The Assignment Manifold: A Smooth Model for Image Labeling

Freddie Åström*, Stefania Petra†, Bernhard Schmitzer‡ and Christoph Schnörr§
*HCI, †MIG, §IPA, Heidelberg University, Germany
‡CEREMADE, University Paris-Dauphine, France

Abstract

We introduce a novel geometric approach to the image labeling problem. A general objective function is defined on a manifold of stochastic matrices, whose elements assign prior data that are given in any metric space, to observed image measurements. The corresponding Riemannian gradient flow entails a set of replicator equations, one for each data point, that are spatially coupled by geometric averaging on the manifold. Starting from uniform assignments at the barycenter as natural initialization, the flow terminates at some global maximum, each of which corresponds to an image labeling that uniquely assigns the prior data. No tuning parameters are involved, except for two parameters setting the spatial scale of geometric averaging and scaling globally the numerical range of features, respectively. Our geometric variational approach can be implemented with sparse interior-point numerics in terms of parallel multiplicative updates that converge efficiently.

1. Introduction

Image labeling amounts to determining a *partition* of the image domain by uniquely assigning to each pixel a single element from a finite set of labels. The mutual dependency on other assignments gives rise to a global objective function whose minima correspond to favorable label assignments and partitions. Because the problem of computing globally optimal partitions generally is NP-hard, *relaxations* of the variational problem only define computationally feasible optimization approaches.

Relaxations of the variational image labeling problem fall into two categories: *convex and non-convex relaxations*. The dominant *convex approach* is based on the local-polytope relaxation, a particular linear programming (LP-) relaxation [17]. This has spurred a lot of research on developing specific algorithms for efficiently solving large problem instances, as they often occur in applications. We mention [10] as a prominent example and refer to [7] for a comprehensive evaluation. Yet, models with higher connectivity in terms of objective functions with local potentials that are defined on larger cliques, are still difficult to solve efficiently. A major reason that has been largely motivating our present work is the *non-smoothness* of optimization problems resulting from convex relaxation, corresponding problem splittings and the resulting dual objective functions – the price to pay for convexity.

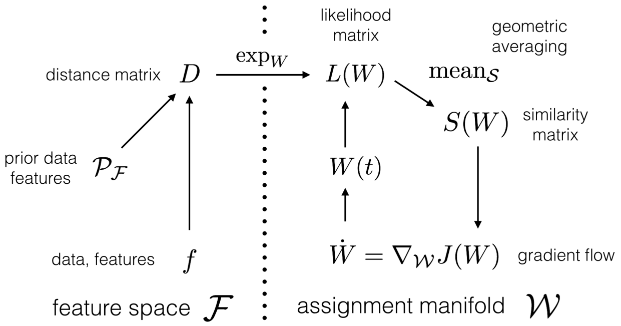

likelihood matrix
geometric averaging
distance matrix
D
$\exp_W$
$L(W)$
$\text{means}_{\mathcal{S}}$
$S(W)$
similarity matrix
prior data features
$\mathcal{P}_{\mathcal{F}}$
$W(t)$
data, features
f
$\dot{W} = \nabla_{\mathcal{W}} J(W)$
gradient flow
feature space $\mathcal{F}$
assignment manifold $\mathcal{W}$

Figure 1. Geometric labeling approach and its components. The feature space $\mathcal{F}$ with a distance function $d_{\mathcal{F}}$, along with observed data and prior data to be assigned, constitute the application specific part. A labeling of the data is determined by a Riemannian gradient flow on the manifold of probabilistic assignments, which terminates at a unique assignment, i.e. a labeling of the given data.

Non-convex relaxations are e.g. based on the mean-field approach [16, Section 5]. They constitute *inner* relaxations of the combinatorially complex feasible set (the so-called marginal polytope) and hence do not require a post-processing step for rounding. However, as for non-convex optimization problems in general, inference suffers from the local-minima problem, and auxiliary parameters introduced for alleviating this difficulty, e.g. by deterministic annealing, can only be heuristically tuned.

Contribution. We introduce a novel approach to the image labeling problem based on a *geometric* formulation. Geometric methods in connection with probabilistic models have been used in computer vision, of course. See, e.g. [14, 4, 5]. Our approach to image labelling is new, however. Figure 1 illustrates the major components of the approach and their interplay. *Labeling* denotes the tasks to assign prior features to given features in any metric space (raw data just constitute a basic specific example). This assignment is determined by solving a Riemannian gradient flow with respect to an appropriate objective function, which evolves on the assignment manifold. The latter key concept encompasses the set of all strictly positive stochastic matrices equipped with a Fisher-Rao product metric. This furnishes a proper geometry for computing local Riemannian means, in order to achieve spatially coherent labelings and to suppress the influence of noise. The Riemannian metric also determines the gradient flow and leads to efficient, sparse interior-point updates that converge in few dozens of outer iterations. Even larger numbers of labels do not significantly slow down the convergence rate. We show that the local Riemannien means can be accurately approximated by closed-form expressions which eliminates inner iterations and hence further speeds up the numerical implementation. For any specified $\varepsilon > 0$, the iterates terminate within a ε-neighborhood of *unique* assignments, which finally determines the labeling.

Our approach is non-convex and *smooth*. Regarding the non-convexity, *no* parameter tuning is needed to escape from poor local minima: For any problem instance, the flow is naturally initialized at the barycenter of the assignment manifold, from which it smoothly evolves and terminates at a labeling.

Organization. We formally detail the components of our approach in Sections 2 and 3. The objective function and the optimization approach are described in Sections 4 and 5. Two academical experiments are reported in Section 6 which illustrate the properties of our approach: (i) the influence of the two user parameters (setting the spatial scale and for scaling the feature range); (ii) spatially consistent image patch assignment.

Our main objective is to introduce and announce a *novel geometric approach* to the image labeling problem of computer vision. Due to lack of space, we omitted all proofs and refer the reader to the report [2] which also provides a more comprehensive discussion of the literature.

