Wasserstein Active Contours

Gabriel Peyré, Jalal M. Fadili, Julien Rabin

Wasserstein Active Contours

Gabriel Peyré[1], Jalal Fadili[2] and Julien Rabin[3]

[1] CNRS and Ceremade, Université Paris-Dauphine, France
[2] CMLA, CNRS-ENS Cachan-UniverSud, France
[3] GREYC, CNRS-ENSICAEN-Université de Caen, France
gabriel.peyre@ceremade.dauphine.fr,
jalal.fadili@greyc.ensicaen.frjulien.rabin@cmla.ens-cachan.fr

Abstract. In this paper, we propose a novel and rigorous framework for region-based active contours that combines the Wasserstein distance between statistical distributions in arbitrary dimension and shape derivative tools. To the best of our knowledge, this is the first variational image segmentation algorithm that involves region-dependent multi-dimensional descriptors based on the optimal transport theory. The distributions are represented owing to non-parametric kernel density estimators (e.g. Parzen), and the exact evolution speed corresponding to the Wasserstein-based segmentation energy is provided. To speed-up the computation and be able to handle high-dimensional features and large-scale data, we introduce a sliced Wasserstein approximation of the original Wasserstein distance. The framework is flexible enough to allow either minimization of the Wasserstein distance to known fixed distributions, or maximization of the distance between the distributions of the regions to be segmented (region competition). Numerical results reported to show the advantages of the proposed optimal transport distance with respect to alternative metrics (such as the Kullback-Leibler divergence). These traditional metrics cannot deal properly with distributions having localized supports, and do not take into account the distance between the modes of the histograms. Additionally, our framework handles distributions in arbitrary dimension, which is crucial to segment color images.

1 Introduction

1.1 Overview of the Literature

Contour-based vs. region-based segmentation methods. Active contours for image segmentation methods can be broadly classified as being either edge-based or region-based. Starting from the seminal work on the snakes model [12], edge-based actives contours are driven towards image edges through the minimization of a boundary integral of functions of features depending on edges. They can be viewed as the computation of a locally [2, 3] or globally [5] minimal path for a Riemannian metric usually derived from the image gradient magnitude.

Region-based active contours (RBAC) were proposed later. In these approaches, region-based terms can be advantageously combined with boundary-based ones. The evolution equation is generally deduced from a general criterion that includes both region integrals and boundary integrals; see e.g. [16, 21, 4, 18, 25]. The main issue

when dealing with RBAC models is the computation of the velocity vector in the evolution equation from the energy functional, especially when the descriptors are region-dependent, as will be the case in this work. This is mostly due to the fact that the set of image regions does not have the structure of a vector space, preventing us from using in a straightforward way gradient descent methods. To circumvent this difficulty, we here take benefit from the shape derivation principles, see [6, 1, 10].

Statistical segmentation A number of authors have proposed RBAC energy functionals involving statistical region-based terms. These are typically functions of the distribution of some image attributes within the region. The distribution can be either parametric or non-parametric, see for instance [15, 25, 9, 23, 8]. In the non-parametric approach, the energy functional usually involves a distance (or divergence), e.g. the Kullback-Leibler divergence, between non-parametric kernel estimates (e.g. Parzen kernel) of the underlying densities. To be able to handle localized distributions, this requires a proper choice of the smoothing kernel bandwidth, see for instance [13].

Optimal transport for imaging. To avoid the drawbacks faced with traditional statistical distances, the authors in [17] propose to use the Wasserstein distance. In their formulation, the distribution estimated from patches, but does not take into account the explicit dependence of the Wasserstein distance to the region.

The Wasserstein distance originates from the theory of optimal transport [24]. It defines a natural metric between probability distributions, that takes into account the relative distance between the different modes. In the case of unimodal distributions, this distance reduces to the distance between the modes, thus enabling a precise discrimination of the features.

This Wasserstein distance have found a wide range of applications such as the comparison of histogram features for image retrieval [22, 19], histogram equalization [7] and color transfer [20].

1.2 Relation to Previous Work

To the best of our knowledge, the work of [17] is the first, and so far the only one, to clearly address the statistical segmentation problem using a Wasserstein distance. Their work clearly emphasizes the usefulness of optimal transport methods to deal with statistically localized features. Our work, however, departs significantly from theirs in many important ways. First, unlike their work which focused on scalar features with the L^1 Wasserstein distance, we consider a general setting in arbitrary dimension. Secondly, the distance in our energy functional is explicitly region-dependent. Using shape derivative tools, we also provide the exact derivative of the Wasserstein distance with respect to the domain boundary and then deduce the active contour velocity field. In contrast, the work [17] use a patch-based local histogram estimation to avoid taking into account the dependence of the statistics to the region. While this solution appears appealing by its simplicity, and provides good results on synthetic and natural images, it faces an important dilemma when choosing the patch size. On the one hand, using large patches is important to get consistent estimates in the setting where the distributions have large variances, which is typically the case for textures or noisy images. But this comes at the price of loss of accuracy near the boundaries as the patches overlap between the inside and the outside regions.

1.3 Contributions

Our contributions are threefold. (i) We propose a novel rigorous framework for region-based active contours that combines the Wasserstein distance in arbitrary dimension and shape derivative tools. (ii) We state theoretical results regarding variational minimization of the Wasserstein distance with respect to a domain. Such results might be of independent interest. (iii) We propose an approximate transport distance to speed up the computations for large-scale images carrying multi-dimensional features.

