Annals of Mathematical Sciences and Applications
Volume 0, Number 0, 1, 2017

New region force for variational models in image segmentation and high dimensional data clustering

Ke Wei,* Ke Yin,† Xue-Cheng Tai‡ and Tony F. Chan

We propose an effective framework for multi-phase image segmentation and semi-supervised data clustering by introducing a novel region force term into the Potts model. Assume the probability that a pixel or a data point belongs to each class is known a priori. We show that the corresponding indicator function obeys the Bernoulli distribution and the new region force function can be computed as the negative log-likelihood function under the Bernoulli distribution. We solve the Potts model by the primal-dual hybrid gradient method and the augmented Lagrangian method, which are based on two different dual problems of the same primal problem. Empirical evaluations of the Potts model with the new region force function on benchmark problems show that it is competitive with existing variational methods in both image segmentation and semi-supervised data clustering.

1. Introduction

Image segmentation plays an important role in image processing and appears in a wide range of applications, including computer vision [35], stereo [24, 25] and 3D reconstruction [41]. Given an image $I(x)$ defined over a domain $\Omega \in \mathbb{R}^2$, the task is to partition Ω into different subdomains so that $I(x)$ has different properties over each subdomain. According to different criteria, image segmentation can be divided into two-phase segmentation vs. multi-phase segmentation, automatic segmentation vs. user-assisted segmentation, and discrete approach based segmentation vs. continuous approach based segmentation, just to name a few. In this paper we study the multi-phase image segmentation problem where the number of partitions, denoted by K, is known a priori.

*The work of this author was supported by the National Science Foundation under grant number DTRA-DMS 1322393.

†Corresponding author.

‡This author acknowledges the support from Norwegian Research Council through ISP-Matematikk (Project no. 239033/F20).

In the spatially discrete setting, a digital image is usually modeled as a graph, and the solution to the multi-phase image segmentation problem can be computed from the min-cut or max-flow solutions of the graph, see [23, 26, 7, 6] and references therein. In contrast, variational approaches have been widely studied in the spatially continuous setting, where image segmentation is typically formulated as a continuous energy-functional minimization problem over the image domain. We consider the Potts model for multi-phase image segmentation. In the simplest form, the Potts model attempts to partition an image by minimizing an energy-functional which combines a region force term and an edge force term. It has several advantages compared with the graph-based approaches: (i) it can avoid the metrication errors owing to the crucial rotation invariance property; (ii) a wide range of reliable numerical algorithms are available, and those algorithms can be easily implemented and accelerated; (iii) it requires less memory in computation; (iv) it is easy to use GPU and other parallel computing systems. The active contours which first appeared as the snake model in [22] is another well known approach in variational models. In the active contours approach, the boundaries of each subdomain are modeled as curves, which can be evolved by minimizing an energy-functional.

Data clustering (or classification) is a fundamental task in machine learning which is about partitioning a large data set into a number of clusters that can be well interpreted from a practical perspective. In general, data clustering can be roughly divided into three groups: unsupervised clustering, supervised clustering and semi-supervised clustering. In this paper we consider the multi-class semi-supervised data clustering problem in which the number of clusters is given and there are a few labelled data points in each cluster. The goal is to infer labels for the rest of data points from the already labelled ones. In practice, data points are typically modeled as vertices of a weighted graph where the weights on the edges describe the affinity between each pair of data points. Many algorithms have been developed under the graphical model. For example, the idea of geometric diffusion was developed for semi-supervised clustering in a seminal paper by Coifman et al. [15] . The propagation of labels in geometric diffusion is driven by a diffusion kernel on the weighted graph of the data points. Moreover, the diffusion map based on the eigenvectors of the graph Laplacian embeds the data points into a feature space with the diffusion distance as a new metric. Variational approaches have also been extended from image segmentation in spatially continuous domain to data clustering on weighted graphs, which will be the focus of this paper. In [9], the Mumford-Shah-Potts model [33, 36] was demonstrated to be effective for data clustering, where

the Cheeger cut, formulated as the sum of a modified total variation of the cluster indicator functions, is used. The Cheeger cut can be interpreted as the perimeter of a cluster normalized by the imbalance of the cluster sizes and hence acts as the edge force in the model. In [21, 27, 19, 30], the authors attempt to extend the Chan-Vese model [13] to data clustering, where the edge force is the total variation of the indicators functions. Furthermore, after approximating the total variation via the phase field representations, a graph-based Merriman-Bence-Osher (MBO [31]) scheme is developed to solve the diffusion equations with double-well potential. In all the aforementioned variational approaches, the region force term is either based on the distance between the data points and the cluster centroids or based on the mismatch of the labels over the already labelled data.

The main contributions of this work are summarized in the following.