Basic Notation. We set $[n] = \{1, 2, \dots, n\}$ and $\mathbb{1} = (1, 1, \dots, 1)^\top$. Vectors $v^1, v^2, \dots$ are indexed by lowercase letters and superscripts, whereas subscripts v_i, $i \in [n]$, index vector components. $\langle u, v\rangle = \sum_{i\in[n]} u_i v_i$ denotes the Euclidean inner product and for matrices $\langle A, B\rangle := \mathrm{tr}(A^\top B)$. For strictly positive vectors we often write pointwise operations more efficiently in vector form. For example, for $0 < p \in \mathbb{R}^n$ and $u \in \mathbb{R}^n$, the expression $\frac{u}{\sqrt{p}}$ denotes the vector $(u_1/\sqrt{p_1}, \dots, u_n/\sqrt{p_n})^\top$.

2. The Assignment Manifold

In this section, we define the feasible set for representing and computating image labelings in terms of assignment matrices $W \in \mathcal{W}$, the assignment manifold $\mathcal{W}$. The basic building block is the open probability simplex $\mathcal{S}$ equipped with the Fisher-Rao metric. We refer to [1] and [6] for background reading.

2.1. Geometry of the Probability Simplex

The relative interior $\mathcal{S} = \mathring{\Delta}_{n-1}$ of the probability simplex $\Delta_{n-1} = \{p \in \mathbb{R}^n_+ \colon \langle \mathbb{1}, p\rangle = 1\}$ becomes a differentiable Riemannian manifold when endowed with the Fisher-Rao metric, which in this particular case reads

$$\langle u, v\rangle_p := \Big\langle \frac{u}{\sqrt{p}}, \frac{v}{\sqrt{p}}\Big\rangle, \; \forall u, v \in T_p\mathcal{S}, \tag{1a}$$

$$T_p\mathcal{S} = \{v \in \mathbb{R}^n \colon \langle \mathbb{1}, v\rangle = 0\}, \; p \in \mathcal{S}. \tag{1b}$$

with tangent spaces denotes by $T_p\mathcal{S}$. The *Riemannian gradient* $\nabla_{\mathcal{S}} f(p) \in T_p\mathcal{S}$ of a smooth function $f \colon \mathcal{S} \to \mathbb{R}$ at $p \in \mathcal{S}$ is the tangent vector given by

$$\nabla_{\mathcal{S}} f(p) = p\big(\nabla f(p) - \langle p, \nabla f(p)\rangle \mathbb{1}\big). \tag{2}$$

We also regard the scaled sphere $\mathcal{N} = 2\mathbb{S}^{n-1}$ as manifold with Riemannian metric induced by the Euclidean inner product of $\mathbb{R}^n$. The following diffeomorphism ψ between $\mathcal{S}$ and the open subset $\psi(\mathcal{S}) \subset \mathcal{N}$, henceforth called *sphere-map*, was suggested e.g. by [9, Section 2.1] and [1, Section 2.5]. $\psi \colon \mathcal{S} \to \mathcal{N}$ and $p \mapsto s = \psi(p) := 2\sqrt{p}$, The sphere-map enables to compute the geometry of $\mathcal{S}$ from the geometry of the 2-sphere. The sphere-map ψ is an isometry, i.e. the Riemannian metric is preserved. Consequently, lenghts of tangent vectors and curves are preserved as well. In particular, geodesics as critical points of length functionals are mapped by ψ to geodesics. We denote by

$$d_{\mathcal{S}}(p, q) \qquad \text{and} \qquad \gamma_v(t), \tag{3}$$

respectively, the *Riemannian distance* on $\mathcal{S}$ between two points $p, q \in \mathcal{S}$, and the *geodesic* on $\mathcal{S}$ emanating from $p = \gamma(0)$ in the direction $v = \dot{\gamma}(0) \in T_p\mathcal{S}$. The *exponential mapping* for $\mathcal{S}$ is denoted by

$$\mathrm{Exp}_p \colon V_p \to \mathcal{S}, \quad v \mapsto \mathrm{Exp}_p(v) = \gamma_v(1), \tag{4}$$

and $V_p = \{v \in T_p\mathcal{S} \colon \gamma_v(t) \in \mathcal{S},\ t \in [0, 1]\}$. The *Riemannian mean* $\mathrm{mean}_{\mathcal{S}}(\mathcal{P})$ of a set of points $\mathcal{P} = \{p^i\}_{i\in[N]} \subset \mathcal{S}$ with corresponding weights $w \in \Delta_{N-1}$ minimizes the objective function

$$\mathrm{mean}_{\mathcal{S}}(\mathcal{P}) = \arg\min_{p\in\mathcal{S}} \frac{1}{2} \sum_{i\in[N]} w_i d^2_{\mathcal{S}}(p, p^i). \tag{5}$$

We use uniform weights $w = \frac{1}{N}\mathbb{1}_N$ in this paper. The following fact is not obvious due to the non-negative curvature of the manifold $\mathcal{S}$. It follows from [8, Thm. 1.2] and the radius of the geodesic ball containing $\psi(\mathcal{S}) \subset \mathcal{N}$.

Lemma 1. *The Riemannian mean* (5) *is unique for any data* $\mathcal{P} = \{p^i\}_{i\in[n]} \subset \mathcal{S}$ *and weights* $w \in \Delta_{n-1}$.

We call the computation of Riemannian means *geometric averaging* (cf. Fig. 1).

2.2. Assignment Matrices and Manifold

A natural question is how to extend the geometry of $\mathcal{S}$ to stochastic matrices $W \in \mathbb{R}^{m\times n}$ with $W_i \in \mathcal{S}$, $i \in [m]$, so as to preserve the information-theoretic properties induced by this metric (that we do not discuss here – cf. [15, 1]).

This problem was recently studied by [12]. The authors suggested three natural definitions of manifolds. It turned out that all of them are slight variations of taking the product of $\mathcal{S}$, differing only by the scaling of the resulting product metric. As a consequence, we make the following

Definition 1 (Assignment Manifold)**.** *The manifold of assignment matrices, called assignment manifold, is the set*

$$\mathcal{W} = \{W \in \mathbb{R}^{m\times n} \colon W_i \in \mathcal{S},\ i \in [m]\}. \tag{6}$$

According to this product structure and based on (1a)*, the Riemannian metric is given by*

$$\langle U, V\rangle_W := \sum_{i\in[m]} \langle U_i, V_i\rangle_{W_i}, \qquad U, V \in T_W\mathcal{W}. \tag{7}$$

Note that $V \in T_W\mathcal{W}$ means $V_i \in T_{W_i}\mathcal{S}$, $i \in [m]$.

