text{(A.75b)} \qquad \overset{(5.15a)}{\leq} \theta_k c_1 \underbrace{\langle R_{S^k}(\Omega S^k), R_{S^k}(d^k) \rangle_{S^k}}_{\leq 0} - \frac{h_k}{2} \langle \Omega R_{S^k}(\Omega S^k), S^* - S^{k+1} \rangle \\
&\text{(A.75c)} \qquad \overset{(5.11)}{\leq} -\theta_k c_1 (\langle R_{S^k}(\Omega S^k), R_{S^k}(\Omega S^k) \rangle_{S^k} \\
&\text{(A.75d)} \qquad\quad + \frac{\theta_k c_1 h_k}{2} |\langle R_{S^k}(\Omega S^k), R_{S^k} \Omega R_{S^k} \Omega S^k \rangle_{S^k}|) + \frac{h_k}{2} |\langle \Omega R_{S^k}(\Omega S^k), S^* - S^{k+1} \rangle| \\
&\text{(A.75e)} \qquad \overset{(A.22),(A.74)}{\leq} -\frac{\theta_k c_1}{2} \langle R_{S^k}(\Omega S^k), R_{S^k}(\Omega S^k) \rangle_{S^k} + \frac{\lambda^2(\Omega) n h_k}{4} \\
&\text{(A.75f)} \qquad = \ -\frac{\theta_k c_1}{2} \| \operatorname{grad} J(S^k) \|_{S^k}^2 + \frac{\lambda^2(\Omega) n h_k}{4} \\
&\text{(A.75g)} \qquad \leq \ 0,
\end{aligned}
$$

where the last inequality holds due to assumption (6.9). Now we focus on the last remaining term occurring in (A.73). Using the index sets (6.2) with respect to the limit point $S^* \in \overline{\mathcal{W}}$ along with $S^k(x) \in \mathcal{S}$, we get

$$
\begin{aligned}
&\text{(A.76a)} \quad -\langle S^k, \Omega S^* \rangle - 2 J(S^*) \overset{(2.46)}{=} -\sum_{x \in \mathcal{V}} \langle S^k(x), \Omega S^*(x) \rangle + \sum_{x \in \mathcal{V}} \langle S^*(x), \Omega S^*(x) \rangle \\
&\text{(A.76b)} \qquad = \ -\sum_{x \in \mathcal{V}} \sum_{j \in [c]} S_j^k(x) (\Omega S^*)_j(x) + \sum_{x \in \mathcal{V}} \underbrace{\sum_{j \in [c]} S_j^k(x)}_{=1} \langle S^*(x), \Omega S^*(x) \rangle
\end{aligned}
$$

$$= -\sum_{x\in\mathcal{V}}\sum_{j\in[c]} S_j^k(x)\left((\Omega S^*)_j(x) - \langle S^*(x), \Omega S^*(x)\rangle\right) \tag{A.76c}$$

$$\overset{(6.2)}{=} -\sum_{x\in\mathcal{V}}\Bigg(\sum_{j\in J_-(S^*(x))} S_j^k(x)\left((\Omega S^*)_j(x) - \langle S^*(x), \Omega S^*(x)\rangle\right) \tag{A.76d}$$

$$+ \sum_{j\in J_+(S^*(x))} S_j^k(x)\left((\Omega S^*)_j(x) - \langle S^*(x), \Omega S^*(x)\rangle\right)\Bigg). \tag{A.76e}$$

As a result, combining (A.75) and (A.76) $\forall k \geq K$ and using $J(S^*) - J(S^{k+1}) < 0$, (A.73) becomes

$$D_{\mathrm{KL}}(S^*, S^{k+1}) - D_{\mathrm{KL}}(S^*, S^k) \leq \theta_k\Bigg(J(S^*) - J(S^{k+1}) - \sum_{x\in\mathcal{V}}\Bigg(\sum_{j\in J_-(S^*(x))} S_j^k(x)((\Omega S^*)_j(x) \tag{A.77a}$$

$$- \langle S^*(x), \Omega S^*(x)\rangle) + \sum_{j\in J_+(S^*(x))} S_j^k(x)\left((\Omega S^*)_j(x) - \langle S^*(x), \Omega S^*(x)\rangle\right)\Bigg)\Bigg). \tag{A.77b}$$