(i) We introduce a new region force term into the Potts model for both image segmentation and high dimensional data clustering. Compared with the snake model, a region force term was introduced for image segmentation in the Chan-Vese model [11]. Some earlier works have tried to introduce region force into variational models for data clustering, see for example [21, 27, 2]. The region force introduced in [45] overcomes some of the difficulties and shows good numerical performance. In the present work, we derive a new region force function and show its applications to multi-phase image segmentation and semi-supervised data clustering.

(ii) The variational model we use in this paper is the Potts model. Using graph total variation, one can easily extend this model to high dimensional data clustering [8][29]. Following [4, p.116], [46, p.386] and [47, 42], the Potts model has two different dual formulations. Related to these two dual formulations, we present two numerical algorithms. One is for the first dual formulation using a primal-dual algorithm, while the other one is for the second dual formulation using an augmented Lagrangian algorithm.

(iii) Numerical experiments show the good performce of the new region force function and demonstrate the effectiveness of the numerical algorithms. The tests for image segmentation show that the new region force function is as effective as the widely used L_2 fidelity in the literature. However, the L_2 fidelity or the Euclidean distance is not applicable for semi-supervised data clustering problems when there exists complex geometry within the data, while our region force function still works very well. The numerical results for semi-supervised data clustering show that our approach can achieve higher classification accuracy than other existing variational methods. Meanwhile, it is much easier to implement the numerical algorithms for our approach.

The remainder of this paper is organized as follows. The Potts model and the corresponding primal-dual formulations for image segmentation and data clustering are presented in Sections 2.1 and 2.2. The new region force function is introdcued in Section 2.2.2. In Section 3, we present the numerical algorithms for the Potts model, and Section 4 contains the numerical simulations. We conclude this paper in Section 5 with some additional remarks about future directions.

2. Variational models and primal-dual formulations

For image segmentation, the snake models only consider edge force [22]. The Chan-Vese model [11] introduced a region force into variational image segmentation. There were efforts to extend these region force for data clustering, [27, 21]. However, these extensions have problems for data with complex geometries. In [45, 2], the authors successfully introduced a region force for data clustering. In the following, we shall continue in this direction and will introduce a new region force for both image segmentation and data clustering. Moreover, we will combine them with efficient algorithms based on two different primal-dual formulations of the primal problem.

To make the connections between image segmentation and data clustering clear, we first present the model for traditional image segmentation. Since data clustering can be formulated as a graph partitioning problem, essentially the same mathematical model can be established for it based on the graph total variation. Therefore, the new region force function equally applies for image segmentation and data clustering.

2.1. Multi-phase image segmentation

Let us start with image segmenation. Given a gray scale image function $I : \Omega \mapsto R$, the two-phase Chan-Vese [11] model is trying to solve the following minimization problem:

$$\min_{\phi,c_1,c_2} \lambda_1 \int_\Omega |I(x) - c_1|^2 H(\phi) \mathrm{d}x + \lambda_2 \int_\Omega |I(x) - c_2|^2 (1 - H(\phi)) \mathrm{d}x + \mu \int_\Omega |\nabla H(\phi)| \mathrm{d}x,$$

where (i) ϕ is a level set function whose zero level curves set represents the segmentation boundary, (ii) $H(\cdot)$ is the Heaviside function, (iii) c_1 and c_2 are two real numbers, and (iv) λ_1 and λ_2 and μ are positive numbers.

In this work, we shall use a more general model that extends the above mode, i.e., the so-called Potts model. The Potts model for multi-phase image segmentation tries to minimize the following energy-functional:

$$\min_{\{\Omega_k\}_{k=1}^K} \sum_{k=1}^{K} \int_{\Omega_k} f_k(x)\mathrm{d}x + R\left(\{\Omega_k\}_{k=1}^K\right), \tag{1}$$

where $\{\Omega_k\}_{k=1}^K$ is a partition of Ω such that $\cup_{k=1}^K \Omega_k = \Omega$ and $\Omega_k \cap \Omega_{k'} = \emptyset$ for $k \neq k'$. The integrand $f_k(x)$ in (1) is usually referred to as the *region force function* or the *fidelity term*. In case that $f_k(x) = |I(x) - c_k|^2$, we recover the Chan-Vese model [11] or the piecewise constant Mumford-Shah model [33]. The regularization term $R\left(\{\Omega_k\}_{k=1}^K\right)$ measures the geometry properties of the boundaries of $\{\Omega_k\}_{k=1}^K$. The regularizer used in this paper is the sum of the weighted length of each boundary, i.e.,

$$R\left(\{\Omega_k\}_{k=1}^K\right) = \sum_{k=1}^{K} |\partial\Omega_k|_\alpha = \sum_{k=1}^{K} \int_{\partial\Omega_k} \alpha(x)ds, \tag{2}$$

where $\alpha(x) \geq 0$ is normally called an edge detector. A popular choice for the edge detector is $\alpha(x) = \frac{\beta}{1+\gamma|\nabla I_\sigma|^2}$ with γ and β being some properly chosen constants and I_σ is a Gaussian smoothing of the image function $I(x)$. In case $\alpha(x) = 1$, the regularizer is the sum of the length of each boundary.