Remark 1. *We call stochastic matrices contained in* $\mathcal{W}$ *assignment matrices, due to their role in the variational approach described next.*

3. Features, Distance Function, Assignment

We refer the reader to Figure 1 for an overview of the following definitions. Let

$$f\colon \mathcal{V} \to \mathcal{F}, \qquad i \mapsto f_i, \qquad i \in \mathcal{V} = [m], \tag{8}$$

denote any given data, either raw image data or features extracted from the data in a preprocessing step. In any case, we call f *feature*. At this point, we do not make any assumption about the *feature space* $\mathcal{F}$ except that a *distance function* $d_{\mathcal{F}}\colon \mathcal{F}\times\mathcal{F} \to \mathbb{R}$, is specified. We assume that a finite subset of $\mathcal{F}$

$$\mathcal{P}_{\mathcal{F}} := \{f_j^*\}_{j\in[n]}, \tag{9}$$

additionally is given, called *prior set*. We are interested in the assignment of the prior set to the data in terms of an *assignment matrix* $W \in \mathcal{W} \subset \mathbb{R}^{m\times n}$, with the manifold $\mathcal{W}$ defined by (6). Thus, by definition, every row vector $0 < W_i \in \mathcal{S}$ is a discrete distribution with full support $\operatorname{supp}(W_i) = [n]$. The element

$$W_{ij} = \Pr(f_j^*|f_i), \qquad i \in [m], \quad j \in [n], \tag{10}$$

quantifies the assignment of prior item f_j^* to the observed data point f_i. We may think of this number as the *posterior probability* that f_j^* generated the observation f_i.

The *assignment task* asks for determining an optimal assignment W^*, considered as “explanation” of the data based on the prior data $\mathcal{P}_{\mathcal{F}}$. We discuss next the ingredients of the objective function that will be used to solve assignment tasks (see also Figure 1).

Distance Matrix. Given $\mathcal{F}$, $d_{\mathcal{F}}$ and $\mathcal{P}_{\mathcal{F}}$, we compute the *distance matrix*

$$D \in \mathbb{R}^{m\times n},\ D_i \in \mathbb{R}^n, \quad D_{ij} = \frac{1}{\rho} d_{\mathcal{F}}(f_i, f_j^*), \tag{11}$$

where $i \in [m]$, $j \in [n]$ and $\rho > 0$ is the first (from two) *user parameters* to be set. This parameter serves two purposes. It accounts for the unknown scale of the data f that depends on the application and hence cannot be known beforehand. Furthermore, its value determines what subset of the prior features f_j^*, $j \in [n]$ effectively affects the process of determining the assignment matrix W. We call ρ *selectivity parameter*. Furthermore, we set

$$W = W(0), \qquad W_i(0) := \frac{1}{n}\mathbb{1}_n, \quad i \in [m]. \tag{12}$$

That is, W is initialized with the uninformative *uniform assignment* that is not biased towards a solution in any way.

Likelihood Matrix. The next processing step is based on the following

Definition 2 (Lifting Map (Manifolds $\mathcal{S}, \mathcal{W}$))**.** *The lifting mapping is defined by* $\exp\colon T\mathcal{S} \to \mathcal{S}$*,*

$$(p, u) \mapsto \exp_p(u) = \frac{pe^u}{\langle p, e^u\rangle}, \tag{13}$$

and $\exp\colon T\mathcal{W} \to \mathcal{W}$ *with*

$$(W, U) \mapsto \exp_W(U) = \begin{pmatrix} \exp_{W_1}(U_1) \\ \dots \\ \exp_{W_m}(U_m) \end{pmatrix}, \tag{14}$$

where $U_i, W_i, i \in [m]$ *index the row vectors of the matrices* U, W*, and where the argument decides which of the two mappings* exp *applies.*

Remark 2. *The lifting mapping generalizes the well-known softmax function through the dependency on the base point* p*. In addition, it approximates geodesics and accordingly the exponential mapping* Exp*, as stated next. We therefore use the symbol* exp *as mnemomic. Unlike* Exp_p *in* (4)*, the mapping* $\exp_p$ *is defined on the entire tangent space, which is convenient for numerical computations.*

Proposition 1. *Let*

$$v = \big(\operatorname{Diag}(p) - pp^\top\big)u, \qquad v \in T_p\mathcal{S}. \tag{15}$$

Then $\exp_p(ut)$ *given by* (13) *solves*

$$\dot{p}(t) = p(t)u - \langle p(t), u\rangle p(t), \qquad p(0) = p, \tag{16}$$

and provides a first-order approximation of the geodesic $\gamma_v(t)$ *from* (3), (4)

$$\exp_p(ut) \approx p + vt, \; \|\gamma_v(t) - \exp_p(ut)\| = \mathcal{O}(t^2). \tag{17}$$

Given D and W, we lift the vector field D to the manifold $\mathcal{W}$ by

$$L = L(W) := \exp_W(-U) \in \mathcal{W}, \tag{18}$$

and $U_i = D_i - \frac{1}{n}\langle \mathbb{1}, D_i\rangle \mathbb{1}, \; i \in [m]$, with $\exp_W$ defined by (14). We call L *likelihood matrix* because the row vectors are discrete probability distributions which separately represent the similarity of each observation f_i to the prior data $\mathcal{P}_{\mathcal{F}}$, as measured by the distance $d_{\mathcal{F}}$ in (11). Note that the operation (18) depends on the assignment matrix $W \in \mathcal{W}$.

Similarity Matrix. Based on the likelihood matrix L, we define the *similarity matrix*

$$S = S(W) \in \mathcal{W}, \; S_i = \operatorname{mean}_{\mathcal{S}}\{L_j\}_{j \in \tilde{\mathcal{N}}_{\mathcal{E}}(i)}, \; i \in [m], \tag{19}$$

where each row is the Riemannian mean (5) of the likelihood vectors, indexed by the neighborhoods as specified by the underying graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\tilde{\mathcal{N}}_{\mathcal{E}}(i) = \{i\} \cup \mathcal{N}_{\mathcal{E}}(i)$, and $\mathcal{N}_{\mathcal{E}}(i) = \{j \in \mathcal{V} \colon ij \in \mathcal{E}\}$.

