MAP Image Labeling Using Wasserstein Messages and Geometric Assignment

Freddie Åström(✉), Ruben Hühnerbein, Fabrizio Savarino, Judit Recknagel, and Christoph Schnörr

Image and Pattern Analysis Group, RTG 1653,
Heidelberg University, Heidelberg, Germany
freddie.astroem@iwr.uni-heidelberg.de

Abstract. Recently, a smooth geometric approach to the image labeling problem was proposed [1] by following the Riemannian gradient flow of a given objective function on the so-called assignment manifold. The approach evaluates user-defined data term and additionally performs Riemannian averaging of the assignment vectors for spatial regularization. In this paper, we consider more elaborate graphical models, given by both data and pairwise regularization terms, and we show how they can be evaluated using the geometric approach. This leads to a novel inference algorithm on the assignment manifold, driven by local Wasserstein flows that are generated by pairwise model parameters. The algorithm is massively edge-parallel and converges to an integral labeling solution.

Keywords: Image labeling · Graphical models · Message passing · Wasserstein distance · Assignment manifold · Riemannian gradient flow · Replicator equation · Multiplicative updates

1 Introduction

Overview. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a grid graph embedded into the image domain $\Omega \subset \mathbb{R}^2$. Vertices $i, j, \ldots \in \mathcal{V}$ index grid positions. A random variable $x_i \in \mathcal{X}$ is assigned to each position i, which takes values in a finite set $\mathcal{X}$ of so-called labels. *Image labeling* commonly denotes the minimization problem

$$\min_{x \in \mathcal{X}^{|\mathcal{V}|}} E(x), \qquad E(x) = \sum_{i \in \mathcal{V}} E_i(x_i) + \sum_{ij \in \mathcal{E}} E_{ij}(x_i, x_j) \tag{1.1}$$

for a given objective function that comprises local functions E_i, E_{ij}, which define a data term and a regularizer, respectively. The data term is typically based on local predictors of the labels, that are trained offline based on observed image features. The latter pairwise terms measure the similarity of labels assigned to adjacent pixel positions and thus enforce spatially smooth label assignments.

The image labeling problem covers a broad range of applications. Accordingly, methods for approximately solving the combinatorial problem (1.1) have attracted a lot of research activities – see [7] for a recent survey.

F. Lauze et al. (Eds.): SSVM 2017, LNCS 10302, pp. 373–385, 2017.
DOI: 10.1007/978-3-319-58771-4_30

The basic *convex relaxation* of (1.1) is based on a reformulation of the objective function in terms of local vectors

$$\theta_i = \big(\theta_i(x_i)\big)_{x_i \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}, \qquad \theta_{ij} = \big(\theta_{ij}(x_i,x_j)\big)_{x_i,x_j \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|^2}, \tag{1.2}$$

whose components are equal to the function values $E_i(x_i), E_{ij}(x_i,x_j)$. Defining in a similar way local indicator vectors $\mu_i \in \{0,1\}^{|\mathcal{X}|}, \mu_{ij} \in \{0,1\}^{|\mathcal{X}|^2}$ yields the linear representation $E_i(x_i) = \langle \theta_i, \mu_i \rangle$ and $E_{ij}(x_i,x_j) = \langle \theta_{ij}, \mu_{ij} \rangle$. Collecting all local terms into vectors θ and μ, respectively, and relaxing the integrality constraint, yields the so-called *local polytope relaxation* [14,16]

$$\min_{\mu} \langle \theta, \mu \rangle \quad \text{subject to} \quad \mu \in \mathcal{P}, \qquad \mathcal{P} = \Big\{ \mu \colon \sum_{x_i \in \mathcal{X}} \mu_{ij}(x_i,x_j) = \mu_j(x_j), \tag{1.3a}$$

$$\sum_{x_j \in \mathcal{X}} \mu_{ij}(x_i,x_j) = \mu_i(x_i), \quad \mu_{ij} \geq 0, \quad \mu_i \in \Delta_c, \quad \forall i \in \mathcal{V},\ \forall ij \in \mathcal{E} \Big\}, \tag{1.3b}$$

where $\Delta_c \subset \mathbb{R}^c, c = |\mathcal{X}|$ denotes the $(c-1)$-dimensional probability simplex. It is well known [17] that this relaxation is only exact for *acyclic* graphs $\mathcal{G}$. For cyclic graphs and image grid graphs, in particular, minimizers $\mu^* \notin \{0,1\}^{\dim(\mu)}$ are not integral, in general. As a consequence, some rounding method is applied to μ^* as postprocessing.

The recent work [1] proposed a *smooth non-convex* approach to the image labeling problem. It is entirely defined in terms of local vectors $W_i \in \operatorname{rint}(\Delta_c)$, $i \in \mathcal{V}$ that live on the relative interior of the simplex which is turned into a simple manifold (see Sect. 2). These vectors are determined by local information, analogous to the data terms $E_i(x)$ of (1.1). In addition, spatial regularization is enforced by computing Riemannian means of the vectors W_i within a local neighborhood around each pixel location i. By construction, the algorithm returns *integral* solutions that make a postprocessing step obsolete. The work [2] studies the multiplicative numerical scheme used in [1], along with a variant, and provides a convergence analysis.