1.4 Notations

In the sequel, we consider a feature map $I : u \in \Sigma \to \mathbb{R}^d$, where $u \in \Sigma$ indexes the pixel location, and d is the dimension of the feature (for instance $d = 3$ for a color image). We consider a histogram binning grid $\Omega \subset \mathbb{R}^d$.

In the following, we consider Σ as a continuous domain (equipped with the Lebesgue measure) and Ω as a finite discrete grid (equipped with the counting measure). We thus define the Hilbert spaces $\mathrm{L}^2(\Omega)$ and $\mathrm{L}^2(\Sigma)$ endowed with the inner products

$$\langle A,\, B\rangle_\Omega = \sum_{x\in\Omega} A(x)B(x) \quad \text{and} \quad \langle f,\, g\rangle_\Sigma = \int_\Sigma f(u)g(u)\mathrm{d}u.$$

The set of statistical distributions on Ω is

$$\mathcal{D}(\Omega) = \left\{ A \setminus A(x) \geqslant 0 \quad \text{and} \quad \sum_{x\in\Omega} A(x) = \langle A,\, \mathbf{1}\rangle_\Omega = 1 \right\} \subset \mathrm{L}^2(\Omega).$$

where $\mathbf{1}(x) = 1$, $\forall x \in \Omega$.

2 Non-parametric Statistical Segmentation

Before detailing our Wasserstein distance segmentation approach, we first introduce in this section the tools needed for the estimation of the region distribution at points taking values in the finite discrete grid Ω, together with the required tools from the theory of shape gradients. Although this is standard material, we detail the exact derivatives since this is crucial for our numerical scheme.

2.1 Kernel Density Estimator

Given a fixed feature map $I : \Sigma \to \mathbb{R}^d$, and a non-negative weight function $w \in \mathrm{L}^2(\Sigma)$, the kernel density estimator of the distribution underlying I is given by

$$\forall\, x \in \Omega, \quad P(w) = \frac{\Psi(w)}{\langle C,\, w\rangle_\Sigma} \in \mathcal{D}(\Omega)\ , \tag{1}$$

where the mapping $\Psi : \mathrm{L}^2(\Sigma) \to \mathrm{L}^2(\Omega)$ and its adjoint are defined as

$$\Psi(w) : x \mapsto \int_\Sigma w(u)\psi_s(x - I(u))\mathrm{d}u, \quad \Psi^*(f) : u \mapsto \sum_{x\in\Omega} f(x)\psi_s(x - I(u)),$$

$$\text{and} \quad C = \Psi^*(\mathbf{1}) \in \mathrm{L}^2(\Sigma) \quad \text{i.e.} \quad C(u) = \sum_{x \in \Omega} \psi_s(x - I(u)).$$

Formally, P is a mapping $P : \mathrm{L}^2(\Sigma) \to \mathcal{D}(\Omega)$.

ψ_s is a non-negative symmetric smooth localized window called the kernel, and s is its bandwidth. A common kernel function is a Gaussian kernel, and the corresponding estimator is termed (with an abuse of terminology), the Parzen estimator. There are a plenty of other choices such as the Epanechnikov kernel. The choice of s is even more crucial and results from a traditional bias-variance tradeoff, and should be adapted to the discretization grid Ω and smoothness of the underlying density.

As we will show in the next section, in order to optimize energies with respect to a domain, by the chain rule, we will end up needing to to derive Ψ and its adjoint w.r.t. w. This is precisely the purpose of the following proposition.

Proposition 1 *We have*

$$DP(w) : \delta \in \mathrm{L}^2(\Sigma) \mapsto \frac{1}{\langle C,\, w \rangle_\Sigma} \big(\Psi(\delta) - P(w) \langle C,\, \delta \rangle_\Sigma \big) \in \mathrm{L}^2(\Omega) \ , \tag{2}$$

$$DP(w)^* : \mu \in \mathrm{L}^2(\Omega) \mapsto \frac{1}{\langle C,\, w \rangle_\Sigma} \big(\Psi^*(\mu) - C \langle P(w),\, \mu \rangle_\Omega \big) \in \mathrm{L}^2(\Sigma). \tag{3}$$

2.2 Statistical Distance-based Segmentation

Let's consider the problem of variational segmentation of the image domain in two regions $\Sigma = \Gamma \cup \Gamma^c$, where Γ is a regular bounded open set. Γ and its complement Γ^c share the same boundary $\partial\Gamma$ (denoted also C for short), with normals pointing in opposite directions. The goal is to find a (local) minimizer of an energy including both region (Wasserstein fidelity) and boundary (regularity) functionals. The key principle is to construct a PDE from the energy criterion that changes the shape of the current boundary curve according to some velocity field which can be thought of as a direction of descent of the energy criterion.

Shape derivatives of statistical distances. Let's define the region functional

$$E(\Gamma, B) = W(P(\chi_\Gamma), B) \tag{4}$$

for any fixed $B \in \mathcal{D}(\Omega)$, where $\chi_\Gamma(u)$ is the characteristic function of Γ, i.e. $\chi_\Gamma(u) = 1$ if $u \in \Gamma$, and 0 otherwise.