Note that S depends on W because L does so by (18). The *size* of the neighbourhoods $|\tilde{\mathcal{N}}_{\mathcal{E}}(i)|$ is the *second user parameter*, besides the selectivity parameter ρ for scaling the distance matrix (11). Typically, each $\tilde{\mathcal{N}}_{\mathcal{E}}(i)$ indexes the same local "window" around pixel location i. We then call the window size $|\tilde{\mathcal{N}}_{\mathcal{E}}(i)|$ *scale parameter*. In basic applications, the distance matrix D will not change once the features and the feature distance $d_{\mathcal{F}}$ are determined. On the other hand, the likelihood matrix $L(W)$ and the similarity matrix $S(W)$ have to be recomputed as the assignment W evolves, as part of any numerical algorithm used to compute an optimal assignment W^*. We point out, however, that more general scenarios are conceivable – without essentially changing the overall approach – where $D = D(W)$ depends on the assignment as well and hence has to be updated too, as part of the optimization process.

4. Objective Function, Optimization

We specify next the objective function as criterion for assignments and the gradient flow on the assignment manifold, to compute an optimal assignment W^*. Finally, based on W^*, the so-called assignment mapping is defined.

Objective Function. Getting back to the interpretation from Section 3 of the assignment matrix $W \in \mathcal{W}$ as *posterior probabilities*,

$$W_{ij} = \Pr(f_j^* | f_i), \tag{20}$$

of assigning prior feature f_j^* to the observed feature f_i, a natural *objective function* to be maximized is

$$\max_{W \in \mathcal{W}} J(W), \qquad J(W) := \langle S(W), W\rangle. \tag{21}$$

The functional J together with the feasible set $\mathcal{W}$ formalizes the following objectives:

1. Assignments W should *maximally correlate* with the feature-induced similarities $S = S(W)$, as measured by the inner product which defines the objective function $J(W)$.

2. Assignments of prior data to observations should be done in a *spatially coherent* way. This is accomplished by *geometric averaging* of likelihood vectors over local spatial neighborhoods, which turns the likelihood matrix $L(W)$ into the similarity matrix $S(W)$, *depending* on W.

3. Maximizers W^* should define *image labelings* in terms of rows $\overline{W}_i^* = e^{k_i} \in \{0,1\}^n, \; i, k_i \in [m]$, that are indicator vectors. While the latter matrices are not contained in the assignment manifold $\mathcal{W}$, which we notationally indicate by the overbar, we compute in practice assignments $W^* \approx \overline{W}^*$ arbitrarily close to such points. It will turn out below that the *geometry enforces* this approximation.

As a consequence of 3. and in view of (20), such points W^* *maximize posterior probabilities* akin to the interpretation of MAP-inference with discrete graphical models by minimizing corresponding energy functionals. The mathematical structure of the optimization task of our approach, however, and the way of fusing data and prior information, are quite different. The following Lemma states point 3. above more precisely.

Lemma 2. *Let* $\overline{\mathcal{W}}$ *denote the closure of* $\mathcal{W}$. *We have*

$$\sup_{W \in \mathcal{W}} J(W) = m, \tag{22}$$

and the supremum is attained at the extreme points

$$\overline{\mathcal{W}}^* := \big\{\overline{W}^* \in \{0,1\}^{m \times n} \colon \overline{W}_i^* = e^{k_i}, \; i \in [m], \; k_1, \dots, k_m \in [n]\big\} \subset \overline{\mathcal{W}}, \tag{23}$$

corresponding to matrices with unit vectors as row vectors.

Assignment Mapping. Regarding the feature space $\mathcal{F}$, no assumptions were made so far, except for specifying a distance function $d_{\mathcal{F}}$. We have to be more specific about $\mathcal{F}$ only if we wish to *synthesize* the approximation to the given data f, in terms of an assignment W^* that optimizes (21) and the prior data $\mathcal{P}_{\mathcal{F}}$. We denote the corresponding approximation by

$$u \colon \mathcal{W} \to \mathcal{F}^{|\mathcal{V}|}, \; W \mapsto u(W), \; u^* := u(W^*), \tag{24}$$

and call it *assignment mapping*.

A simple example of such a mapping concerns cases where vector-valued prototypical features f^{*j}, $j \in [n]$ are assigned to data vectors f^i, $i \in [m]$: the mapping $u(W^*)$ then simply replaces each data vector by the convex combination of prior vectors assigned to it,

$$u^{*i} = \sum_{j \in [n]} W^*_{ij} f^{*j}, \qquad i \in [m]. \tag{25}$$

And if W^* approximates a global maximum $\overline{W}^*$ as characterized by Lemma 2, then each f_i is uniquely replaced ("labelled") by some $u^{*k_i} = f^{*k_i}$.

Optimization Approach. The optimization task (21) does not admit a closed-form solution. We therefore compute the assignment by the *Riemannian gradient ascent flow* on the manifold $\mathcal{W}$,

$$\begin{aligned} \dot{W}_{ij} &= \big(\nabla_{\mathcal{W}} J(W)\big)_{ij} \\ &= W_{ij}\Big(\big(\nabla_i J(W)\big)_j - \langle W_i, \nabla_i J(W)\rangle\Big), \; j \in [n], \end{aligned} \tag{26}$$

using the initialization (12) with

$$\begin{aligned} \nabla_i J(W) &:= \frac{\partial}{\partial W_i} J(W) \\ &= \Big(\frac{\partial}{\partial W_{i1}} J(W), \dots, \frac{\partial}{\partial W_{in}} J(W)\Big), \quad i \in [m], \end{aligned} \tag{27}$$

which results from applying (2) to the objective (21). The flows (26), for $i \in [m]$, are *not* independent as the product structure of $\mathcal{W}$ (cf. Section 2.2) might suggest. Rather, they are coupled through the gradient $\nabla J(W)$ which reflects the interaction of the distributions W_i, $i \in [m]$, due to the geometric averaging which results in the similarity matrix (19).