By Lemma 6.2, there exist $\varepsilon > 0$ and $k_0 \in \mathbb{N}$ such that $\forall S^k \in \mathcal{W}$ with $k \geq k_0$ and $\|S^k - S^*\| < \varepsilon$ inequality (6.3) is satisfied, where

$$Q(S) = \sum_{x\in\mathcal{V}}\sum_{j\in J_+(S^*(x))} S_j(x).$$

Introducing the mapping

$$V\colon \mathcal{W} \to \mathbb{R}_+, \qquad V(S) = D_{\mathrm{KL}}(S^*, S) + M^*\, Q(S)$$

with $M^* > 1$ as in Lemma 6.2, we obtain

$$\begin{aligned} V(S^{k+1}) - V(S^k) &= D_{\mathrm{KL}}(S^*, S^{k+1}) - D_{\mathrm{KL}}(S^*, S^k) + M^*\left(Q(S^{k+1}) - Q(S^k)\right) \\ &\overset{\substack{(6.2a)\\(A.77)}}{\leq} \theta_k\Big(J(S^*) - J(S^k) - \sum_{x\in\mathcal{V}}\sum_{j\in J_-(S^*(x))} S_j^k(x)\left((\Omega S^*)_j(x) - \langle S^*(x), \Omega S^*(x)\rangle\right)\Big). \end{aligned} \tag{A.78}$$

By Lemma 6.1 $J(S)$ is constant on the set of limit points of the sequence (S^k) and the right-hand side of (A.78) is strictly negative unless S^k is a stationary point of $J(S)$. Consequently, (A.78) is *strictly* negative $\forall k \geq k_0$ with $\|S^k - S^*\| < \varepsilon$. Consider $U_\delta = \{S \in \overline{\mathcal{W}}\colon V(S) < \delta\}$ with δ small enough such that $U_\delta \subset \{S \in \overline{\mathcal{W}}\colon \|S - S^*\| < \epsilon\}$. Then, as $S^* \in \Lambda$ is a limit point, there exists an index $K \geq k_0$ such that $S^K \in U_\delta$ and $(S^k)_{k\geq K} \subset U_\delta$ due to $V(S^{K+1}) < V(S^K) < \delta$ by (A.78). Therefore, for $k \geq K$ we conclude

$$0 \leq D_{\mathrm{KL}}(S^*, S^k) \leq V(S^k) \to 0 \quad \text{for} \quad k \to \infty, \tag{A.79}$$

which shows $S^k \to S^*$. ∎

Proof of Theorem 6.6. For $\varepsilon > 0$ let $k \in \mathbb{N}$ be such that $S^k \in B_\varepsilon(S^*)$. Then, with $S^{k+\frac{1}{2}}, S^{k+1} \in \mathcal{W}$ given by (5.12) and taking into account assumption (6.10), we have for any $x \in \mathcal{V}$ with $S^*(x) = e_{j^*(x)}$

$$\|S^{k+1}(x) - S^*(x)\|_1 = \sum_{j \in [c] \setminus j^*(x)} S_j^{k+1}(x) + 1 - S_{j^*(x)}^{k+1}(x) \tag{A.80a}$$

$$= 2 - 2S_{j^*(x)}^{k+1}(x) \tag{A.80b}$$

$$\overset{(5.12)}{=} 2 - 2\frac{S_{j^*(x)}^k(x) e^{\theta_k (\Omega S^k)_{j^*(x)}(x) + \frac{\theta_k h_k}{2} (\Omega R_{S^k}(\Omega S^k))_{j^*(x)}(x)}}{\langle S^k(x), e^{\theta_k (\Omega S^k)(x) + \frac{\theta_k h_k}{2} \Omega R_{S^k}(\Omega S^k)(x)} \rangle} \tag{A.80c}$$