Contribution. The objective of the present paper is to adopt and extend the approach of [1] in order to evaluate established graphical models of the form (1.1), which abound in the literature. This raises the question as to how to take into account the regularizing terms of (1.1). This will be accomplished (i) by regularized Wasserstein distances between adjacent assignment vectors W_i, W_j, $ij \in \mathcal{E}$ (these vectors replace μ_i, μ_j in our approach) that are directly based on the given model parameters θ_{ij}, and (ii) by evolving the corresponding Riemannian gradient on the assignment manifold, as proposed in [1]; see Fig. 1 for an illustration. The resulting approach adds a novel inference algorithm for the image labeling problem to the literature [7]. It may be seen as a sparse interior point algorithm that is exact on acyclic graphs (Lemma 1), and simultaneously performs relaxation and rounding to integrality in a smooth geometric fashion on cyclic graphs. See the Remarks 1, 2 and Sect. 3 for additional detailed comments that classify and position our work.

Related Work. Optimal transport and the Wasserstein distance have become a major tool of image modeling and analysis [8]. Regarding the finite-dimensional formulation in terms of linear programs, we apply the standard device of enhancing convexity through entropic regularization, which increases smoothness in the dual domain. We refer e.g. to [13] and [3, Ch. 9] for basic related work and the connection to matrix scaling algorithms and the history. A smoothed version of the basic Sinkhorn algorithm has become popular in machine learning due to [4], and smoothed Wasserstein distances have been comprehensively investigated in [5, 10] for computing Wasserstein barycenters and interpolation. Our approach to image labeling, in conjunction with the geometric approach of [1], is novel.

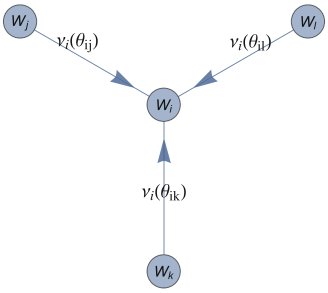

W_j
$\nu_i(\theta_{\mathrm{ij}})$
W_l
$\nu_i(\theta_{\mathrm{il}})$
W_i
$\nu_i(\theta_{\mathrm{ik}})$
W_k

Fig. 1. Illustration of a key aspect of our approach for a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Marginals $W_v \in \mathcal{S}$, $v \in \{i, j, k, l\} \subset \mathcal{V}$ assigned to pixel positions v evolve on the assignment manifold $\mathcal{S}$. Regularization parameters θ_e, $e \in \{ij, ik, il\} \subset \mathcal{E}$ of a given graphical model define Wasserstein distances between pairs of marginals that are incident to edge e. These distances generate *Wasserstein messages* $\nu_v(\theta_e)$ that drive the evolution of W_i, in addition to ordinary local data terms based on observed data.

Organization. Section 2 introduces components of the approach of [1] on which our approach is based on. Our approach is detailed in Sect. 3. Numerical experiments validate and illustrate our approach in Sect. 4.

Basic Notation. $\langle \cdot, \cdot \rangle$ denotes the canonical inner product inducing the Euclidean norm $\|v\| = \langle v, v \rangle^{1/2}$ for vectors or the Frobenius norm $\|A\|_F = \langle A, A \rangle^{1/2}$ in the cases of matrices. $\mathbb{1} = (1, 1, \dots, 1)^\top$ of appropriate dimension. Functions apply componentwise to vectors, e.g. $e^v = (e^{v_1}, \dots, e^{v_n})^\top$. The componentwise multiplication of vectors is denoted as $p \cdot q = (p_1 q_1, \dots, p_n q_n)^\top$. Likewise, we write $\frac{q}{p} := p^{-1} \cdot q$ for the componentwise subdivision by a strictly positive vectors. The set $\mathcal{L}_c = \{e_1, \dots, e_c\}$ collects the c unit vectors as extreme points of the probability simplex $\Delta_c = \{p \in \mathbb{R}^c_+ : \langle \mathbb{1}, p \rangle = 1\}$. The indicator function of a closed convex set C is denoted by $\delta_C(x) = 0$ if $x \in C$, and $\delta_C(x) = +\infty$ otherwise. We set $n = |\mathcal{V}|$ and $[c] = \{1, 2, \dots, c\}$ for $c \in \mathbb{N}$.

2 Image Labeling on the Assignment Manifold

We collect components of the approach [1] that are required to introduce our approach in Sect. 3.

Analogous to the vectors μ_i of (1.3), the basic variables are vectors $W_i \in \mathcal{S} := \operatorname{rint}(\Delta_c)$, $i \in \mathcal{V}$, with $\mathcal{S}$ denoting the relative interior of the probability simplex equipped with the Fisher-Rao metric. A label is assigned to pixel i whenever the vectors W_i are ε-close to some unit vector from the set $\mathcal{L}_c$.

Let f_i, $i \in \mathcal{V}$ denote observed data and f_j^*, $j \in [c]$ given labels. The choice of a distance function $d(\cdot, \cdot)$ defines the distance vectors

$$D_i = \big(d(f_i, f_1^*), \dots, d(f_i, f_c^*)\big)^\top \in \mathbb{R}^c \tag{2.1}$$

and in turn the likelihood vectors

$$L_i(W_i) = \exp_{W_i}(-U_i/\rho) := \frac{W_i \cdot e^{-U_i/\rho}}{\langle W_i, e^{-U_i/\rho}\rangle} \in \mathcal{S}, \qquad U_i = D_i - \frac{1}{c}\langle \mathbb{1}, D_i\rangle \mathbb{1}, \tag{2.2}$$

at every pixel $i \in \mathcal{V}$, where $\rho > 0$ is a user parameter. Next similarity vectors

$$S_i(W) = \operatorname{mean}_{\mathcal{S}}\{L_j\}_{j \in \overline{\mathcal{N}}(i)} \tag{2.3}$$

are computed as approximate Riemannian means of the likelihood vectors over *closed* local neighborhoods $\overline{\mathcal{N}}(i) = \mathcal{N}(i) \cup \{i\}$ containing the center pixel i. The matrix $W \in \mathcal{W} \in \mathbb{R}^{n \times c}$ collects the vectors W_i as row vectors and is an element of the so-called assignment manifold $\mathcal{W} = \mathcal{S} \times \dots \times \mathcal{S}$ ($n = |\mathcal{V}|$ times). Similarly, the vectors $S_i(W)$ are collected as rows of the similarity matrix $S(W) \in \mathcal{W}$.