5. Algorithm, Implementation

We discuss in this section specific aspects of the implementation of the variational approach.

Assignment Normalization. Because each vector W_i approaches some vertex $\overline{W}^* \in \overline{\mathcal{W}}^*$ by construction, and because the numerical computations are designed to evolve on $\mathcal{W}$, we avoid numerical issues by checking for each $i \in [m]$ every entry W_{ij}, $j \in [n]$, after each iteration of the algorithm (32) below. Whenever an entry drops below $\varepsilon = 10^{-10}$, we rectify W_i by

$$W_i \leftarrow \frac{1}{\langle \mathbb{1}, \tilde{W}_i\rangle} \tilde{W}_i, \; \tilde{W}_i = W_i - \min_{j \in [n]} W_{ij} + \varepsilon, \tag{28}$$

and $\varepsilon = 10^{-10}$. In other words, the number ε plays the role of 0 in our impementation. Our numerical experiments show that this operation removes any numerical issues without affecting convergence in terms of the termination criterion specified at the end of this section.

Computing Riemannian Means. Computation of the similarity matrix $S(W)$ due to Eq. (19) involves the computation of Riemannian means. Although a corresponding fixed-point iteration (that we omit here) converges quickly, carrying out such iterations as a subroutine, at each pixel and iterative step of the outer iteration (32), increases runtime (of non-parallel implementations) noticeably. In view of the approximation of the exponential map $\mathrm{Exp}_p(v) = \gamma_v(1)$ by (17), it is natural to approximate the Riemannian mean as well.

Lemma 3. *Replacing in the optimality condition of the Riemannian mean (5) (see, e.g. [6, Lemma 4.8.4]) the inverse exponential mapping Exp_p^{-1} by the inverse $\exp_p^{-1}$ of the lifting map (13), yields the closed-form expression*

$$\frac{\mathrm{mean}_g(\mathcal{P})}{\langle \mathbb{1}, \mathrm{mean}_g(\mathcal{P})\rangle}, \qquad \mathrm{mean}_g(\mathcal{P}) := \Big(\prod_{i \in [N]} p^i\Big)^{\frac{1}{N}} \tag{29}$$

as approximation of the Riemannian mean $\mathrm{mean}_{\mathcal{S}}(\mathcal{P})$, with the geometric mean $\mathrm{mean}_g(\mathcal{P})$ applied componentwise to the vectors in $\mathcal{P}$.

Optimization Algorithm. A thorough analysis of various discrete schemes for numerically integrating the gradient flow (26), including stability estimates, is beyond the scope of this paper. Here, we merely adopt the following basic strategy from [11], that has been widely applied in the literature (in different contexts) and performed remarkably well in our experiments. Approximating the flow (26) for each vector W_i, $i \in [m]$, and $W_i^{(k)} := W_i(t_i^{(k)})$, by the time-discrete scheme

$$\begin{aligned} &\frac{W_i^{(k+1)} - W_i^{(k)}}{t_i^{(k+1)} - t_i^{(k)}} \\ &\quad = W_i^{(k)}\big(\nabla_i J(W^{(k)}) - \langle W_i^{(k)}, \nabla_i J(W^{(k)})\rangle \mathbb{1}\big), \end{aligned} \tag{30}$$

and choosing the adaptive step-sizes $t_i^{(k+1)} - t_i^{(k)} = \frac{1}{\langle W_i^{(k)}, \nabla_i J(W^{(k)})\rangle}$, yields the multiplicative updates

$$W_i^{(k+1)} = \frac{W_i^{(k)}\big(\nabla_i J(W^{(k)})\big)}{\langle W_i^{(k)}, \nabla_i J(W^{(k)})\rangle}, \qquad i \in [m]. \tag{31}$$

We further simplify this update in view of the explicit expression of the gradient of the objective function with components $\partial_{W_{ij}} J(W) = \langle T^{ij}(W), W\rangle + S_{ij}(W)$, that comprise two terms. The first one in terms of a matrix T^{ij} (that we do not further specify here) contributes the derivative of $S(W)$ with respect to W_i, which is significantly smaller than the second term $S_{ij}(W)$, because $S_i(W)$ results from *averaging* (19) the likelihood vectors $L_j(W_j)$ over spatial neighborhoods and hence changes slowly. As a consequence, we simply drop this first term which.

Thus, for computing the numerical results reported in this paper, we used the fixed-point iteration

$$W_i^{(k+1)} = \frac{W_i^{(k)}\big(S_i(W^{(k)})\big)}{\langle W_i^{(k)}, S_i(W^{(k)})\rangle}, \; W_i^{(0)} = \frac{1}{n}\mathbb{1}, \; i \in [m] \tag{32}$$

together with the approximation due to Lemma 3 for computing Riemannian means, which define by (19) the similarity matrices $S(W^{(k)})$. Note that this requires to recompute the likelihood matrices (18) as well, at each iteration k.

Termination Criterion. Algorithm (32) was terminated if the average entropy

$$-\frac{1}{m}\sum_{i\in[m]}\sum_{j\in[n]} W_{ij}^{(k)} \log W_{ij}^{(k)} \tag{33}$$

dropped below a threshold. For example, a threshold value 10^{-3} means in practice that, up to a tiny fraction of indices $i \subset [m]$ that should not matter for a subsequent further analysis, all vectors W_i are very close to unit vectors, thus indicating an almost unique assignment of prior items f_j^*, $j \in [n]$ to the data f_i, $i \in [m]$. This termination criterion was adopted for all experiments.

Application to Patches: Patch Assignment

Let f^i denote a patch of raw image data (or, more generally, a patch of features vectors) $f^{ij} \in \mathbb{R}^d$, $j \in \mathcal{N}_p(i)$, $i \in [m]$, centered at location $i \in [m]$ and indexed by $\mathcal{N}_p(i) \subset \mathcal{V}$ (subscript p indicates neighborhoods for patches). With each entry $j \in \mathcal{N}_p(i)$, we associate the Gaussian weight

$$w_{ij}^p := G_\sigma(\|x^i - x^j\|), \qquad i, j \in \mathcal{N}_p(i), \tag{34}$$

where the vectors $x^i, x^j \in \mathbb{R}^d$ correspond to the locations in the image domain indexed by $i, j \in \mathcal{V}$.