Let $\phi_k(x)(1 \leq k \leq K)$ be an indicator function associated with the k-th sub-domain,

$$\phi_k(x) = \begin{cases} 1 & x \in \Omega_k \\ 0 & x \notin \Omega_k. \end{cases}$$

It is true that $\int_{\partial\Omega_k} \alpha(x)ds = \int_\Omega \alpha(x)|\nabla\phi_k(x)|\mathrm{d}x$, so we can rewrite (1) with the regularizer $R\left(\{\Omega_k\}_{k=1}^K\right) = \sum_{k=1}^K |\partial\Omega_k|_\alpha$ as[1]

$$\min_{\substack{\phi_k\in\{0,1\} \\ [\phi_k]\in S}} \sum_{k=1}^{K} \int_\Omega f_k\phi_k\mathrm{d}x + \sum_{k=1}^{K} \int_\Omega \alpha(x)|\nabla\phi_k|\mathrm{d}x, \tag{3}$$

[1]Throughout this paper we omit the independent variable notation x when there is no risk of confusion, and we use $|\cdot|$, $|\cdot|_1$, and $|\cdot|_\infty$ to denote the l_2-norm, l_1-norm and l_∞-norm, respectively.

where

$$[\phi_k] = (\phi_1, \cdots, \phi_K),$$

and

$$S = \left\{ [\phi_k] : \ \sum_{k=1}^{K} \phi_k = 1, \ 0 \le \phi_k \le 1 \right\}.$$

One can immediately see that (3) is a non-convex optimization problem. Therefore there does not exist a tractable way to compute its global solution reliably. In a seminal paper, Chan, Esedoglu and Nicolova [12] proposed to relax the binary value constraint on ϕ_k to $0 \le \phi_k \le 1$. Based on this relaxation, (3) can be transformed into the following convex programming:

$$\text{(P)} \qquad \min_{\substack{0 \le \phi_k \le 1 \\ [\phi_k] \in S}} \sum_{k=1}^{K} \int_\Omega f_k \phi_k \mathrm{d}x + \sum_{k=1}^{K} \int_\Omega \alpha(x) |\nabla \phi_k| \mathrm{d}x.$$

We will refer to (P) as the *primal problem*. Despite its convexity, numerical algorithms for the primal problem usually suffer from slow convergence rate due to the non-smoothness of the TV term. In this paper, we will present two numerical methods for (P): a primal-dual hybrid gradient descent method and an augmented Lagrangian method. In order to describe those two methods, we first give two dual formulations of (P), one of which leads to the continuous max-flow approach studied in [48].

The first dual formulation of (P) can be obtained using the following equality:

$$\int_\Omega \alpha(x) |\nabla \phi_k| \mathrm{d}x = \max_{|q_k| \le \alpha(x)} \int_\Omega \phi_k \operatorname{div} q_k \mathrm{d}x, \ \ k = 1, \cdots, n, \tag{4}$$

and the min-max theorem [17, Chapter 6, Proposition 2.4]. That is,

$$\begin{aligned}
&\min_{\substack{0 \le \phi_k \le 1 \\ [\phi_k] \in S}} \sum_{k=1}^{K} \int_\Omega f_k \phi_k \mathrm{d}x + \sum_{k=1}^{K} \int_\Omega \alpha(x) |\nabla \phi_k| \mathrm{d}x \\
= &\min_{\substack{0 \le \phi_k \le 1 \\ [\phi_k] \in S}} \max_{|q_k| \le \alpha(x)} \sum_{k=1}^{K} \int_\Omega f_k \phi_k \mathrm{d}x + \sum_{k=1}^{K} \int_\Omega \phi_k \operatorname{div} q_k \mathrm{d}x \\
= &\max_{|q_k| \le \alpha(x)} \min_{\substack{0 \le \phi_k \le 1 \\ [\phi_k] \in S}} \sum_{k=1}^{K} \int_\Omega \phi_k (f_k + \operatorname{div} q_k) \mathrm{d}x
\end{aligned}$$

$$\text{(D1)} \qquad = \max_{|q_k| \le \alpha(x)} \int_\Omega \min_{k=1,\cdots,K} (f_k + \operatorname{div} q_k) \mathrm{d}x,$$

The above dual formulation for (P) was first observed in [4, p.116], where a gradient decent method was developed to solve the smoothed dual problem [4, p.120]. In this work, we shall use the recent developed primal-dual algorithms related to the ones in [18, 10, 50] to solve it, see Algorithm 1.