The counterpart of the convex relaxation (1.3) for determining a labeling is the smooth non-convex problem

$$\sup_{W \in \mathcal{W}} J(W), \; J(W) = \langle S(W), W\rangle, \quad \dot{W}(t) = \nabla_{\mathcal{W}} J(W), \; W(0) = \frac{1}{c}\mathbb{1}\mathbb{1}^\top, \tag{2.4}$$

together with the Riemannian gradient flow on the right-hand side of (2.4): The objective is to determine the assignment matrix W so as to maximize the correlation (inner product) with the similarity matrix, that incorporates the given data and spatial regularization depending on W, too.

Numerical approximations of the gradient flow (2.4) yield assignment vectors W_i, $i \in \mathcal{V}$ that are ε-close to some unit vector from the set $\mathcal{L}_c$, which holds in all experiments. For further details we refer to [1]. A postprocessing step for rounding, as with the convex approach (1.3), is *not* required.

3 Application to Graphical Models

Our approach to the evaluation of a given graphical model of the form (1.1), using the geometric approach of Sect. 2, involves the steps: (1) smooth approximation of the LP relaxation (1.3), (2) adopting the geometry of the assignment manifold, and (3) labeling through numerical optimization. Finally, (4), we reconsider and discuss again Fig. 1.

(1) Smooth Approximation of the LP Relaxation. Setting

$$\mathcal{P}_{\mathcal{V}} = \{\mu_{\mathcal{V}} = (\mu_1, \dots, \mu_n) \colon \mu_i \in \Delta_c,\ i \in \mathcal{V}\}, \tag{3.1a}$$

$$\Pi(\mu_i, \mu_j) = \{\mu_{ij} \geq 0 \colon \mu_{ij} \text{ satifies the marginal. constraints (1.3)}\} \tag{3.1b}$$

we rewrite problem (1.3) in the form

$$\min_{\mu_{\mathcal{V}} \in \mathcal{P}_{\mathcal{V}}} \Big(\sum_{i \in \mathcal{V}} \langle \theta_i, \mu_i \rangle + \sum_{ij \in \mathcal{E}} \min_{\mu_{ij} \in \Pi(\mu_i, \mu_j)} \langle \theta_{ij}, \mu_{ij} \rangle \Big), \tag{3.2}$$

which involves the local Wasserstein distances

$$d_W(\mu_i, \mu_j; \theta_{ij}) = \min_{\mu_{ij} \in \Pi(\mu_i, \mu_j)} \langle \theta_{ij}, \mu_{ij} \rangle, \quad ij \in \mathcal{E}. \tag{3.3}$$

Lemma 1. *Problems (3.2) and (1.3) are equivalent, that is (3.2) is the convex local polytope relaxation of the graphical model (1.1).*

Proof. This follows from the equation

$$\begin{aligned} \min_{\mu \in \mathcal{P}} \langle \theta, \mu \rangle &= \min_{\mu \in \mathcal{P}} \big(\langle \theta_{\mathcal{V}}, \mu_{\mathcal{V}} \rangle + \langle \theta_{\mathcal{E}}, \mu_{\mathcal{E}} \rangle \big) && (3.4a) \\ &= \min_{\mu_{\mathcal{V}}} \big(\langle \theta_{\mathcal{V}}, \mu_{\mathcal{V}} \rangle + \min_{\mu_{\mathcal{E}}} \sum_{ij \in \mathcal{E}} (\langle \theta_{ij}, \mu_{ij} \rangle + \delta_{\Pi(\mu_i, \mu_j)}(\mu_{ij}) \big) && (3.4b) \end{aligned}$$

where $\mu_{\mathcal{E}} = (\dots, \mu_{ij}, \dots)$ and similarly $\theta_{\mathcal{E}}$ collect the local vectors indexed by edges, analogous to $\mu_{\mathcal{V}}$ given by (3.1a) and $\theta_{\mathcal{V}}$ for the vertices. □

Using the entropy function $H(\mu_{ij}) = -\sum_{x_i, x_j} \mu_{ij}(x_i, x_j) \log \mu_{ij}(x_i, x_j)$, where μ satisfies (3.1), our approach is to smooth the convex but non-smooth (piecewise-linear) local functions $d_W(\mu_i, \mu_j; \theta_{ij})$ by entropy regularization,

$$d_{W,\tau}(\mu_i, \mu_j; \theta_{ij}) = \min_{\mu_{ij} \in \Pi(\mu_i, \mu_j)} \{\langle \theta_{ij}, \mu_{ij} \rangle - \tau H(\mu_{ij})\}, \quad ij \in \mathcal{E}, \quad \tau > 0, \tag{3.5}$$

and to minimize the resulting *smooth convex* functional

$$\min_{\mu_{\mathcal{V}} \in \mathcal{P}_{\mathcal{V}}} E_\tau(\mu_{\mathcal{V}}), \qquad E_\tau(\mu_{\mathcal{V}}) = \Big\{ \sum_{i \in \mathcal{V}} \langle \theta_i, \mu_i \rangle + \sum_{ij \in \mathcal{E}} d_{W,\tau}(\mu_i, \mu_j; \theta_{ij}) \Big\} \tag{3.6}$$

by adopting and suitably extending the geometric approach of Sect. 2.