The prior information is given in terms of n prototypical patches

$$\mathcal{P}_\mathcal{F} = \{f^{*1}, \dots, f^{*n}\}, \tag{35}$$

and a corresponding distance $d_\mathcal{F}(f^i, f^{*j})$, $i \in [m]$, $j \in [n]$. There are many ways to choose this distance depending on the application at hand.

Given an optimal assignment matrix W^*, it remains to specify how prior information is assigned to every location $i \in \mathcal{V}$, resulting in a vector $u^i = u^i(W^*)$ that is the overall result of processing the input image f. Location i is affected by patches that overlap with i. Let us denote the indices of these patches by

$$\mathcal{N}_p^{i\leftarrow j} := \{j \in \mathcal{V} \colon i \in \mathcal{N}_p(j)\}. \tag{36}$$

Every such patch is centered at location j to which prior patches are assigned by

$$\mathbb{E}_{W_j^*}[\mathcal{P}_\mathcal{F}] = \sum_{k\in[n]} W_{jk}^* f^{*k}. \tag{37}$$

Let location i be indexed by i_j in patch j (local coordinate inside patch j). Then, by summing over all patches indexed by $\mathcal{N}_p^{i\leftarrow j}$ whose supports include location i, and by weighting the contributions to location i by the corresponding weights (34), we obtain the vector

$$\begin{aligned} u^i &= u^i(W^*) \\ &= \frac{1}{\sum_{j'\in\mathcal{N}_p^{i\leftarrow j}} w_{j' i_j}^p} \sum_{j\in\mathcal{N}_p^{i\leftarrow j}} w_{j i_j}^p \sum_{k\in[n]} W_{jk}^* f^{*k i_j} \end{aligned} \tag{38}$$

that is assigned by W^* to location i. This expression looks more clumsy than it actually is. In words, the vector u^i assigned to location i is the convex combination of vectors contributed from patches overlapping with i, that itself are formed as convex combinations of prior patches. In particular, if we consider the common case of *equal* patch supports $\mathcal{N}_p(i)$ for every i, that additionally are *symmetric* with respect to the center location i, then $\mathcal{N}_p^{i\leftarrow j} = \mathcal{N}_p(i)$. As a consequence, due to the symmetry of the weights (34), the first sum of (38) sums up all weights w_{ij}^p. Hence, the normalization factor on the right-hand side of (38) equals 1, because the low-pass filter w^p preserves the zero-order moment (mean) of signals. Furthermore, it then makes sense to denote by $(-i)$ the location i_p corresponding to i in patch j. Thus (38) becomes

$$u^i = u^i(W^*) = \sum_{j\in\mathcal{N}_p(i)} w_{j(-i)}^p \sum_{k\in[n]} W_{jk}^* f^{*k(-i)}. \tag{39}$$

Introducing in view of (37) the shorthand

$$\mathbb{E}_{W_j^*}^i[\mathcal{P}_\mathcal{F}] := \sum_{k\in[n]} W_{jk}^* f^{*k(-i)} \tag{40}$$

for the vector assigned to i by the convex combination of prior patches assigned to j, we finally rewrite (38) due the symmetry $w_{j(-i)}^p = w_{ji}^p = w_{ij}^p$ in the more handy form[1]

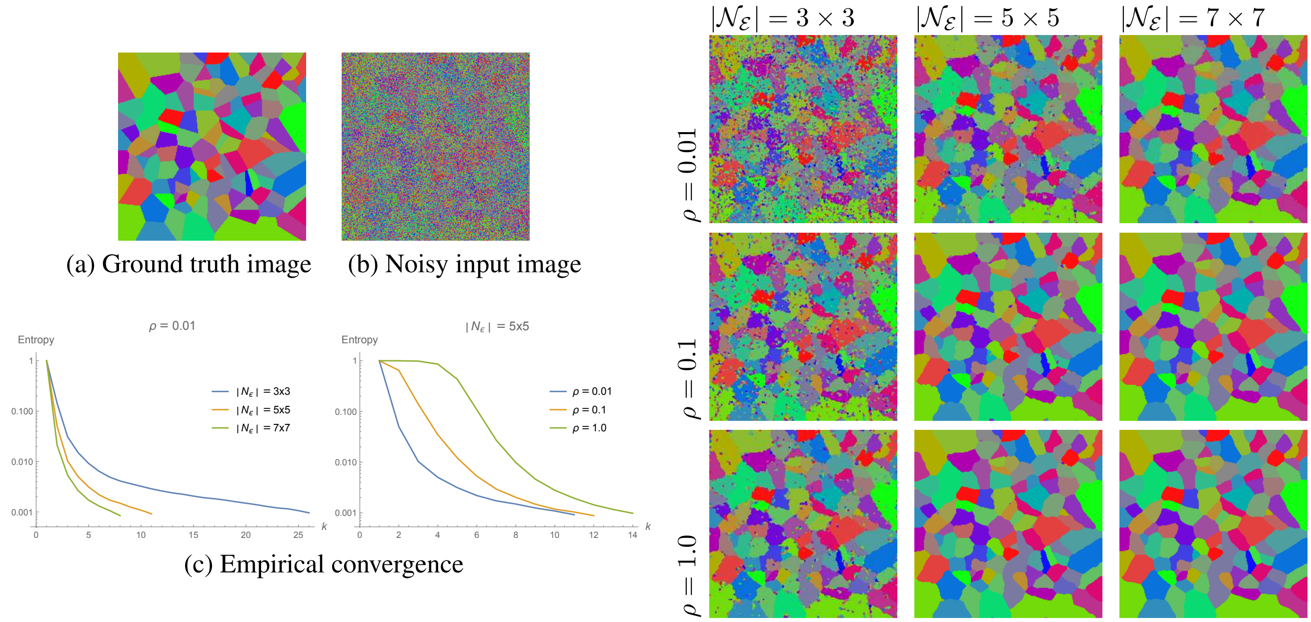
$$u^i = u^i(W^*) = \mathbb{E}_{w^p}\big[\mathbb{E}_{W_j^*}^i[\mathcal{P}_\mathcal{F}]\big]. \tag{41}$$

The inner expression represents the assignment of prior vectors to location i by fitting prior patches to all locations $j \in \mathcal{N}(i)$. The outer expression fuses the assigned vectors. If they were all the same, the outer operation would have no effect, of course.