The second dual formulation of (P) is given by

$$\text{(D2)} \qquad \max_{\lambda} \int_\Omega \lambda \mathrm{d}x \quad \text{subject to} \quad \begin{cases} h_k \le f_k, |q_k| \le \alpha(x) \\ \operatorname{div} q_k - \lambda + h_k = 0. \end{cases}$$

In fact, the above problem is a continuous max-flow problem with flow conservation in a system where the image region is copied K times. The equality $\operatorname{div} q_k - \lambda + h_k = 0$ represents flow conservation in each of the copied regions. The vector function q_k is the flow inside each copy and the scalar function h_k is the flow between the copies with upper flow constraint $h_k \le f_k$, see [46, p.386]. A more comprehensive exploration about the connection between other continuous min-cut and max-flow problems were also discussed in [42]. In order to derive the above max-flow model, we begin with introducing K auxiliary variables $h_k,\ k = 1, \cdots, K$,

$$\begin{aligned} &\min_{\substack{0 \le \phi_k \le 1 \\ [\phi_k] \in S}} \sum_{k=1}^K \int_\Omega f_k \phi_k \mathrm{d}x + \sum_{k=1}^K \int_\Omega \alpha(x) |\nabla \phi_k| \mathrm{d}x \\ &= \min_{\substack{0 \le \phi_k \le 1 \\ [\phi_k] \in S}} \max_{h_k \le f_k} \sum_{k=1}^K \int_\Omega h_k \phi_k \mathrm{d}x + \sum_{k=1}^K \int_\Omega \alpha(x) |\nabla \phi_k| \mathrm{d}x \\ &= \min_{\substack{\phi_k \in \mathbb{R} \\ [\phi_k] \in S}} \max_{h_k \le f_k} \sum_{k=1}^K \int_\Omega h_k \phi_k \mathrm{d}x + \sum_{k=1}^K \int_\Omega \alpha(x) |\nabla \phi_k| \mathrm{d}x. \end{aligned} \tag{5}$$

To show the last line, suppose there exists a $\phi_k < 0$. We can then take the corresponding h_k to be negative infinity so that the inner maximum problem can be arbitrarily large. However, this case can be excluded since a minimization over ϕ_k is followed.

By introducing another variable λ and utilizing (4), one can further see

that (5) is equivalent to

$$(6)\quad \min_{\phi_k\in\mathbb{R}} \max_{\substack{\lambda\\ h_k\le f_k\\ |q_k|\le\alpha(x)}} \int_\Omega (1-\sum_{k=1}^K \phi_k)\lambda \mathrm{d}x + \sum_{k=1}^K \int_\Omega h_k\phi_k \mathrm{d}x + \sum_{k=1}^K \int_\Omega \phi_k \operatorname{div} q_k \mathrm{d}x.$$

At last, the application of the min-max theorem implies that the optimal value of the above min-max problem is equal to the optimal value of the following max-min problem

$$\begin{aligned}
&\max_{\substack{\lambda\\ h_k\le f_k\\ |q_k|\le\alpha(x)}} \min_{\phi_k\in\mathbb{R}} \int_\Omega (1-\sum_{k=1}^K \phi_k)\lambda \mathrm{d}x + \sum_{k=1}^K \int_\Omega h_k\phi_k \mathrm{d}x + \sum_{k=1}^K \int_\Omega \phi_k \operatorname{div} q_k \mathrm{d}x\\
=&\max_{\substack{\lambda\\ h_k\le f_k\\ |q_k|\le\alpha(x)}} \min_{\phi_k\in\mathbb{R}} \int \lambda \mathrm{d}x + \sum_{k=1}^K \int_\Omega \phi_k(\operatorname{div} q_k - \lambda + h_k)\mathrm{d}x\\
=&\max_{\substack{\lambda\\ h_k\le f_k\\ |q_k|\le\alpha(x)}} \int_\Omega \lambda \mathrm{d}x \quad \text{subject to} \quad \operatorname{div} q_k - \lambda + h_k = 0,
\end{aligned}$$

where the last line gives the dual problem in (D2) after rearrangement.

The pair of min-max problems in (P) and (D2) are continuous analogue of the min-cut and max-flow problems in graph theory. In the discrete case, it is well-known that the min-cut problem is equivalent to the max-flow problem. The above analysis suggests this is also true in the spatially continuous setting. The interested reader can find more details about the continuous max-flow approaches in [48, 1, 3] and references therein.