If we introduce an artificial time τ for the evolution, and consider $m \in [0,1] \mapsto \mathsf{C}(m,\tau)$ to be a parametric representation of the boundary $\partial\Gamma$ at time τ, a gradient flow of this boundary reads

$$\frac{\partial\, \mathsf{C}(m,\tau)}{\partial \tau} = \mathbf{v}_\Gamma(\mathsf{C}(m,\tau)) \quad \text{and} \quad \mathsf{C}(\cdot, 0) = \mathsf{C}_0\,. \tag{5}$$

Here $\mathbf{v}_\Gamma$ is the so-called shape gradient. It ensures that $\mathsf{C}(m,\tau)$ converges when τ increases to a stationary point (hopefully a local minimium) of $E(\Gamma, B)$. The following proposition gives the expression for the shape gradient.

Proposition 2 *The shape gradient* $\mathbf{v}_\Gamma$ *ensuring that* (5) *converges to a stationary point of* $E(\Gamma, B)$ *is*

$$\mathbf{v}_\Gamma(u) = G_{\Gamma,B}(u)\,\mathbf{N}_u \quad \textit{where} \quad G_{\Gamma,B} = [DP(\chi_\Gamma)^*]\left(\nabla_1 W(P(\chi_\Gamma), B)\right), \tag{6}$$

where $\mathbf{N}_u$ *is the unit inward normal to* $\partial\Gamma$ *at* u*, and* $\nabla_1 W$ *is the (sub)gradient of* W *with respect to its first variable, and* DP^* *is given in* (3)*.*

Proof. See for instance[1, Theorem 6.1][11, Theorem 2].

Level set implementation. For the numerical implementation of the minimization of the energy with respect to the domain Γ, we use the level set method applied to active contours. The key idea is to introduce an auxiliary $\varphi : \Sigma \to \mathbb{R}$, which is often chosen to be the signed distance to $\partial\Gamma$. Thus Γ is represented as

$$\Gamma = \{u \in \Sigma \setminus \varphi(u) > 0\} \quad \text{and} \quad \partial\Gamma = \{u \in \Sigma \setminus \varphi(u) = 0\}.$$

The energy (4) is rewritten as $W(P(H(\varphi)), B)$ where $H = \chi_{[0,+\infty)}$ is the Heaviside function. Introducing an artificial time τ, the evolution equation (5) associated to the energy (4) then becomes

$$\frac{\partial\varphi(u,\tau)}{\partial\tau} = |\nabla\varphi(u,\tau)| G_{\Gamma,B}(u) \quad \text{where} \quad \Gamma = \{u \setminus \varphi(u,\tau) > 0\}. \tag{7}$$

Note that the velocity function $G_{\Gamma,B}$ is computed only on the curve $\mathsf{C}(\cdot,\tau)$, but we can extend its expression to the whole image domain Σ. However, the signed distance function φ is not a solution of the PDE (7), and in practice, it must be periodically re-initialized so that it remains a distance function. This is important to ensure numerical stability of the method.

2.3 Statistical Segmentation by Region Competition

In the same vein as [11, 9], we restrict our attention in this paper to a non-parametric variational segmentation method that seeks the maximization of the distance between the respective distributions in Γ and Γ^c, i.e. region competition. Of course, our approach can be applied to other energy functionals just as well, e.g. those with terms that favor region homogeneity. The energy functional to be minimized reads

$$\min_\Gamma \mathcal{E}(\Gamma) = -W(P(\chi_\Gamma), P(\chi_{\Gamma^c})) + \lambda r(\mathsf{C}). \tag{8}$$

where $r(\mathsf{C})$ is a boundary regularity term, e.g. the curve length. Written using the level set formalism, this corresponds to the solution of

$$\min_\varphi \; -W(P(H(\varphi)), P(H(-\varphi))) + \lambda R(\varphi). \tag{9}$$

where $R(\varphi)$ is a suitable regularization associated to $r(\mathsf{C})$. For instance, if $r(\mathsf{C})$ is the length, then $R(\varphi)$ is the TV regularization $R(\varphi) = \int |\nabla\varphi(u)| \mathrm{d}u$.

The equivalent level set evolution PDE (7) that drives an initial contour to a stationary point (hopefully a local minimizer) of (8) is

$$\frac{\partial \varphi(u,\tau)}{\partial \tau} = |\nabla \varphi(u,\tau)| \left(G_{\Gamma^c, B_\Gamma}(u) - G_{\Gamma, B_{\Gamma^c}}(u) + \lambda \kappa \right) \tag{10}$$

where $\Gamma = \{u \setminus \varphi(u,\tau) > 0\}$ is the domain at time τ (we have dropped the dependency on τ for the sake of clarity), where the histogram inside and outside Γ are

$$B_\Gamma = P(H(\varphi(\cdot,\tau)) \quad \text{and} \quad B_{\Gamma^c} = P(H(-\varphi(\cdot,\tau)),$$

where G_{Γ,B_Γ} is the velocity defined in (6), and $\kappa = \operatorname{div}\left(\frac{\nabla \varphi}{|\nabla \varphi|}\right)$ is the mean curvature of the boundary. In practice, this PDE is discretized with a sufficiently small time step (for instance computed using a line search).