$$= 2 - \frac{2 S_{j^*(x)}^k(x)}{S_{j^*(x)}^k(x) + \sum_{j \neq j^*(x)} S_j^k(x) e^{-\theta_k H_j(x)}}, \tag{A.80d}$$

with the shorthand

$$H_j(x) := (\Omega S^k)_{j^*(x)}(x) - (\Omega S^k)_j(x) + \frac{h_k}{2} \left((\Omega R_{S^k}(\Omega S^k))_{j^*(x)}(x) - (\Omega R_{S^k}(\Omega S^k))_j(x) \right). \tag{A.81}$$

We consider the first two terms of the right-hand side of (A.81). Since $S^k(x) \in B_\varepsilon(S^*)$, we have

$$S_{j^*(x)}^k(x) > 1 - \frac{\varepsilon}{2}, \qquad S_j^k(x) < \frac{\varepsilon}{2} \quad \forall \quad j \neq j^*(x) \tag{A.82}$$

and get

$$\begin{aligned} (\Omega S)_{j^*(x)}(x) - (\Omega S)_j(x) &\overset{(2.48)}{=} \sum_{y \in \mathcal{N}(x)} \Omega(x,y) S_{j^*(x)}(y) - \sum_{y \in \mathcal{N}(x)} \Omega(x,y) S_j(y) \\ &= \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) = j^*(x)}} \Omega(x,y) S_{j^*(x)}(y) + \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) \neq j^*(x)}} \Omega(x,y) S_{j^*(x)}(y) \\ &\quad - \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) = j}} \Omega(x,y) S_j(y) - \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) \neq j}} \Omega(x,y) S_j(y). \end{aligned} \tag{A.83a}$$

Skipping the nonnegative second term and applying the constraint $S_j(y) < 1$ for indices $j^*(y) = j$, it follows with (A.82)

$$(\Omega S)_{j^*(x)}(x) - (\Omega S)_j(x) > \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) = j^*(x)}} \Omega(x,y) S_{j^*(x)}(y) - \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) = j}} \Omega(x,y) - \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) \neq j}} \Omega(x,y) S_j(y) \tag{A.83b}$$

$$\overset{(A.82)}{>} \left(1 - \frac{\varepsilon}{2}\right) \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) = j^*(x)}} \Omega(x,y) - \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) = j}} \Omega(x,y) - \frac{\varepsilon}{2} \sum_{\substack{y \in \mathcal{N}(x) \\ j^*(y) \neq j}} \Omega(x,y) \tag{A.83c}$$

and after rewriting the last sum as $1-\sum_{\substack{y\in\mathcal{N}(x)\\ j^*(x)=j}}\Omega(x,y)$ and using $S^*(x)=e_{j^*(x)}$

$$\geq \left(1-\frac{\varepsilon}{2}\right)\left((\Omega S^*)_{j^*(x)}-(\Omega S^*)_j\right)(x)-\frac{\varepsilon}{2}. \tag{A.83d}$$

Now we consider the last two terms of the right-hand side of (A.81), starting with the expression $R_{S^k}(\Omega S^k)$. As $\overline{B_\varepsilon(S^*)}$ is compact, the maximum

$$\rho^*=\max_{S\in\overline{B_\varepsilon(S^*)}}\rho(S),\qquad \rho(S)=\max_{x\in\mathcal{V}}\max_{l\in[c]\setminus j^*(x)}\left((\Omega S)_{j^*(x)}-(\Omega S)_l\right)(x) \tag{A.84}$$

is attained. For $j\in[c]$ with $(R_{S^k}(\Omega S^k))_j(x)<0$, we get

$$\left(R_{S^k}(\Omega S^k)\right)_j(x)=S_j^k(x)\left((\Omega S^k)_j(x)-\langle S^k(x),(\Omega S^k)(x)\rangle\right) \tag{A.85a}$$

$$=S_j^k(x)\left(\sum_{l\neq j}S_l^k(x)\left((\Omega S^k)_j(x)-(\Omega S^k)_l(x)\right)\right). \tag{A.85b}$$