2.2. Semi-supervised clustering

2.2.1. Discrete Potts model for data clustering. Before describing the discrete Potts model for data clustering, we briefly review some concepts related to the graphic model. Our description follows that in [20] (also adopted in [19] and numerous other works). In the graphic model data feature vectors are represented by vertices of a weighted graph $G = (V, E, w)$, where V represents the set of vertices, E represents the set of edges connecting different vertices, and w represents the set of weights on the edges. The graph G is typically sparse in real applications. For instance, in image segmentation each pixel is only connected with its four nearest neighbor

pixels or pixels in a local image patch. In data clustering problems, data points are often assumed to be uniformly distributed on a low dimensional manifold endowed with a Riemannian metric $d(\cdot,\cdot)$. Each data point on the manifold is usually connected with s-nearest neighbors for a small s, and together they form a local patch of the manifold. Therefore the graph G can be constructed by s-Nearest-Neighbor (s-NN). In practice, the number of neighbor points s may be determined by the dimension or co-dimension of the underlying manifold.

There are several interesting weight functions in the literautre, for example the radial basis function (RBF [37])

$$w(x_i, x_j) = \exp(-d(x_i, x_j)^2/(2\epsilon)), \tag{7}$$

and the Zelnik-Manor and Perona function (ZMP [49])

$$w(x_i, x_j) = \exp\left(-d(x_i, x_j)^2/(\sigma(x_i)\sigma(x_j))\right), \tag{8}$$

where ϵ in (7) is a tuning parameter and $\sigma(\cdot)$ in (8) measures the local variance within the data. Another popular weight function in natural language processing is the cosine similarity function [39]

$$w(x_i, x_j) = \cos(x_i, x_j) = \frac{\langle x_i, x_j\rangle}{|x_i||x_j|}. \tag{9}$$

Let $W = (w_{ij})$ be a weight matrix constructed from the weight function and $D = (d_{ii})$ be a diagonal matrix with the i-th diagonal entry being equal to the l_1-norm of the i-th row of W. The normalized affinity matrix defined via $\widehat{W} = D^{-1/2}WD^{-1/2}$ will be used later in the computation of the new region force function.

We will introduce more nations in order to describe the discrete Potts model for data clustering. Let $N = |V|$, the total number of vertices of the graph G. For any $u \in L^2(V)$, the gradient of u at the vertex x_i, denoted by $\nabla u(x_i)$, is defined as

$$\nabla u(x_i) = (\partial_{x_1} u(x_i), \ldots, \partial_{x_N} u(x_i)),$$

where

$$\partial_{x_j} u(x_i) = w_{ij}(u(x_j) - u(x_i)).$$

Here we assume $w_{ij} = 0$ and consequently $\partial_{x_j} u(x_i) = 0$ if $x_i x_j \notin E$. For any $q = (q(x_i)(x_j)) \in L^2(V, L^2(V))$, the divergence of $q(x_i)$, denoted by $\operatorname{div} q(x_i)$, is defined as

$$\operatorname{div} q(x_i) = \sum_{j=1}^{N} w_{ij}(q(x_j)(x_i) - q(x_i)(x_j)).$$

The computation of the divergence of $q(\cdot)$ over all the vertices of G can be proceeded in the following matrix form

$$\operatorname{div} q = (W \circ (q^T - q))\mathbf{1},$$

where $\circ$ denotes the Hadamard product. Moreover, one can easily verify that the divergence operator is the adjoint of the gradient operator which satisfies

$$\langle \nabla u, q \rangle = \langle u, \operatorname{div} q \rangle . \tag{10}$$

Now we are ready to describe the discrete Potts model. Suppose we want to partition the data points into K clusters, denoted by $V_1, \cdots, V_K$. If the corresponding membership function $\phi_k(x_i)$ for the k-th cluster is defined as

$$\phi_k(x_i) = \begin{cases} 1 & \text{if } x_i \in V_k \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

then the discrete counterpart of the Potts model in (3) can be written as

$$\min_{\substack{\phi_k \in \{0,1\} \\ [\phi_k] \in S}} \sum_{k=1}^{K} \sum_{x_i \in V} f_k(x_i)\phi_k(x_i) + \sum_{k=1}^{K} \sum_{x_i \in V} \alpha(x_i) |\nabla \phi_k(x_i)|_1, \tag{12}$$

where $f_k(\cdot)$ is a region force function and $|\nabla \phi_k(x_i)|_1$ is the anisotropic version of the total variation,

$$\begin{aligned} \sum_{x_i \in V} \alpha(x_i) |\nabla \phi_k(x_i)|_1 &= \sum_{x_i \in V} \sum_{x_j \in V} \alpha(x_i) w_{ij} |\phi_k(x_j) - \phi_k(x_i)| \\ &= |\operatorname{diag}(\alpha) W \operatorname{diag}(\phi_k) - \operatorname{diag}(\alpha) \operatorname{diag}(\phi_k) W|_1. \end{aligned}$$