3 Wasserstein Distance

Previous work, including [11, 9, 23], have specialized the above statistical framework to the case of point-wise statistical metrics. Among the most popular is the Kullback-Leibler distance (i.e. symmetrized divergence)

$$W(A,B) = \tilde{W}(A,B) + \tilde{W}(B,A) \quad \text{where} \quad \tilde{W}(A,B) = \sum_{x \in \Omega} A(x) \log\left(\frac{A(x)}{B(x)}\right). \tag{11}$$

While this distance enjoys many appealing statistical/information theoretic properties, it suffers from some drawbacks. The most prominent one is that it assumes that both distributions share the same support. In particular $W(A,B) = +\infty$ if $A(x) = 0$ and $B(x) \neq 0$ for some $x \in \Omega$, and vice versa. Furthermore, such a metric does not care about the relative positions of highly localized modes in the distribution. Put differently, this means that $W(A,B) = W(\sigma(A), \sigma(B))$ where $\sigma(A)(x) = A(\sigma(x))$ for any permutation σ of the grid points. Both issues make such a metric not very robust when comparing localized distributions. A practical but artificial way to somehow alleviate these difficulties is to use a large smoothing bandwidth s in (1), but this is likely to yield oversmooth distribution estimates.

Our proposal is to use the Wasserstein distance which is not prone to the above drawbacks, since it takes into account the relative distances between the grid points.

3.1 Discrete Optimal Transport

We consider two distributions $A, B \in \mathcal{D}(\Omega)$. An optimal transport cost on $\mathcal{D}(\Omega)$ is defined as

$$W(A,B) = \min_{P \in \mathcal{P}(A,B)} \langle C,\, P \rangle_{\Omega \times \Omega} = \sum_{(x,y) \in \Omega^2} C(x,y) P(x,y) \tag{12}$$

where C is a fixed cost matrix and $\mathcal{P}(A,B)$ is the polytope of stochastic matrices with marginals A and B:

$$\mathcal{P}(A,B) = \left\{ P \setminus P(x,y) \geqslant 0, \sum_{y \in \Omega} P(x,y) = A(x), \sum_{x \in \Omega} P(x,y) = B(y) \right\}.$$

When the cost is $C(x,y) = \|x-y\|^p$, $W(A,B)^{1/p}$ is often called the L^p Wasserstein distance, and is a distance on $\mathcal{D}(\Omega)$. For the sake of simplicity, we will refer to W as being the Wasserstein distance, and will not use the exponent $1/p$.

3.2 Wasserstein Subdifferential

The Wasserstein distance can be re-written using the dual problem to (12)

$$W(A,B) = \max_{(u,v)\in\mathcal{Q}} \left\{ \langle u,\, A\rangle_\Omega + \langle v,\, B\rangle_\Omega = \sum_{x\in\Omega} u(x)A(x) + v(x)B(x) \right\} \tag{13}$$

$$\text{where} \quad \mathcal{Q} = \left\{ (u,v) \setminus \forall\, (x,y) \in \Omega^2,\ u(x) + v(y) \leqslant C(x,y) \right\},$$

see [24]. Note that $\mathcal{Q}$ is independent of A and B. Rephrased in the language of convex analysis, the Wasserstein distance $W(A,B)$ is nothing but the support function of the closed convex set $\mathcal{Q}$. Thus, using properties of Legendre-Fenchel conjugacy, closedness, convexity and subdiffentiability properties of $W(A,B)$ can be established.

Proposition 3 *The function $A \in \mathcal{P}(\Omega) \mapsto W(A,B) \in \mathbb{R}$ is closed and convex for any fixed $B \in \mathcal{P}$. Its subdifferential at some $A \in \mathcal{P}(\Omega)$ is such that*

$$u \in \partial_1 W(A,B) \iff \exists v \text{ s.t. } (u,v) \in \operatorname*{Argmax}_{(q,p)\in\mathcal{Q}} \langle q,\, A\rangle_\Omega + \langle p,\, B\rangle_\Omega \ .$$

Proof. Closeness and convexity follow from [14, Theorem X.1.1.2]. The subdifferential characterization is a consequence of Fenchel identity [14, Theorem X.1.4.1].

In plain words, a subgradient of W at A, for a given B, belongs to the set of (global) maximizers of (13). A maximizer can be computed by solving the linear program (13) using dedicated methods in roughly $O(|\Omega|^3)$ operations [22].

It is worth noting that these subgradients are defined up to an additive constant. Indeed if (u,v) is a solution to (13), then so is $(u+\alpha, v-\alpha)$. Nevertheless, this is of no importance for the computation of the velocity vector (6) which is invariant to constant perturbations of u, since for any $w \in \mathrm{L}^2(\Omega)$, $[DP(w)^*](\mathbf{1}) = 0$.

When the subdifferential is a singleton $\partial_1 W(A,B) = \{u\}$ (meaning that (13) has a strict global maximizer), then $A \mapsto W(A,B)$ is Fréchet differentiable at A and $\nabla_1 W(A,B) = u$. In the other cases, one can take any subgradient of $W(A,B)$ in lieu of $\nabla_1 W(A,B)$ in (6).

4 Sliced Wasserstein Distance

The computation of $W(A,B)$ and $\nabla_1 W(A,B)$ for the Wasserstein distance (12) is however demanding for large-scale and high-dimensional histograms. Numerically, the computation time is acceptable (less than a second on a standard laptop computer) for only $|\Omega| \ll 10^3$. For the segmentation application we are targeting, typically 3-D color histograms with at least 10^4 grid points, solving this linear program is not an option.

To speed up the computation, we follow [20] and consider an alternative distance that mimics the properties of the Wasserstein distance, but is faster to compute. In the following, we restrict our attention to the L^p Wasserstein distance, meaning that $C(x,y) = \|x-y\|^p$.