Taking into account (6.12) for $S^k\in B_\varepsilon(S^*)$, we have $(\Omega S^k)_{j^*(x)}(x)>(\Omega S^k)_l(x)\ \ \forall\, l\in[c]\setminus j^*(x)$ by (6.11), and due to $R_{S^k}(\Omega S^k)_j(x)<0$, we conclude $j\neq j^*(x)$ in the preceding equation. Consequently, applying the second inequality in (A.82) further yields

$$\left(R_{S^k}(\Omega S^k)\right)_j(x)\overset{(A.82)}{>}\frac{\varepsilon}{2}\sum_{l\neq j}S_l^k(x)\left((\Omega S^k)_j-(\Omega S^k)_l\right)(x) \tag{A.85c}$$

$$\overset{(6.10)}{\geq}\frac{\varepsilon}{2}\sum_{l\neq j}S_l^k(x)\left((\Omega S^k)_j-(\Omega S^k)_{j^*(x)}\right)(x) \tag{A.85d}$$

$$=\frac{\varepsilon}{2}(1-S_j^k(x))\left((\Omega S^k)_j-(\Omega S^k)_{j^*(x)}\right)(x) \tag{A.85e}$$

$$\overset{(A.84)}{\geq}-\frac{\varepsilon}{2}\rho^*. \tag{A.85f}$$

In view of the last two terms of the right-hand side of (A.81), we introduce the index sets

$$\begin{aligned}\mathcal{N}_-^j(x)&:=\{y\in\mathcal{N}(x)\colon\ (R_S(\Omega S))_j(y)<(R_S(\Omega S))_{j^*(x)}(y)\},\\ \mathcal{N}_+^j(x)&:=\{y\in\mathcal{N}(x)\colon\ (R_S(\Omega S))_j(y)>(R_S(\Omega S))_{j^*(x)}(y)\}\end{aligned} \tag{A.86}$$

and estimate

$$(\Omega R_{S^k}(\Omega S^k))_{j^*(x)}(x)-(\Omega R_{S^k}(\Omega S^k))_j(x)=\sum_{y\in\mathcal{N}(x)}\Omega(x,y)\left(R_{S^k}(\Omega S^k)_{j^*(x)}-R_{S^k}(\Omega S^k)_j\right)(y) \tag{A.87a}$$

$$\geq\sum_{y\in\mathcal{N}_+^j(x)}\Omega(x,y)\left(R_{S^k}(\Omega S^k)_{j^*(x)}-R_{S^k}(\Omega S^k)_j\right)(y). \tag{A.87b}$$

Regarding the term $(\cdots)$ in parentheses, using $\mathbb{1}^\top R_{S^k}=0^\top$ and consequently $\sum_{l\in[c]}(R_{S^k}(\Omega S^k))_l(y)=0$ for $y\in\mathcal{N}_+^j(x)$, it follows that

(A.88a)
$$R_{S^k}(\Omega S^k)_{j^*(x)}(y) - R_{S^k}(\Omega S^k)_j(y) = 2(R_{S^k}(\Omega S^k))_{j^*(x)}(y) + \sum_{\substack{l\in[c]\\ l\notin\{j^*(x),j\}}} (R_{S^k}(\Omega S^k))_l(y)$$

(A.88b)
$$\geq 2c \min_{l\in[c]\setminus j^*(y)} (R_{S^k}(\Omega S^k))_l(y)$$

(A.88c)
$$\overset{\text{(A.87)}}{>} -\varepsilon c\rho^*.$$

Consequently, applying (A.88) and $\Omega(x,y)\leq 1$, inequality (A.87) becomes

(A.89)
$$\left(\left(\Omega R_{S^k}(\Omega S^k)\right)_{j^*(x)} - \left(\Omega R_{S^k}(\Omega S^k)\right)_j\right)(x) > -\varepsilon|\mathcal{N}(x)|c\rho^*.$$