In addition, one also has

$$\sum_{x_i \in V} \alpha(x_i) |\nabla \phi_k(x_i)|_1 = \sum_{x_i \in V} \max_{|q_k(x_i)|_\infty \le \alpha(x_i)} \langle \nabla \phi_k(x_i), q_k(x_i) \rangle$$

$$
\begin{aligned}
&= \max_{|q_k|_\infty \le \alpha(x_i)} \langle \nabla \phi_k, q_k \rangle \\
&= \max_{|q_k|_\infty \le \alpha(x_i)} \langle \phi_k, \operatorname{div} q_k \rangle \\
&= \max_{|q_k|_\infty \le \alpha(x_i)} \sum_{x_i \in V} \phi_k(x_i) \operatorname{div} q_k(x_i), \qquad (13)
\end{aligned}
$$

where in the first line we use the fact that ℓ_∞-norm is the dual norm of ℓ_1-norm, and in the fourth line we apply (10).

It is evident that (12) is a non-convex problem and the application of the same convex relaxation technique as in (P) leads to the following primal problem of the Potts model for data clustering

$$
(\bar{\mathrm{P}}) \qquad \min_{\substack{\phi_k \in [0,1] \\ [\phi_k] \in S}} \sum_{k=1}^{K} \sum_{x_i \in V} f_k(x_i) \phi_k(x_i) + \sum_{k=1}^{K} \sum_{x_i \in V} \alpha(x_i) |\nabla \phi_k(x_i)|_1.
$$

Using a variant of (13) and the same min-max argument as in Section 2.1, we are also able to obtain two different dual formulations for ($\bar{\mathrm{P}}$), which are listed below:

$$
(\bar{\mathrm{D}}1) \quad \max_{|q_k|_1 \le \alpha(x_i)} \sum_{x_i \in V} \min_k (f_k(x_i) + \operatorname{div} q_k(x_i)),
$$

$$
(\bar{\mathrm{D}}2) \quad \max \sum_{x_i \in V} \lambda(x_i) \quad \text{subject to} \quad \begin{cases} h_k(x_i) \le f_k(x_i), |q_k(x_i)|_\infty \le \alpha(x_i) \\ \operatorname{div} q_k(x_i) - \lambda(x_i) + h_k(x_i) = 0. \end{cases}
$$

In this paper, we investigate the discrete Potts model for semi-supervised data clustering. Suppose there exists a small fraction $S_k \subset V_k$ in each cluster such that the label of the data points in S_k is given. The goal is to determine the labels for the rest of the data points from those labelled ones. In our approach, the labelled data points will be used to compute the probabilities in the new region force function presented in the next section.

2.2.2. Effective region force under the Bernoulli model. In [21], Hu, Sunu and Bertozzi extended the Chan-Vese model to data clustering by combing a region force function of the form $f_k(x_i) = |x_i - c_k|^2$ with a special type of edge force function. Here c_k denotes the centroid of the k-th cluster which can be computed as the weighted average of the data points in each cluster. The quadratic region force function defined using the Euclidean distance between the data points and the cluster centroids penalizes

the heterogeneity of the data points within each cluster. It is effective for the Gaussian mixture model, where the data points are homogeneous in a visually smooth region. However, for many data clustering problems, there exists complex geometry within the data points and typically the clusters cannot be distinguished by the centroid of each cluster, for example in the three-circles synthetic data set. In this section, we present a different region force function which can be obtained as the negative log-likelihood function under the Bernoulli model.

Let $p_k(x_i)$ denote the probability of a given data point x_i belonging to the k-th cluster V_k. If $p_k(x_i)$ is known a priori, then the binary value label function $\phi_k(x_i)$ defined in (11) is a random variable which satisfies the Bernoulli distribution and

$$\begin{aligned}\mathbb{P}\{\phi_k(x_i)\} &= \begin{cases} p_k(x_i) & \text{if } \phi_k(x_i) = 1 \\ 1 - p_k(x_i) & \text{if } \phi_k(x_i) = 0 \end{cases} \\ &= (p_k(x_i))^{\phi_k(x_i)}(1 - p_k(x_i))^{1-\phi_k(x_i)}.\end{aligned}$$

Therefore, the negative log-likelihood function over all the data points is given by

$$\begin{aligned}-\sum_{x_i \in V} \log(\mathbb{P}\{\phi_k(x_i)\}) = &\sum_{x_i \in V} \{-\log(p_k(x_i))\phi_k(x_i) + \log(1 - p_k(x_i))\phi_k(x_i)\} \\ &+ Const.,\end{aligned}$$

where the last term is a constant as we assume $p_k(x_i)$ is given. Without any prior information imposed on $\phi_k(x_i)$, we are interested in a realization which minimizes the negative log-likelihood function under the constraint $\sum_{k=1}^{K} \phi_k(x_i) = 1$ for all x_i. This can be achieved by computing the solution of the following minimization problem:

$$\begin{aligned}&\min_{\phi_k(x_i)} \sum_{k=1}^{K} \sum_{x_i \in V} \{-\log(p_k(x_i))\phi_k(x_i) + \log(1 - p_k(x_i))\phi_k(x_i)\} \\ &\text{s. t. } \phi_k(x_i) \in \{0, 1\} \text{ and } \sum_{k=1}^{K} \phi_k(x_i) = 1 \text{ for all } i = 1, \cdots, n.\end{aligned}$$