4.1 Sliced Distance Approximation

The sliced Wasserstein distance reads

$$SW(A,B) = \sum_{\theta \in \Theta} W(A_\theta, B_\theta) \tag{14}$$

where Θ is a finite subset of the unit sphere in $\mathbb{R}^d$, and $A_\theta \in \mathcal{D}(\Omega_\theta)$ is the projected distribution in the direction θ, defined on 1-D grid points $\Omega_\theta = \{x_\theta = \langle x, \theta \rangle\}_{x \in \Omega}$, that has the same values as A, i.e. $\forall\, x \in \Omega, A_\theta(x_\theta) = A(x)$. The sliced Wasserstein distance is thus a sum of 1-D Wasserstein distances between the projected distributions. Note that the grid Ω_θ that supports these distributions is non-uniform, so that some care is needed to perform the computation, as detailed in the following section.

The method described in [20] considered probability distributions $A(x)$ defined on a varying grid $x \in \Omega$ (that depends on A) with constant weights $A(x) = 1/|\Omega|$. But this entails non-convexity of the function $A \mapsto W(A,B)$. In contrast, we consider here distributions $A \in \mathcal{D}(\Omega)$ on a fixed grid Ω but with varying values $A(x)$, which thus preserves convexity of the sliced Wasserstein distance, whose subdifferential $\partial_1 SW(A,B)$ can be easily computed as we will show shortly.

Proposition 4 *The function $A \mapsto SW(A,B)$ is closed and convex and its subdifferential at A is such that*

$$\forall\, x \in \Omega, \quad \partial_1 SW(A,B)(x) = \sum_{\theta \in \Theta} \partial_1 W(A_\theta, B_\theta)(x_\theta)\theta.$$

Proof. Closedness and convexity are preserved under the sum. The formula is a consequence of subdifferential properties and calculus rules [14] and linearity of the projection $A \mapsto A_\theta$.

4.2 Wasserstein Distance on a Non-Uniform 1-D Grid

The ability to compute (14) and to (sub)differentiate it, is conditioned to that of computing quickly the Wasserstein distance between two distributions $A, B \in \mathcal{D}(\Omega)$ defined on a 1-D grid $\Omega \subset \mathbb{R}$. We assume that the grid points $\Omega = \{x_i\}_{i=1}^N$ are sorted in increasing order, $x_i \leqslant x_{i+1}$. This is achieved in $O(|\Omega| \log(|\Omega|))$ operations, which will turn out to be main computational cost of the discrete 1-D Wasserstein distance and its (sub)gradients.

L^p Wasserstein Distance. The L^p Wasserstein distance on the real line for any $p \geqslant 1$ reads

$$W(A,B) = \int_0^1 |R_A^{-1}(t) - R_B^{-1}(t)|^p \mathrm{d}t \ ,$$

$R_A(s) = \int_{-\infty}^s A(x)\mathrm{d}x$, is the cumulative distribution function (CDF) and $R_A^{-1}(t) = \inf\{s \setminus R_A(s) \geqslant t\}$ its pseudo-inverse. The latter is well defined as the CDF is non-decreasing. When $A \in \mathcal{D}(\Omega)$ is discrete, the CDF is equal to $R_A(s) = \sum_{x_i \leqslant s} A(x_i)$.

Wasserstein distance sub-differential. The following proposition gives the sub-gradients of the Wasserstein distance for discrete distributions.

Proposition 5 *Let $A, B \in \mathcal{D}(\Omega)$. For $p \geqslant 1$, the sub-gradients of $A \mapsto W(A,B)$ are written as*

$$\nabla_1 W(A,B) : x_i \mapsto \sum_{j \geqslant i} |x_j - \tilde{x}_j|^p - |x_{j+1} - \tilde{x}_j|^p \tag{15}$$

where

$$\begin{cases} \tilde{x}_j = x_k \quad \textit{if} \quad R_B(x_{k-1}) < R_A(x_j) < R_B(x_k), \\ \tilde{x}_j \in [x_k, x_{k+1}] \quad \textit{if} \quad R_A(x_j) = R_B(x_k). \end{cases}$$

Proof. For simplicity, we treat the case where the set of values of R_A and R_B are disjoints, and $A \mapsto W(A,B)$ is thus differentiable. The general case is obtained by replacing R_B^{-1} by a set-valued mapping. We can write

$$W(A,B) = \tilde{W}(R_A, B) \quad \text{where} \quad \tilde{W}(R,B) = \sum_{i=0}^{N-1} \int_{R(x_i)}^{R(x_{i+1})} |x_{i+1} - R_B^{-1}(s)|^p \mathrm{d}s.$$

Deriving with respect to $R_A(x_j)$ gives

$$\nabla_1 \tilde{W}(R_A, B) : x_j \mapsto |x_j - R_B^{-1}(R_A(x_j))|^p - |x_{j+1} - R_B^{-1}(R_A(x_j))|^p.$$

Using the chain rule, introducing R^* the adjoint of $A \mapsto R_A$, we obtain

$$\nabla_1 W(A,B) = R^*\left(\nabla_1 \tilde{W}(R_A, B)\right) \quad \text{where} \quad R^*(U) : x_i \mapsto \sum_{j \geqslant i} U(x_j),$$

which gives (15).