Substituting this estimate and (A.83) into (A.81) yields for any $x\in\mathcal{V}$ and $j\in[c]\setminus\{j^*(x)\}$

(A.90)
$$H_j(x) \geq \left(1-\frac{\varepsilon}{2}\right)((\Omega S^*)_{j^*(x)} - (\Omega S^*)_j)(x) - \frac{\varepsilon}{2} - \frac{\overline{h}c}{2}\varepsilon|\mathcal{N}(x)|\rho^*, \quad \overline{h} = \max_{k\geq k_0} h_k.$$

Thus, returning to (A.80), we finally obtain $\forall\varepsilon$ satisfying (6.15) and using

(A.91)
$$H^*(x) := \min_{j\neq j^*(x)} H_j(x) > 0$$

the bound

(A.92a)
$$\|S^{k+1}(x) - S^*(x)\|_1 \leq 2 - \frac{2S^k_{j^*(x)}(x)}{S^k_{j^*(x)}(x) + \sum_{j\neq j^*(x)} S^k_j(x)e^{-\theta_k H^*(x)}}$$

(A.92b)
$$= \frac{2\left(1-S^k_{j^*(x)}(x)\right)e^{-\theta_k H^*(x)}}{S^k_{j^*(x)}(x) + \left(1-S^k_{j^*(x)}(x)\right)e^{-\theta_k H^*(x)}}$$

(A.92c)
$$\overset{S^k_{j^*(x)}(x)=e_{j^*(x)}}{=} \|S^k(x) - S^*\|_1 \underbrace{\frac{e^{-\theta_k H^*(x)}}{S^k_{j^*(x)}(x) + \left(1-S^k_{j^*(x)}(x)\right)e^{-\theta_k H^*(x)}}}_{=:\xi(x)<1,\ \text{if } H^*(x)>0.}$$

(A.92d)
$$=: \|S^k(x) - S^*\|_1 \cdot \xi(x)$$

with $\xi(x)<1$, since $H^*(x)>0$ by (A.91). Induction over $k>k_0$ yields

(A.93)
$$\|S^{k+1}(x) - S^*(x)\|_1 < \xi^{k-k_0}(x)\|S^{k_0}(x) - S^*(x)\|_1,$$

which proves (6.16). ∎

REFERENCES

[1] G. Gilboa and S. Osher, *Nonlocal linear image regularization and supervised segmentation*, Multiscale Model. Simul., 6 (2007), pp. 595–630.