Remark 1. The role of smoothing employed here should *not* be merely considered as a pure *numerical* techniqe (cf. e.g. [9]) for handling the non-smooth convex programs (1.3) and (3.2), because solving the *non-tight* relaxation (1.3) is *not* our focus. Rather, we are interested in approximately solving the original combinatorial labeling problem (1.1), which will be achieved by applying a geometric optimization strategy to (3.6) that converges to *integer-valued* solutions, i.e. labelings. Thus, smoothing in the case of (3.6) is a strategy for taming the combinatorial labeling problem, that is *independently* applied and does *not* conflict with the geometric numerical strategy for computing integral (non-fuzzy) labelings.

(2) Geometric Minimization Approach. In this section, we will consider (i) the computation of the partial gradients $\nabla_{\mu_i} E_\tau(\mu_\mathcal{V})$ and (ii) a natural way for incorporating this information into the geometric approach of Sect. 2.

Regarding (i), we use the following classical result, which is an extension of Danskin's theorem due to Rockafellar.

Theorem 1 [6,11]. *Let $f(z) = \max_{w\in W} g(z,w)$, where W is compact and the function g is differentiable and $\nabla_z g(z,w)$ depending continuously on (z,w). If in addition $g(z,w)$ is convex in z, and if $\overline{z}$ is a point such that $\arg\max_{w\in W} g(\overline{z},w) = \{\overline{w}\}$, then f is differentiable at $\overline{z}$ with*

$$\nabla f(\overline{z}) = \nabla_z g(\overline{z},\overline{w}). \tag{3.7}$$

We apply Theorem 1 to the function $d_{W,\tau}$ of (3.6). To this end, let

$$A\colon \mathbb{R}^{c^2} \to \mathbb{R}^c, \quad A\mu_{ij} = \left(\begin{smallmatrix}\mu_i\\ \mu_j\end{smallmatrix}\right) \tag{3.8}$$

denote the linear mapping so that (3.8) is equal to the marginalizations constraints (3.1b). Furthermore, we introduce the projection

$$\mathit{\Pi}_0\colon \mathbb{R}^c \to T\mathcal{S}, \quad \mathit{\Pi}_0(v) = (I - \frac{1}{c}\mathbb{1}\mathbb{1}^\top), \qquad T\mathcal{S} = \{v\in\mathbb{R}^c\colon \langle \mathbb{1}, v\rangle = 0\} \tag{3.9}$$

onto the tangent space of the manifold $\mathcal{S}$ of Sect. 2, i.e. the subspace of zero-mean vectors (which does not depend on a base point of $\mathcal{S}$).

Corollary 1. *Let μ_i, μ_j be given. Then the gradient $\nabla d_{W,\tau}(\mu_i,\mu_j;\theta_{ij})$ of the function (3.5) is given by the unique solution $(\overline{\nu}_i,\overline{\nu}_j)$ of the equation*

$$\begin{pmatrix}\mu_i\\ \mu_j\end{pmatrix} = A\exp\Big[\frac{1}{\tau}\Big(A^\top\begin{pmatrix}\overline{\nu}_i\\ \overline{\nu}_j\end{pmatrix} - \theta_{ij}\Big)\Big], \qquad \overline{\nu}_i,\overline{\nu}_j \in T\mathcal{S}. \tag{3.10}$$

Proof. The function $d_{W,\tau}$ is defined by the convex problem

$$d_{W,\tau}(\mu_i,\mu_j;\theta_{ij}) = \min_{\mu_{ij}}\big(\langle\theta_{ij},\mu_{ij}\rangle + f_\tau(\mu_{ij})\big) \quad \text{s.t.} \quad A\mu = \left(\begin{smallmatrix}\mu_i\\ \mu_j\end{smallmatrix}\right) \tag{3.11}$$

with $f_\tau(\mu_{ij}) = -\tau H(\mu_{ij}) + \delta_{\mathbb{R}^{c^2}_+}(\mu_{ij})$ and dual problem

$$\max_{\nu_i,\nu_j} g(\mu_i,\mu_j,\nu_i,\nu_j) = \max_{\nu_i,\nu_j}\big[\langle\mu_i,\nu_i\rangle + \langle\mu_j,\nu_j\rangle - f_\tau^*\big(A^\top\left(\begin{smallmatrix}\nu_i\\ \nu_j\end{smallmatrix}\right) - \theta_{ij}\big)\big] \tag{3.12}$$

and convex conjugate function $f_\tau^*(\nu_{ij}) = \tau\sum_{x_i,x_j} e^{\frac{\nu_{ij}(x_i,x_j)-\tau}{\tau}}$. Compactness of the set of maximizers follows from the continuity of f_τ^* (closed level sets) and $\lim_{\tau\to 0} f_\tau^*(\nu_{ij}) = \delta_{\mathbb{R}^{c^2}_-}(\nu_{ij})$, and uniqueness is due to the constraint $\nu_i,\nu_j\in T\mathcal{S}$, which removes the ambiguity $(\nu_i + c\mathbb{1}, \nu_j - c\mathbb{1})$ in the argument of f_τ^* of (3.12), for arbitrary constants $c\in\mathbb{R}$. Since g is linear in μ_i,μ_j, Theorem 1 applies, and Eq. (3.7) becomes (3.10), where we omitted the immaterial scaling factor caused by τ in the numerator of f_τ^*. □

We wish to point out that strong duality between the primal (3.11) and its dual (3.12) holds since μ_i in fact are $W_i \in \mathcal{W}, i \in \mathcal{V}$ which are strictly positive.