5 Numerical Results

5.1 Comparison of Wasserstein and Kullback-Leibler Distances

We first illustrate the difficulty of point-wise statistical metrics such as the Kullback-Leibler symmetrized divergence (11) on a simple instructive example with localized 1-D ($d = 1$) distributions. Figure 1, left, shows an example of a gray-scale image with three concentric regions delimited by two circles of increasing radii $r_0 < r_1$. The distribution of the intensity values within each region Gaussian with a mean value in $\{0.05, 0.8, 0.95\}$ of small variances, so that the resulting image is a mixture of three localized and barely overlapping Gaussians. As can be seen from the histogram in Figure 1 top-right, the correct segmentation should group together the two outer regions which have close means.

Figure 2, shows the energy landscape for circular regions of radius r. The leftmost figure shows this energy for the Wasserstein distance, which provides the proper segmentation. Indeed, circle of radius r_0 is the only local, and hence global, maximum of the L^2 ($p = 2$) Wasserstein distance between the inside and outside regions. The situation is radically different with the Kullback-Leibler divergence (rightmost figure),

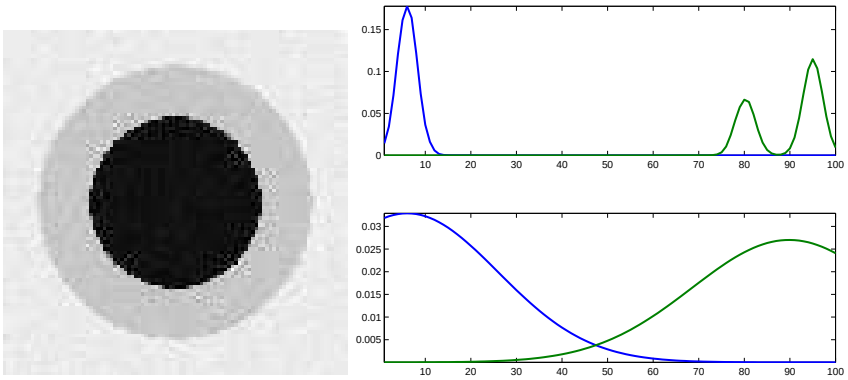

Fig. 1. *Left: gray-scale image example with three concentric regions with radii* $r_0 < r_1$*. Each region has its intensity values Gaussian-distributed with means* $\{0.05, 0.8, 0.95\}$ *and the same variance. Right: estimated densities using the* $P(\chi_\Gamma)$ *with the optimal partition* $\Gamma = \Gamma_{r_0}$ *and using the Gaussian kernel with bandwidth* $s = 10^{-2}$ *(top) and* $s = 0.2$ *(bottom).*

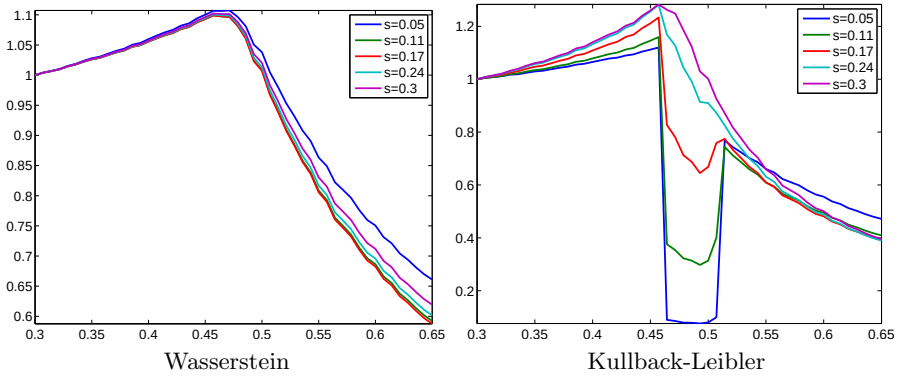

Fig. 2. *Energy* $W(P(\chi_{\Gamma_r}), P(\chi_{\Gamma_r^c}))$ *for a centered circle* Γ_r *of radius* r *as a function of* r*. Each curve corresponds to a different smoothing bandwidth* s *in the Parzen kernel estimator.*

where a spurious local minimum for $r = r_1$ persists, unless an extremely large kernel bandwidth is used. The weakness of point-wise statistical metrics is thus apparent when localized feature histograms come into play, and the smoothing impacts significantly the spatial localization. Note that a similar conclusion is arrived at in [17] using a different segmentation method and a L^1 Wasserstein metric.

5.2 Synthetic Image Segmentation

We consider a synthetic segmentation example with a realization of two 2-D Gaussian mixtures (hence $d = 2$)

$$I(x) \sim \begin{cases} \frac{1}{2}\mathcal{N}(\mu_1, \rho^2) + \frac{1}{2}\mathcal{N}(\nu_1, \rho^2) & \text{if} \quad x \in \Gamma_0, \\ \frac{1}{2}\mathcal{N}(\mu_2, \rho^2) + \frac{1}{2}\mathcal{N}(\nu_2, \rho^2) & \text{if} \quad x \notin \Gamma_0. \end{cases}$$

where the standard deviation of the Gaussians is $\rho = 0.05$, and the means $\mu_i, \nu_i \in \mathbb{R}^2$ are located on the four corners of a square, see Figure 3, last row. Note that the distributions are 2-D, hence the image colors are located in the red/green 2-D plane.