The above minimization problem provides us a new region force function for the Potts model. That is, we can set

$$f_k(x_i) = -\log(p_k(x_i)) + \log(1 - p_k(x_i)) \tag{14}$$

so that

$$f_k(x_i)\phi_k(x_i) = -\log(p_k(x_i))\phi_k(x_i) + \log(1-p_k(x_i))\phi_k(x_i)$$

in (12). Since $\log(t) \le t-1$ for all $t>0$, we have

$$-\log(p_k(x_i)) + \log(1-p_k(x_i)) \le \frac{1-2p_k(x_i)}{p_k(x_i)}. \tag{15}$$

The numerator in (15) gives the region force function proposed in [45],

$$f_k(x_i) = 1 - 2p_k(x_i). \tag{16}$$

Numerical simulations in Section 4 demonstrate that the region force functions in (14) and (16) are equally effective when used in the Potts model for image segmentation and semi-supervised clustering.

2.2.3. Compute the probability. We now describe how to compute $p_k(x_i)$, the probability that x_i belongs to V_k, in the region force function. The idea behind the computation is simple. If a data point is "much closer" to the labelled data points in a cluster V_k (i.e., the data points in S_k), then with high probability this data point should belong to the k-th cluster. So $p_k(x_i)$ should be proportional to the "closeness" between x_i and S_k. Similar ideas can be found in [44, 43], where a novel learning algorithm based on the random Markov chain model was proposed for semi-supervised clustering.

Recall from Section 2.2.1 that $\widehat{W}$ is the normalized affinity matrix for the data points. Let $\widehat{W}^m = (\widehat{w}_{ij}^{(m)})$ be the m-th power of $\widehat{W}$. In [15], the m-th diffusion distance between two data points x_i and x_j is defined as

$$d^{(m)}(x_i, x_j) = \widehat{w}_{ii}^{(m)} + \widehat{w}_{jj}^{(m)} - 2\widehat{w}_{ij}^{(m)},$$

where $\widehat{w}_{ii}^{(m)}$ (and $\widehat{w}_{jj}^{(m)}$) describes the probability that a random walk starting from x_i (and x_j) returns back to x_i (and x_j) after m steps, and $\widehat{w}_{ij}^{(m)}$ describes the probability that a random walk starting from x_i arrives at x_j after m steps. Thus, $\widehat{w}_{ij}^{(m)}$ measures the closeness between two data points. Based on the diffusion distance, we compute $p_k(x_i)$ in semi-supervised clustering as follows:

$$p_k(x_i) = \frac{\frac{1}{|S_k|}\sum_{j\in S_k} r_{ij}}{\sum_{k'=1}^{K}\frac{1}{|S_{k'}|}\sum_{j\in S_{k'}} r_{ij}}, \tag{17}$$

where

$$r_{ij} = (\widehat{w}_{ij}^{(m)})^2/(\widehat{w}_{ii}^{(m)}\widehat{w}_{jj}^{(m)}),$$

and with a slight abuse of notation $|\cdot|$ denotes the cardinality of a finite set. In the numerical simulations, we take $m = 1$ or 2 and set $p_k(x_i) = 1/K$ when the denominator in (17) is zero.

The region force functions presented in Section 2.2.2 are also applicable for the multi-phase image segmentation problem. That is, we can take $f_k(x)$ to be either $-\log(p_k(x)) + \log(1 - p_k(x))$ or $1 - 2p_k(x)$ in (3) and (P). Assume the image density of each subdomain obeys the Gaussian random model given by $I(x) \sim \mathcal{N}(c_k, \sigma^2)$. For each pixel x in the image domain, the probability of x belonging to the k-th subdomain, denoted by $p_k(x)$, should be proportional to $\exp(-|I(x) - c_k|/2\sigma^2)$. Therefore we can compute $p_k(x)$ as follows:

$$p_k(x) = \frac{\exp(-|I(x) - c_k|/2\sigma^2)}{\sum_{k'=1}^{K} \exp(-|I(x) - c_{k'}|/2\sigma^2)}. \tag{18}$$

3. Algorithms

In this section, we present two numerical algorithms for computing the solutions to the primal problems (P) and ($\bar{\text{P}}$). Since ($\bar{\text{P}}$) is just a discrete version of (P), we only present the algorithms for (P) but note one can easily extend them for ($\bar{\text{P}}$). As stated previously, computing the solution to (P) directly suffers from the non-smoothness of the TV term. Alternatively, we solve (P) by the primal-dual hybrid gradient method and the alternating direction method of multipliers (ADMM) which are targeting the min-cut problem (P) and the max-flow problem (D2), respectively.