[2] A. Elmoataz, O. Lezoray, and S. Bougleux, *Nonlocal discrete regularization on weighted graphs: A framework for image and manifold processing*, IEEE Trans. Image Process., 17 (2008), pp. 1047–1059.
[3] G. Gilboa and S. Osher, *Nonlocal operators with applications to image processing*, Multiscale Model. Simul., 7 (2009), pp. 1005–1028.
[4] A. Buades, B. Coll, and J. M. Morel, *Image denoising methods. A new nonlocal principle*, SIAM Rev., 52 (2010), pp. 113–147.
[5] A. Elmoataz, M. Toutain, and D. Tenbrinck, *On the p-Laplacian and ∞-Laplacian on graphs with applications in image and data processing*, SIAM J. Imaging Sci., 8 (2015), pp. 2412–2451.
[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.
[7] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural ordinary differential equations*, in Proceedings of Neural Information Processing Systems, 2018.
[8] M. Wainwright and M. Jordan, *Graphical models, exponential families, and variational inference*, Found. Trends Mach. Learn., 1 (2008), pp. 1–305.
[9] J. Kappes, B. Andres, F. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, and C. Rother, *A comparative study of modern inference techniques for structured discrete energy minimization problems*, Int. J. Comput. Vis., 115 (2015), pp. 155–184.
[10] B. Merriman, J. Bence, and S. Osher, *Motion of multiple junctions: A level set approach*, J. Comput. Phys., 112 (1994), pp. 334–363.
[11] Y. van Gennip, N. Guillen, B. Osting, and A. L. Bertozzi, *Mean curvature, threshold dynamics, and phase field theory on finite graphs*, Milan J. Math., 82 (2014), pp. 3–65.
[12] A. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, SIAM Rev., 58 (2016), pp. 293–328.
[13] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[14] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[15] H. Garcke, *Curvature driven interface evolution*, Jahresber. Dtsch. Math.-Ver., 115 (2013), pp. 63–100.
[16] V. Caselles, A. Chambolle, and M. Novaga, *Total variation in imaging*, in Handbook of Mathematical Methods in Imaging, O. Scherzer, ed., Springer, New York, 2015, pp. 1455–1499.
[17] F. Åström, S. Petra, B. Schmitzer, and C. Schnörr, *Image labeling by assignment*, J. Math. Imaging Vision, 58 (2017), pp. 211–238.
[18] C. Schnörr, *Assignment flows*, in Variational Methods for Nonlinear Geometric Data and Applications, P. Grohs, M. Holler, and A. Weinmann, eds., Springer, New York, 2020, pp. 235–260.
[19] A. Zern, A. Zeilmann, and C. Schnörr, *Assignment flows for data labeling on graphs: Convergence and stability*, Inform. Geom., 5 (2022), pp. 355–404, https://doi.org/10.1007/s41884-021-00060-8.
[20] A. Zeilmann, F. Savarino, S. Petra, and C. Schnörr, *Geometric numerical integration of the assignment flow*, Inverse Problems, 36 (2020), 034004.
[21] A. Zern, M. Zisler, S. Petra, and C. Schnörr, *Unsupervised assignment flow: Label learning on feature manifolds by spatially regularized geometric assignment*, J. Math. Imaging Vision, 62 (2020), pp. 982–1006.
[22] M. Zisler, A. Zern, S. Petra, and C. Schnörr, *Self-assignment flows for unsupervised data labeling on graphs*, SIAM J. Imaging Sci., 13 (2020), pp. 1113–1156.
[23] R. Hühnerbein, F. Savarino, S. Petra, and C. Schnörr, *Learning adaptive regularization for image labeling using geometric assignment*, J. Math. Imaging Vision, 63 (2021), pp. 186–215.
[24] A. Zeilmann, S. Petra, and C. Schnörr, *Learning linear assignment flows for image labeling via exponential integration*, in Scale Space and Variational Methods in Computer Vision, Lecture Notes in Comput. Sci. 12679, A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, and L. Simon, eds., Springer, New York, 2021, pp. 385–397.
[25] A. Zeilmann, S. Petra, and C. Schnörr, *Learning linearized assignment flows for image labeling*, J. Math. Imaging Vision, 65 (2023), pp. 164–184.
[26] B. Boll, A. Zeilmann, S. Petra, and C. Schnörr, *Self-Certifying Classification by Linearized Deep Assignment*, preprint, https://arxiv.org/abs/2201.11162, 2022.