6. Experiments

In this section, we show results on empirical convergence rate and the influence of the fix-point iteration (32). Additionally, we show results on a patch-based multi-class labeling problem of orientation estimation by labeling.

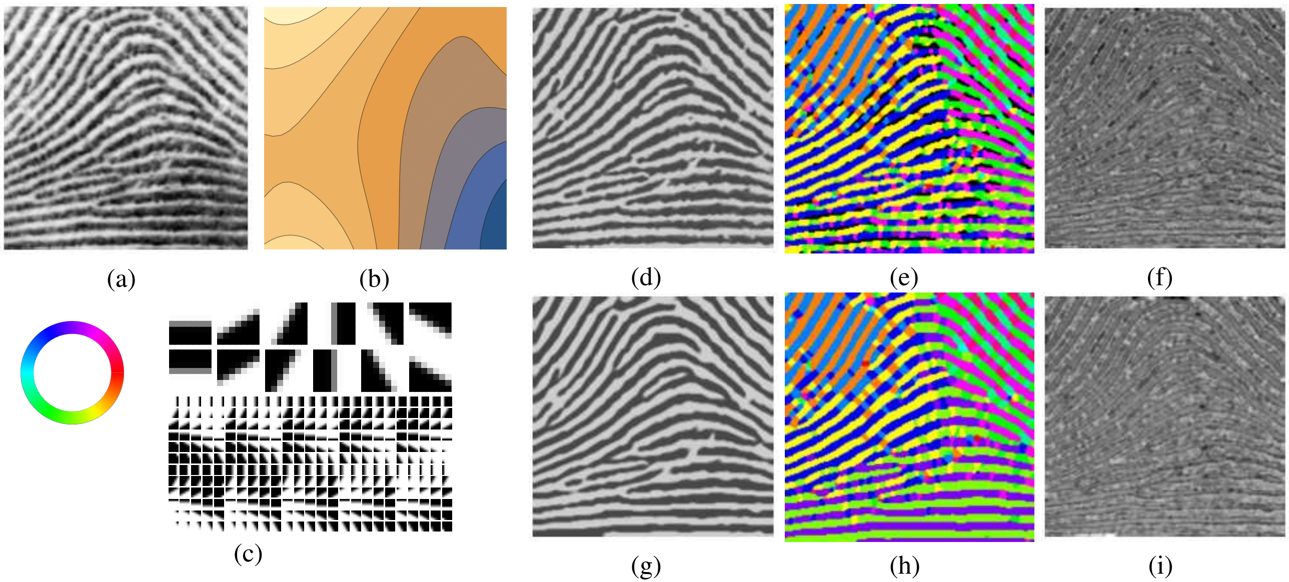
[1] For locations i close to the boundary of the image domain where patch supports $\mathcal{N}_p(i)$ shrink, the definition of the vector w^p has to be adapted accordingly.



(a) Ground truth image (b) Noisy input image

(c) Empirical convergence

Figure 2. **Parameter influence on labeling.** Panels (a) and (b) show a ground-truth image and noisy input data. Right show the assignments $u(W^*)$ for various parameter values where W^* maximizes the objective function (21). The spatial scale $|\mathcal{N}_{\mathcal{E}}|$ increases from left to right. The results illustrate the compromise between sensitivity to noise and to the geometry of signal transitions. Panel (c) shows the average entropy of the assignment vectors $W_i^{(k)}$ as a function of the iteration counter k and the two parameters ρ and $|\mathcal{N}_{\varepsilon}|$,



(a) (b) (d) (e) (f)

(c) (g) (h) (i)

Figure 3. Analysis of the local signal structure of image (a) by patch assignment. This process is twofold non-local: (i) through the assignment of 3×3, (d)-(f) and 7×7 patches (g)-(i), respectively, and (ii) due to the gradient flow (26) that promotes the spatially coherent assignment of patches corresponding to different orientations of signal transitions, in order to maximize the similarity objective (21).

6.1. Parameters, Empirical Convergence Rate

Figure 2 shows a color image and a noisy version of it. All results were computed using the assignment mapping (25) *without* rounding. This shows that the termination criterion of Section 5, illustrated in panel (c) leads to (almost) unique assignments.

The color images comprise of 31 color vectors forming the prior data set $\mathcal{P}_{\mathcal{F}} = \{f^{1*}, \dots, f^{31*}\}$ and are used to illustrate the labeling problem. The labeling task is to assign these vectors in a spatially coherent way to the input data so as to recover the ground truth image. Every color vector was encoded by the vertices of the simplex Δ_{30}, that is by the unit vectors $\{e^1, \dots, e^{31}\} \subset \{0,1\}^{31}$. Choosing the distance $d_{\mathcal{F}}(f^i, f^j) := \|f^i - f^j\|_1$, this results in unit distances between all pairs of data points and hence enables to assess most clearly the impact of geometric spatial averaging and the influence of the two parameters ρ and $|\mathcal{N}_\varepsilon|$, introduced in Section 3.

In Figure 2, the selectivity parameter ρ increases from top to bottom. If ρ is chosen too small, then there is a tendency to noise-induced oversegmentation, in particular at small spatial scales $|\mathcal{N}_\mathcal{E}|$. The reader familiar with total variation based denoising [13], where a *single* parameter is only used to control the influence of regularization, may ask why *two* parameters are used in the present approach and if they are necessary. Note, however, that depending on the application, the ability to separate the physical and the spatial scale in order to recognize outliers with small spatial support, while performing diffusion at a larger spatial scale may be beneficial. We point out that this separation of the physical and spatial scales (image range vs. image domain) is not possible with total variation based regularization where these scales are coupled through the coarea formula. As a consequence, a single parameter is only needed in total variation. On the other hand, larger values of the total variation regularization parameter lead to the well-known loss-of-contrast effect, which in the present approach can be avoided by properly choosing the parameters ρ, $|\mathcal{N}_\varepsilon|$ corresponding to these two scales.