It remains to explain how we apply the geometric approach from Sect. 2.

For clarity, we adopt the corresponding notation and replace in the remainder of this paper the local vectors μ_i by W_i, $i \in \mathcal{V}$.

Equation (2.2) defines likelihood vectors $L_i(W_i) = \exp_{W_i}\big(-\Pi_0(D_i)/\rho\big)$ in terms of some vector of distances D_i of the data point f_i to given labels. Since the local energy $\langle\theta_i, \mu_i\rangle$ plays a similar role in the functional (3.6), i.e. measuring a local distance to the labels, it is natural to project the gradient to the tangent space and to define the likelihood vectors

$$L_i(W_i) = \exp_{W_i}\big(-\Pi_0(\theta_i)/\rho\big) = \exp_{W_i}(-\theta_i/\rho), \qquad i \in \mathcal{V}, \tag{3.13}$$

where the projection can be omitted because the mapping $\exp_{W_i}$ defined by (2.2) is invariant against the addition of constant vectors $c\mathbb{1}$, $c \in \mathbb{R}$.

Regarding the pairwise energy terms of (3.6), we proceed in a similar way. For every edge $ij \in \mathcal{E}$, we determine the gradient of the corresponding summand, given by the solution $(\overline{\nu}_i, \overline{\nu}_j)$ to (3.10) (with μ_i, μ_j on the left-hand side replaced by W_i, W_j) and define the likelihood vectors

$$L_{ij;i}(W_i) = \exp_{W_i}(-\overline{\nu}_i/\tau), \qquad L_{ij;j}(W_j) = \exp_{W_j}(-\overline{\nu}_j/\tau), \qquad ij \in \mathcal{E}. \tag{3.14}$$

At this point, we have taken into account the *pairwise* parameters of the graphical model (1.1), and we continue with adapting the final steps of the approach proposed in [1]. Assuming an *arbitrary but fixed orientation* for every edge (i.e. $ij \in \mathcal{E}$ implies $ji \notin \mathcal{E}$), we define for every node i the sets of neighbors of i given by edges *incoming* and *outgoing* to/from i,

$$I(i) = \{j \in \mathcal{V} \colon ji \in \mathcal{E}\}, \qquad O(i) = \{j \in \mathcal{V} \colon ij \in \mathcal{E}\}. \tag{3.15}$$

Then, based on the likelihood vectors associated with each node i,

$$\mathcal{L}_i(W) = \{L_i(W_i)\} \cup \big(\cup_{j\in I(i)} L_{ji;i}(W_j)\big) \cup \big(\cup_{j\in O(i)} L_{ij;i}(W_j)\big), \quad i \in \mathcal{V}, \tag{3.16}$$

we compute analogous to (2.3) the similarity vectors

$$S_i(W) = \mathrm{mean}_{\mathcal{S}}\big(\mathcal{L}_i(W)\big)\Big/\big\langle\mathbb{1}, \mathrm{mean}_{\mathcal{S}}\big(\mathcal{L}_i(W)\big)\big\rangle, \quad i \in \mathcal{V} \tag{3.17}$$

and solve the optimization problem (2.4).

Remark 2. The reader may wonder: Why do we not simply encode the pairwise energy terms $E_{ij}(x_i, x_i)$ by $\langle W_i, \theta_{ij} W_j\rangle$ and generate likelihood vectors by the corresponding partial gradients $\theta_{ij} W_j$ and $\theta_{ij}^{\top} W_i$? The reason is that this would closely correspond to the naive mean field approach to labeling, which is plagued by the local minima problem, as the generally *non-convex* quadratic form $\langle W_i, \theta_{ij} W_j\rangle$ indicates. By contrast, our approach couples the marginals W_i, W_j in terms of the given parameters θ_{ij} through the *convex local* smoothed Wasserstein distance $d_{W,\tau}(W_i, W_j; \theta_{ij})$.

(3) Numerical Optimization. Defining the local model parameter matrices

$$\Theta_{ij} \in \mathbb{R}^{c \times c}, \quad E(\Theta_{ij}) = e^{\frac{-\Theta_{ij}}{\tau}}, \qquad (\Theta_{ij})_{kl} = \theta_{ij}(x_k, x_l), \quad x_k, x_l \in \mathcal{X}, \quad (3.18)$$

where the edge-indexed matrix θ_{ij} is not necessarily symmetric, Eq. (3.10) takes the form $\begin{pmatrix} W_i \\ W_j \end{pmatrix} = \mathrm{Diag}(e^{\frac{\nu_i}{\tau}}) E(\Theta_{ij}) \,\mathrm{Diag}(e^{\frac{\nu_j}{\tau}})$, where $\mathrm{Diag}(\cdot)$ denotes the diagonal matrix with the argument vector as entries. The vectors $\overline{\nu}_i, \overline{\nu}_j$ can be determined by Sinkhorn's algorithm, up to a common multiplicative constant. Setting

$$v_i := e^{\frac{\nu_i}{\tau}}, \qquad v_j := e^{\frac{\nu_j}{\tau}}, \quad (3.19)$$

the corresponding fixed point iterations read

$$v_i^{(k+1)} = \frac{W_i}{E(\Theta_{ij})\Big(\frac{W_j}{E(\Theta_{ij})^\top v_i^{(k)}}\Big)}, \qquad v_j^{(k+1)} = \frac{W_j}{E(\Theta_{ij})^\top\Big(\frac{W_i}{E(\Theta_{ij}) v_j^{(k)}}\Big)}, \quad (3.20)$$