[27] Q. Du, M. Gunzburger, R. B. Lehoucq, and K. Zhou, *Analysis and approximation of nonlocal diffusion problems with volume constraints*, SIAM Rev., 54 (2012), pp. 667–696.
[28] Q. Du, M. Gunzburger, R. B. Lehoucq, and K. Zhou, *A nonlocal vector calculus, nonlocal volume-constrained problems, and nonlocal balance laws*, Math. Models Methods Appl. Sci., 23 (2013), pp. 493–540.
[29] F. Savarino and C. Schnörr, *Continuous-domain assignment flows*, European J. Appl. Math., 32 (2021), pp. 570–597.
[30] L. Alvarez, F. Guichard, P. L. Lions, and J. M. Morel, *Axioms and fundamental equations of image processing*, Arch. Ration. Mech. Anal., 123 (1993), pp. 199–257.
[31] J. Weickert, *Anisotropic Diffusion in Image Processing*, B.G. Teubner, Leipzig, 1998.
[32] R. Horst, and N. V. Thoai, *DC programming: Overview*, J. Optim. Theory Appl., 103 (1999), pp. 1–43.
[33] L. T. Hoai An and T. Pham Dinh, *The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems*, Ann. Oper. Res., 133 (2005), pp. 23–46.
[34] A. Beck and M. Teboulle, *Smoothing and first order methods: A unified framework*, SIAM J. Optim., 22 (2012), pp. 557–580.
[35] W. Krichene, A. Bayen, and P. Bartlett, *Adaptive averaging in accelerated descent dynamics*, in Proceedings of Neural Information Processing Systems, 2016.
[36] M. Fazylab, A. Ribeiro, M. Morari, and V. M. Preciado, *Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems*, SIAM J. Optim., 28 (2018), pp. 2654–2689.
[37] F. Andreu-Vaillo, J. M. Mazón, J. D. Rossi, and J. J. Toledo-Melero, *Nonlocal Diffusion Problems*, AMS, Providence, RI, 2010.
[38] Q. Du, *Nonlocal Modeling, Analysis, and Computation*, SIAM, Philadelphia, 2019.
[39] F. Chung and R. P. Langlands, *A combinatorial Laplacian with vertex weights*, J. Combin. Theory Ser. A, 5 (1996), pp. 316–327.
[40] F. Chung, *Spectral Graph Theory*, AMS, Providence, RI, 1997.
[41] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry,* AMS, Providence, RI, 2000.
[42] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer, *Information Geometry*, Springer, New York, 2017.
[43] J. Jost, *Riemannian Geometry and Geometric Analysis*, 7th ed., Springer-Verlag, Berlin, 2017.
[44] D. Sitenko, B. Boll, and C. Schnörr, *Assignment flow for order-constrained OCT segmentation*, Int. J. Comput. Vis., 129 (2021), pp. 3088–3118.
[45] M. Welk and J. Weickert, *PDE evolutions for M-smoothers in one, two, and three dimensions*, J. Math. Imaging Vision, 63 (2020), pp. 157–185.
[46] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain non-convex minimization problems*, Internat. J. Systems Sci., 12 (1981), pp. 989–1000.
[47] F. J. Aragón Artacho, R. Fleming, and P. T. Vuong, *Accelerating the DC algorithm for smooth functions*, Math. Program., 169 (2018), pp. 95–118.
[48] Y. Dai and Y. Yuan, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim., 10 (1999), pp. 177–182.
[49] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.
[50] R. T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
[51] L. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.
[52] H. H. Bauschke and J. M. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.
[53] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM J. Optim., 28 (2018), pp. 2131–2151.
[54] D. N. Phan, H. M. Le, and H. A. L. Thi, *Accelerated difference of convex functions algorithm and its application to sparse binary logistic regression*, in Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 1369–1375.

[55] R. B. Lehoucq, D. C. Sorenson, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998.
[56] H. A. L. Thi and T. P. Dinh,*DC programming and DCA: Thirty years of developments*, Math. Program., 169 (2018), pp. 5–68.
[57] D. Gonzalez-Alvarado, A. Zeilmann, and C. Schnörr, *Quantifying uncertainty of image labelings using assignment flows*, in Pattern Recognition, Lecture Notes in Comput. Sci. 13024, Springer, New York, 2021, pp. 453–466.
[58] F. Savarino, P. Albers, and C. Schnörr, *On the geometric mechanics of assignment flows for metric data labeling*, in Scale Space and Variational Methods in Computer Vision, Lecture Notes in Comput. Sci. 12679, A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, and L. Simon, eds., Springer, New York, 2021, pp. 398–410.
[59] F. Savarino, P. Albers, and C. Schnörr, *On the Geometric Mechanics of Assignment Flows for Metric Data Labeling*, CoRR, abs/2111.02543, 2021.
[60] P. Hartman, *On functions representable as a difference of convex functions*, Pacific J. Math., 9 (1959), pp. 707–713.
[61] A. Graham, *Kronecker Products and Matrix Calculus: With Applications*, Ellis Horwood, 1981.
[62] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.