6.2. Orientation Estimation by Patch Assignment

Figure 3 (a) shows a fingerprint image characterized by two grey values $f^*_{\text{dark}}, f^*_{\text{bright}}$, that were extracted from the histogram of f after removing a smooth function of the spatially varying mean value (panel (b)). The latter was computed by interpolating the median values for each patch of a coarse 16×16 partition of the entire image. Panel (c) shows the dictionary of patches modelling the remaining binary signal transitions. The averaging process was set-up to distinguish only the assignment of patches of *different* patch classes and to treat patches of the same class equally. This makes geometric averaging particularly effective if signal structures conform to a single class on larger spatial connected supports. Moreover, it reduces the problem size to merely 13 class labels: 12 orientations at $k \cdot 30°$, $k \in [12]$ degrees, together with the single constant patch complementing the dictionary.

The distance $d_{\mathcal{F}}(f^i, f^{*j})$ between the image patch centered at i and the j-th prior patch was chosen depending on both the prior patch and the data patch it was compared to. For the constant prior patch, the distance was

$$d_{\mathcal{F}}(f^i, f^{*j}) = \frac{1}{|\mathcal{N}_p(i)|} \|f^i - f^*_i f^{*j}\|_1 \tag{42}$$

with

$$f^*_i = \begin{cases} f^*_{\text{dark}} & \text{if } \operatorname{med}\{f^i_j\}_{j \in \mathcal{N}_p(i)} \leq \frac{1}{2}(f^*_{\text{dark}} + f^*_{\text{bright}}), \\ f^*_{\text{bright}} & \text{otherwise.} \end{cases}$$

For all other prior patches, the distance was

$$d_{\mathcal{F}}(f^i, f^{*j}) = \frac{1}{|\mathcal{N}_p(i)|} \|f^i - f^{*j}\|_1. \tag{43}$$

Panels (d)-(f) and (g)-(i) in Figure 3, show the assignment $u(W^*)$ of the dictionary of 3×3 patches and of 7×7 patches. Panels (e) and (h) depict the class labels of these assignments according to the color code of panel (c) and illustrates the interpretation of the image structure of f from panel (a). While the assignment of patches of size 3×3 is slightly noisy, which becomes visible through the assignment of the constant template marked by black in panel (e), the assignment of 5×5 or 7×7 patches results in a robust and spatially coherent, accurate representation of the local image structure. The corresponding pronounced nonlinear filtering effect is due to the consistent assignment of a large number of patches at each pixel location and fusing the corresponding predicted values. Panels (f) and (i) show the resulting additive image decompositions $f = u(W^*) + v(W^*)$ that seem difficult to achieve when using established convex variational approaches (see, e.g., [3]) that employ various regularizing norms and duality, for this purpose. Finally, we point out that it would be straighforward to add to the dictionary further patches modelling minutiae and other features relevant to fingerprint analysis. We do not consider in this paper any application-specific aspects, however.

7. Conclusion

We presented a novel approach to image labeling, formulated in a smooth geometric setting. The approach contrasts with etablished convex and non-convex relaxations of the image labeling problem through smoothness and geometric averaging. The numerics boil down to parallel sparse updates, that maximize the objective along an interior path in the feasible set of assignments and finally return a labeling. Although an elementary first-order approximation of the gradient flow was only used, the convergence rate seems competitive. In particular, a large number of labels does not slow down convergence as is the case of convex relaxations. All aspects specific to an application domain are represented by a single distance matrix D and a single user parameter ρ. This flexibility and the absence of ad-hoc tuning parameters should promote applications of the approach to various image labeling problems.

Acknowledgement. FÅ, SP and CS thank the German Research Foundation (DFG) for support via grant GRK 1653. BS was supported by the European Research Council (project SIGMA-Vision).

References

[1] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Amer. Math. Soc. and Oxford Univ. Press, 2000. 2, 3

[2] F. Åström, S. Petra, B. Schmitzer, and C. Schnörr. Image Labeling by Assignment. March, 16, 2016. preprint: http://arxiv.org/abs/1603.05285. 2

[3] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-Texture Image Decomposition – Modeling, Algorithms, and Parameter Selection. *Int. J. Comp. Vision*, 67(1):111–136, 2006. 8

[4] K. M. Carter, R. Raich, W. G. Finn, and A. O. H. III. Fine: Fisher information nonparametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2093–2098, Nov 2009. 1

[5] M. Harandi, M. Salzmann, and M. Baktashmotlagh. Beyond gauss: Image-set matching on the riemannian manifold of pdfs. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4112–4120, Dec 2015. 1

[6] J. Jost. *Riemannian Geometry and Geometric Analysis*. Springer, 4th edition, 2005. 2, 5

[7] 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. Comp. Vision*, 115(2):155–184, 2015. 1

[8] H. Karcher. Riemannian Center of Mass and Mollifier Smoothing. *Comm. Pure Appl. Math.*, 30:509–541, 1977. 2

[9] R. Kass. The Geometry of Asymptotic Inference. *Statist. Sci.*, 4(3):188–234, 1989. 2

[10] V. Kolmogorov. Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Trans. Patt. Anal. Mach. Intell.*, 28(10):1568–1583, 2006. 1

[11] V. Losert and E. Alin. Dynamics of Games and Genes: Discrete Versus Continuous Time. *J. Math. Biology*, 17(2):241–251, 1983. 5

[12] G. Montúfar, J. Rauh, and N. Ay. On the Fisher Metric of Conditional Probability Polytopes. *Entropy*, 16(6):3207–3233, 2014. 3

[13] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, Nov. 1992. 8

[14] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 1

[15] N. Čencov. *Statistical Decision Rules and Optimal Inference*. Amer. Math.Soc., 1982. 3

[16] M. Wainwright and M. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learning*, 1(1-2):1–305, 2008. 1

[17] T. Werner. A Linear Programming Approach to Max-sum Problem: A Review. *IEEE Trans. Patt. Anal. Mach. Intell.*, 29(7):1165–1179, 2007. 1