which are iterated until $\|v_i^{(k+1)} - v_i^{(k)}\| \le 10^{-16}$, $\|v_j^{(k+1)} - v_j^{(k)}\| \le 10^{-16}$, which for a reasonable range of $\Theta_{ij} \in [0,1]^{c \times c}$ happens quickly after few iterations. Denoting the iterates after convergence by $v_i^{(\infty)}, v_j^{(\infty)}$, resubstitution into (3.19) and projection onto $T\mathcal{S}$ using (3.9) gives the vectors

$$\overline{\nu}_i = \tau \Pi_0(\log v_i^{(\infty)}), \qquad \overline{\nu}_j = \tau \Pi_0(\log v_j^{(\infty)}). \quad (3.21)$$

which are used to compute the edge likelihood vectors (3.14). These likelihood vectors together with the corresponding vectors (3.13) generated by the data term define (3.16) and in turn the similarity vectors (3.17), which are integrated into the multiplicative scheme of [1] to evolve the marginals by

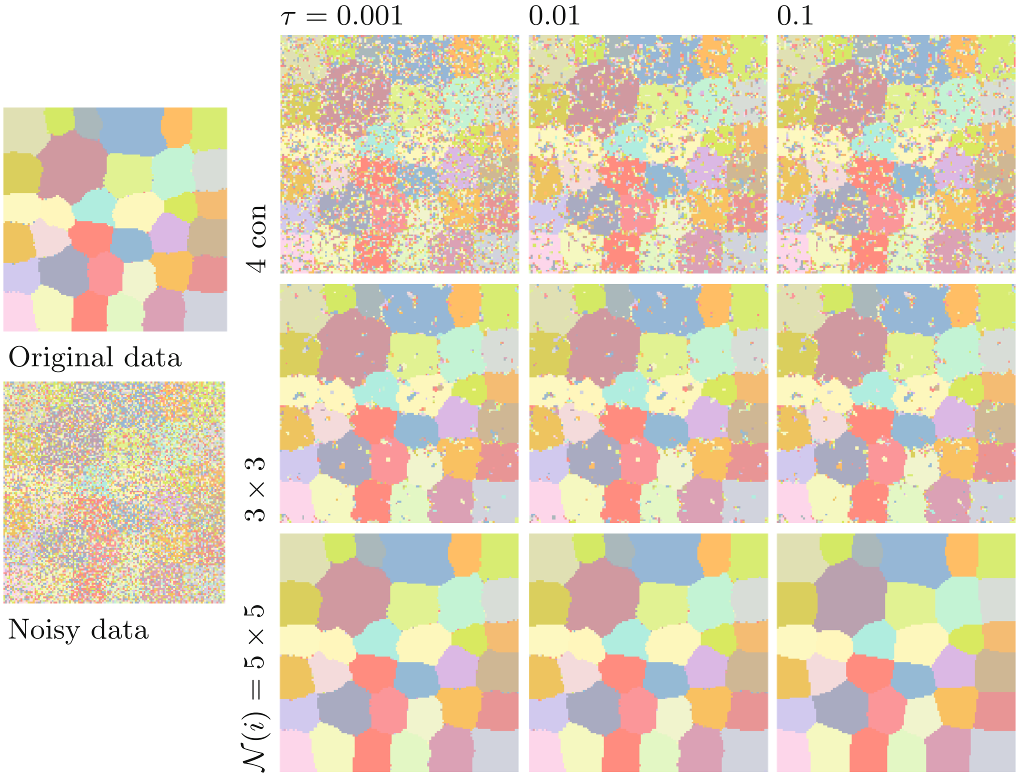
$$W_i^{(k+1)} = \big(W_i^{(k)} \cdot S_i(W^{(k)})\big) / \langle W_i, S_i(W^{(k)}) \rangle, \qquad i \in \mathcal{V}, \quad \forall j, k \in \mathcal{N}(i). \quad (3.22)$$

We adopt the following approximations from [1]: In case an entry of $W_i^{(k+1)}$ drops below $\varepsilon = 10^{-10}$, we set $W_i^{(k+1)} = \varepsilon$ and hence let ε play the role of 0. Furthermore, we approximate the Riemannian mean by the geometric mean, which due to [1, Prop. 3.1] provides a closed form first-order approximation of the geodesics of $\mathcal{S}$ in terms of the mapping $\exp_{W_i}(\cdot)$ defined by (2.2). Finally, we terminate the update scheme (3.22) when the average of the entropies of $W_i^{(k+1)}$ over $i \in \mathcal{V}$ drops below 10^{-3}.

Wasserstein Messages. The rationale behind (3.14) becomes more apparent when rewriting the fixed point Eq. (3.20) *after* convergence in the form

$$v_i^{(\infty)} = W_i / (E(\Theta_{ij}) v_j^{(\infty)}), \qquad v_j^{(\infty)} = W_j / (E(\Theta_{ij})^\top v_i^{(\infty)}). \quad (3.23)$$

This shows that the variables $\overline{\nu}_i, \overline{\nu}_j$ which generate the likelihood vectors (3.14), are *passed along the edges indicent to pixel i* (see Fig. 1). Taking


Original data

Noisy data

Fig. 2. A 4-connected neighborhood denotes the pixels $(x \pm 1, y)$ and $(x, y \pm 1)$ for image coordinates (x, y) at i, and 3×3 and 5×5 mean fully connected neighborhoods with 9 and 25 pixels involved in geometric averaging (cf. (3.17)). For $\rho = 1$, we see that the regularization changes slowly within each scale $\mathcal{N}$ with increasing τ, and increasing the neighborhood sizes increases the spatial regularization.

the log of both sides of the first equation results due to (3.21) in $\overline{\nu}_i = \tau \Pi_0\big(\log W_i - \log(E(\Theta_{ij}) v_j^{(\infty)})\big)$, and in a similar expression for $\overline{\nu}_j$. Comparing this to the general formula for solving Eq. (2.2) for U_i due to [1, App. B.2], $U_i = \rho \Pi_0(\log W_i - \log L_i)$, suggests to identify likelihood vectors that are generated along the edges, as given by (3.14).