We solve the region competition problem (8) using the sliced L^2 (hence $p = 2$) Wasserstein distance (14). Figure (3) shows the evolution of a discretization of the PDE (10) at various time steps. The top row shows the results obtained using $|\Theta| = 8$ directions evenly distributed on the unit circle of $\mathbb{R}^2$, $\Theta = \{(\cos(k\pi/8), \sin(k\pi/8))\}_{k=0}^{7}$. This produces the expected segmentation since the curve converges to $\partial\Gamma_0$. In contrast, Figure (3), bottom row, shows the evolution computed using two axis-aligned directions $\Theta = \{(1,0),(0,1)\} \subset \mathbb{R}^2$. The sliced Wasserstein distance (14) is thus the sum of the 1-D distances on the red and green channels of the image. The contour gets stuck in a stationary point far from the global minimum. This is due to the fact that the two 2-D mixtures are mapped exactly on the same 1-D mixture when projected on either the red or the green channel.

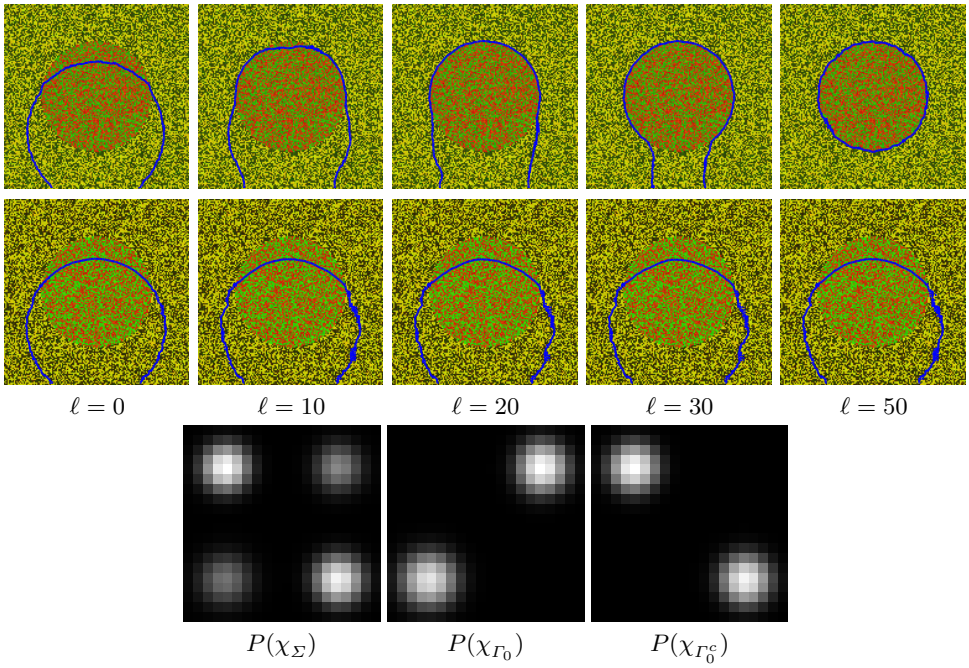

Fig. 3. *Top row: evolution of the segmentation, with ℓ indexing the iterations, for $|\Theta| = 8$ directions. Middle row: same evolution with $|\Theta| = 2$ axis-aligned orientations. Bottom row: 2-D histograms for the whole domain Σ (which is $P(\chi_\Sigma)$) and inside and outside the central disk Γ_0.*

5.3 Natural Images Segmentation

We finally report results on complex color natural images each of size 200×200 pixels. The results are displayed in Figure 4. For these examples we use $|\Theta| = 12$ random directions in 3-D for the computation of the sliced Wasserstein distance (14). In this experiments, we use the L^2 Wasserstein distance, i.e. $p = 2$.

In each case, the initial contour is a circle of radius 0.4 (assuming an image defined over $\Sigma = [0,1]^2$). The figure also depicts 2-D histograms to show how the global color distribution is split by the segmentation algorithm. Only a 2-D slice of the full 3-D histogram is displayed, along the two dominant colors provided by a principal component analysis. The Wasserstein active contour is able to split properly the color space consistently with what an observer would do visually.

6 Conclusion

We have proposed a mathematically grounded way to handle the statistical segmentation problem in arbitrary dimension. Our framework combines wisely the Wasserstein statistical distance and shape derivative tools. It can handle localized distributions, while this property seems to be out of reach for traditional metrics unless a severe smoothing is applied. Such a smoothing would clearly harm the segmentation precision. This approach is quite general and paves the way to many applications. One may for instance think of more advanced features beside simple colors, typically coefficients in the domain of a multi-scale transform or joint statistics between groups of pixels.

References

1. G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. Image segmentation using active contours: Calculus of variations or shape gradients ? *SIAM Applied Mathematics*, 63(6):2128–2154, 2003.
2. V. Caselles, F. Catté, T. Coll, and F. Dibos. A geometric model for active contours in image processing. *Numerische Mathematik*, 66(1):1–31, 1993.
3. V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.
4. T. Chan and L. Vese. Active contours without edges". *IEEE Trans. Image Proc.*, 10(2):266–277, 2001.
5. L. D. Cohen and R. Kimmel. Global Minimum for Active Contour models: A Minimal Path Approach. *International Journal of Computer Vision*, 24(1):57–78, Aug. 1997.
6. M. C. Delfour and J.-P. Zolésio. *Shapes and geometries: analysis, differential calculus, and optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.
7. Julie Delon. Midway image equalization. *Journal of Mathematical Imaging and Vision*, 21(2):119–134, 2004.
8. M. Heiler and C. Schnorr. Natural image statistics for natural image segmentation. *Proc. ICCV*, pages 1259–1266, 2003.
9. A. Herbulot, S. Jehan-Besson, S. Duffner, M. Barlaud, and G. Aubert. Segmentation of vectorial image features using shape gradients and information measures. *Journal of Mathematical Imaging and Vision*, 25(3):365–386, October 2006.