Since $\overline{\nu}_i, \overline{\nu}_j$ are the dual variables corresponding to the local marginalization constraints of (1.3) resp. (3.1b), we call these vectors *Wasserstein messages*, in view of the established message passing schemes [17] that aim at solving the dual LP of (1.3) by fixed point iteration. Unlike the latter schemes, our approach satisfies the marginalization constraints *all the time* during the numerical optimization process, rather than only after convergence to a fixed point (provided this happens with common belief propagation on an acyclic graph).

4 Experiments

Parameter Influence, Convergence Rate. In order to better assess the parameter influence we defined 35 unit vectors, each corresponding to one label, encoded on the simplex. This assures that the unary (or distance function,

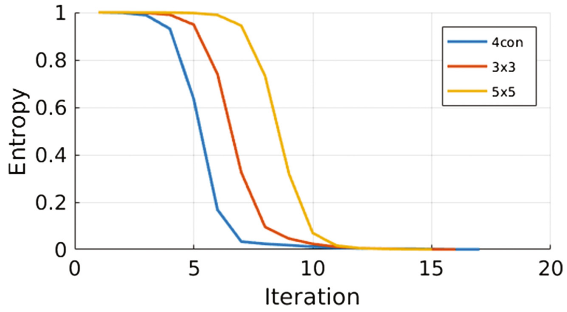


Fig. 3. Entropy as a function of iterations for $\tau = 0.1$ and $\rho = 1$. With an increasing neighborhood size we observe slower convergence which, however, gives a more spatially coherent labeling, as seen in Fig. 2.

cmp. (2.1)) defined as $d(f_i, f_j^*) = \|f_i - f_j^*\|_1$ is not biased towards any single label. Figure 2 shows the influence, for fixed $\rho = 0.2$ for increasing neighborhood size and increasing selectivity τ. In all experiments the termination criteria of 10^{-3} was reached. Figure 3 depicts that the average entropy decrease rapidly for smaller neighborhood sizes. This is reflected in a "noisy" labeling as seen in Fig. 2 since noise is treated as structure due to more conservative information propagation as the regularization is smaller. Increasing the neighborhood size increases the number of iterations until convergence, because the algorithm resolves the 'label competition' through stronger geometric averaging, which results in a smoother labeling. Overall, however, the number of iterations is small.

Inference on Cyclic Graphs. We focus next on the performance of our approach for difficult inference problems on cyclic graphs. To this end, we considered the binary labeling problem on the triangle with three nodes as minimal complete graph, together with many instances of model parameters where the convex LP relaxation (1.3) *fails completely*: It returns the fractional solution $(1/2, 1/2)$ as optimal extreme point of the local polytope for *every* node, which

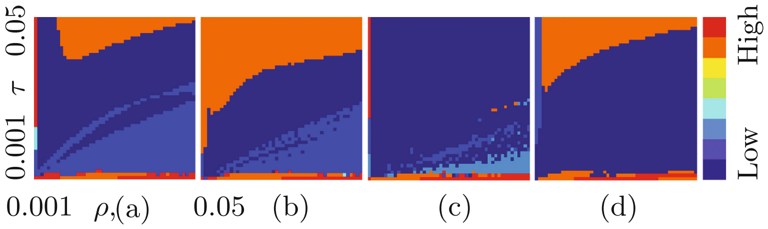


Fig. 4. Energy values for optima computed for the 'frustrated triangle' problem, for various values of τ and ρ. Dark blue color indicates that the globally optimal integral solution was found. The results indicate the existence of an optimal parameter regime for all problem instances. White color indicates that the convergence rate slowed down significantly. The panel labels correspond to the ones in Table 1. (Color figure online)

thus necessitates a postprocessing step to find the optimal binary 0/1 solution, that is equivalent to the original combinatorially hard problem.

Table 1 shows some instances of unary potentials that we used, together with the pairwise potential $\left(\begin{smallmatrix} 0 & -0.1 \\ -0.1 & 0 \end{smallmatrix}\right)$ that favors different labels on adjacent nodes. The latter is impossible on a triangle, which explains the difficulty of this labeling problem. Table 1 shows the energy LP_{frac} of the convex relaxation as lower bound, together with the energy LP_{int} of the optimal binary labeling, determined by exhaustive search.

It is clear that our geometric approach can only find a local binary optimum for this NP-hard problem. Figure 4 shows for 4 problem instances the "energy landscape" resulting after convergence, for varying values of the parameters τ and λ, where the lowest energy corresponding to LP_{int} is encoded with blue. Table 1 displays in the rightmost column the energy within the blue region, which confirms that the optimal *binary* solution was found *without rounding*. The shape of the energy landscape looked roughly the same for all problem instances. A better understanding of how to find parameter values corresponding to the blue region in applications, will be the subject of future work.

Table 1. Unary potentials for the 'frustrated triangle' problem constructed such that the LP relaxation yields a fractional solution $(1/2, 1/2)$ at each vertex. The three right-most columns show the energies LP_{frac} of the LP relaxation, the energy LP_{int} of the globally optimal integral solution, and the energy obtained with our geometric approach within the blue region of parameter values, as displayed by Fig. 4.