10. S. Jehan-Besson, M. Barlaud, and G. Aubert. DREAM^{2}S: Deformable regions driven by an eulerian accurate minimization method for image and video segmentation. *International Journal of Computer Vision*, 53(1):45–70, 2003.
11. S. Jehan-Besson, M. Barlaud, G. Aubert, and O. Faugeras. Shape gradients for histogram segmentation using active contours. In *International Conference on Computer Vision*, 2003.
12. M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.
13. J. Kim, J. W. Fisher, A. J. Yezzi, M. Cetin, and A. S. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Trans. Image Proc.*, 14(10):1486–1502, October 2005.
14. C. Lemaréchal and J.-B. Hiriart-Urruty. *Convex Analysis and Minimization Algorithms I and II*. Springer, 2nd edition, 1996.
15. P. Martin, P. Réfrégier, F. Goudail, and F. Guérault. Influence of the noise model on level set active contour segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:799–803, 2004.
16. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, XLII(4), 1989.
17. K. Ni, X. Bresson, T. Chan, and S. Esedoglu. Local histogram based segmentation using the wasserstein distance. *International Journal of Computer Vision*, 84:97–111, August 2009.
18. N. Paragios and R. Deriche. Geodesic active regions: A new framework to deal with frame partition problems in computer vision. *Journal of Visual Communication and Image Representation*, 13(1/2):249–268, March 2002.
19. J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009.
20. J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. *Proc. SSVM'11*, 2011.
21. R. Ronfard. Region-based strategies for active contour models. *International Journal of Computer Vision*, 13(2):229–251, 1994.
22. Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.
23. G. Unal, A. J. Yezzi, and H. Krim. Information-theoretic active polygons for unsupervised texture segmentation. *International Journal of Computer Vision*, 62(3):199–220, 2005.
24. C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.
25. S. Zhu and A. Yuille. Region competition: unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Trans. Patt. Anal. and Mach. Intell.*, 18(9):884–900, 1996.

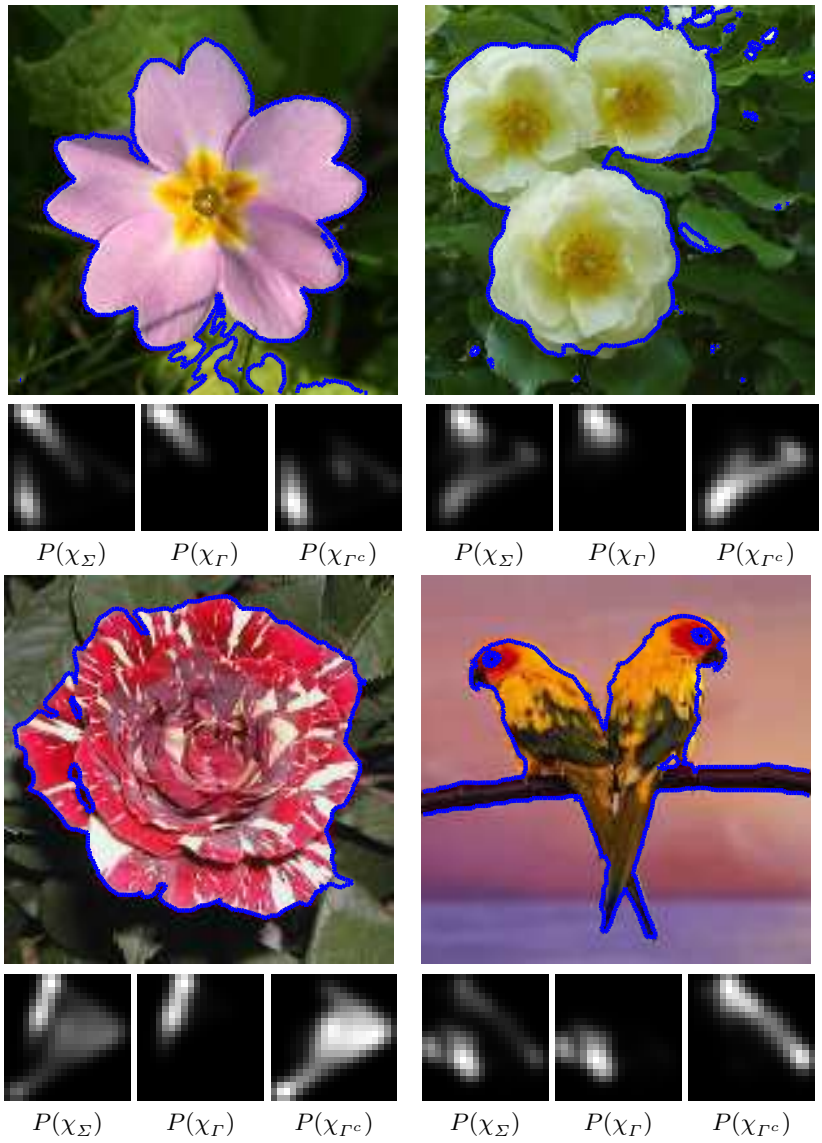

Fig. 4. *Left: example of natural image segmentation. Below each image is displayed the 2-D histogram of the image (in the space of the two dominant color eigenvectors as provided by a PCA) for the whole domain Σ (which is $P(\chi_\Sigma)$) and over the inside and outside the segmented region.*