	Unary	LP_{frac}	LP_{int}	Geometric
(a)	(0.13, 0.10, 0.86, 0.95, 1.03, 1.06)	1.77	1.79	1.79
(b)	(1.68, 1.67, 0.10, 0.16, 1.21, 1.27)	2.75	2.78	2.78
(c)	(0.99, 0.90, 1.49, 1.46, 0.10, 0.16)	2.25	2.26	2.26
(d)	(1.53, 1.49, 0.10, 0.12, 0.25, 0.19)	1.54	1.58	1.58
(e)	(0.10, 0.10, 1.49, 1.40, 0.86, 0.93)	2.14	2.16	2.16
(f)	(1.08, 0.94, 0.12, 0.10, 1.12, 1.14)	1.95	1.96	1.96

Denoising by Labeling. We competitively compared the performance of our approach for the labeling problem depicted by Fig. 5, using a standard data term together with an Ising prior. To this end, we also evaluated the mean field method [14] and loopy belief propagation [15] (Loopy-BP) based on the UGM package [12], and local rounding was used for these methods as a post-processing step to obtain an integral solution. Figure 5 shows the visual reconstruction as well as the corresponding energy values and percentage of correct labels, which reveals a superior performance of our approach.

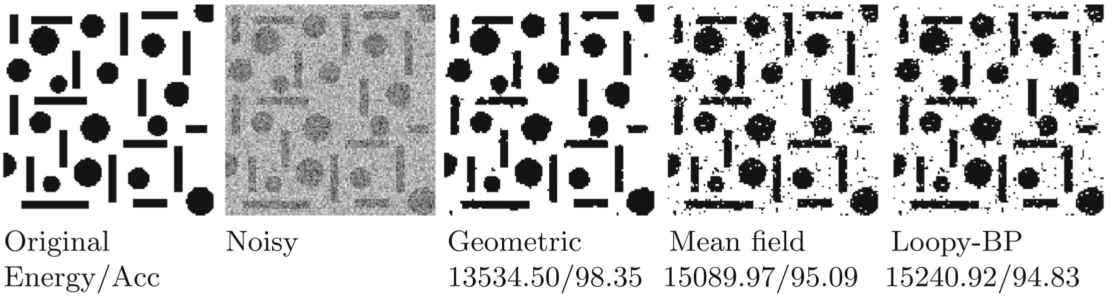


Fig. 5. Noisy binary image recovery. Compared to standard message passing algorithms, our geometric approach shows competitive performance regarding both optimal energy and labeling accuracy. The parameter configuration was $\tau = 0.05$ and $\lambda = 1$, and a 4 connectivity neighborhood was used for geometric spatial regularization.

5 Conclusion

We presented a novel approach which evaluate established graphical models in a smooth geometric setting. Taking into account pairwise potentials, we formulated a novel inference algorithm that propagates "Wasserstein messages" along edges. These messages are lifted to the assignment manifold and drive a Riemannian gradient flow, that terminates at an integral labeling. Our work adds a new inference method to the literature, that simultaneously performs relaxation and rounding to integrality in a smooth geometric fashion.

Acknowledgments. We gratefully acknowledge support by the German Science Foundation, grant GRK 1653.

References

1. Åström, F., Petra, S., Schmitzer, B., Schnörr, C.: Image labeling by assignment. J. Math. Imaging Vis. **58**(2), 211–238 (2017)
2. Bergmann, R., Fitschen, J.H., Persch, J., Steidl, G.: Iterative multiplicative filters for data labeling. Int. J. Comput. Vis. 1–19 (2017). http://dx.doi.org/10.1007/s11263-017-0995-9
3. Brualdi, R.: Combinatorial Matrix Classes. Cambridge University Press, Cambridge (2006)
4. Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In: Proceedings of the NIPS (2013)
5. Cuturi, M., Peyré, G.: A smoothed dual approach for variational wasserstein problems. SIAM J. Imag. Sci. **9**(1), 320–343 (2016)
6. Danskin, J.: The theory of max min with applications. SIAM J. Appl. Math. **14**, 641–664 (1966)
7. Kappes, J., Andres, B., Hamprecht, F., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B., Rother, C.: A comparative study of modern inference techniques for structured discrete energy minimization problems. Int. J. Comput. Vis. **115**(2), 155–184 (2015)

8. Kolouri, S., Park, S., Thorpe, M., Slepcev, D., Rohde, G.: Transport-based analysis, modeling, and learning from signal and data distributions (2016). preprint: https://arxiv.org/abs/1609.04767
9. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. Ser. A **103**, 127–152 (2005)
10. Peyré, G.: Entropic approximation of wasserstein gradient flows. SIAM J. Imag. Sci. **8**(4), 2323–2351 (2015)
11. Rockafellar, R.: On a special class of functions. J. Opt. Theor. Appl. **70**(3), 619–621 (1991)
12. Schmidt, M.: UGM: Matlab code for undirected graphical models, January 2017
13. Schneider, M.: Matrix scaling, entropy minimization, and conjugate duality (II): the dual problem. Math. Program. **48**, 103–124 (1990)
14. Wainwright, M., Jordan, M.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1**(1–2), 1–305 (2008)
15. Weiss, Y.: Comparing the mean field method and belief propagation for approximate inference in MRFs. In: Advanced Mean Field Methods: Theory and Practice, pp. 229–240. MIT Press (2001)
16. Werner, T.: A linear programming approach to max-sum problem: a review. IEEE Trans. Patt. Anal. Mach. Intell. **29**(7), 1165–1179 (2007)
17. Yedidia, J., Freeman, W., Weiss, Y.: Constructing free-energy approximations and generalized belief propagation algorithms. Trans. I. Theor. **51**(7), 2282–2312 (2005)