In Section 2, two dual problems are presented for the primal problem (P). If $[\phi_k^*]$ is the optimal solution of the primal problem and $[q_k^*]$ is the optimal solution of the first dual problem, then $([\phi_k^*], [q_k^*])$ forms a saddle point of the min-max problem

$$\min_{\substack{0\le\phi_k\le 1\\ [\phi_k]\in S}} \max_{|q_k|\le\alpha(x)} \sum_{k=1}^{K} \int_\Omega f_k\phi_k \mathrm{d}x + \sum_{k=1}^{K} \int_\Omega \phi_k \operatorname{div} q_k \mathrm{d}x.$$

A primal-dual hybrid gradient algorithm can be developed for the above min-max problem, see Algorithms 1. In each iteration, the primal and dual variables are updated successively by a projected gradient descent step, followed

by an acceleration using the Nesterov's memory technique. Algorithm 1 is a special case of the general primal-dual algorithms that have been well studied in the literature. The convergence analysis of the primal-dual algorithms can be found in [18, 10, 50, 5].

Algorithm 1 Primal-Dual Hybrid Gradient (PDHG)

1. Update the dual variables $[q_k]$ by

$$q_k^{l+1} = \Pi_{|q_k| \le \alpha(x)}(q_k^l - \beta_k \nabla \phi_k^l).$$

2. Update the primal variables $[\phi_k]$ by

$$[\phi_k^{l+1}] = \Pi_S[\phi_k^l - \gamma_k(\operatorname{div} q_k^l + f_k)].$$

3. Combine two adjacent steps

$$[\phi_k^{l+1}] = \theta[\phi_k^l] + (1-\theta)[\phi_k^{l+1}],$$

where we choose $\theta = -0.5$.

As noted below the second dual formation (D2), the continuous Potts model (P) can be interpreted as a continuous min-cut problem, while the corresponding max-flow problem is given by its dual formulation in (D2). Therefore we can instead solve the dual problem by the ADMM algorithm. First note that the augmented Lagrangian associated with (D2) is

$$\begin{aligned} L(\lambda, [h_k], [q_k], [\phi_k]) &= \int_\Omega \lambda \mathrm{d}x \\ &+ \sum_{k=1}^K \int_\Omega \phi_k(\operatorname{div} q_k - \lambda + h_k)\mathrm{d}x - \frac{c}{2}\sum_{k=1}^K \int_\Omega (\operatorname{div} q_k - \lambda + h_k)^2 \mathrm{d}x. \end{aligned} \tag{19}$$

Here we use $[\phi_k]$ to denote the Lagrangian multipliers because when we use ADMM to solve the dual problem based on the augmented Lagrangian, the Lagrangian multipliers converge to the primal optimal solution. The ADMM algorithm for the augmented Lagrangian functional (19) is presented in Algorithm 2. In the algorithm, each dual variable is updated by solving a minimization subproblem when the other variables are fixed, and the Lagrangian multiplies are updated by a gradient descent step. While the closed-form solutions to the minimization problems with respect to λ and $[h_k]$ can be computed easily, the minimization problem with respect to $[q_k]$ does not have

an explicit solution. However, we can compute its solution approximately using one step projected gradient descent.

Algorithm 2 ADMM for Augmented Lagrangian (ADMM)

1. Update the dual variables
   (a) Update λ by

$$\begin{aligned}\lambda^{l+1} &= \arg\max_{\lambda} \int_{\Omega} \lambda \mathrm{d}x - \frac{c}{2}\sum_{k=1}^{K}\int_{\Omega}(\operatorname{div} q_k^l - \lambda + h_k^l - \frac{\phi_k^l}{c})^2 \,\mathrm{d}x \\ &= \frac{1}{K}\sum_{k=1}^{K}(\operatorname{div} q_k^l + h_k^l - \frac{\phi_k^l}{c}) + \frac{1}{Kc}.\end{aligned}$$

   (b) Update $[h_k]$ by

$$\begin{aligned}h_k^{l+1} &= \arg\max_{h_k \le f_k} -\int_{\Omega}(\mathrm{div} q_k^l - \lambda^{l+1} + h_k - \frac{\phi_k^l}{c})^2 \,\mathrm{d}x \\ &= \min\{\frac{\phi_k^l}{c} + \lambda^{l+1} - \mathrm{div} q_k^l, f_k\}.\end{aligned}$$

   (c) Update $[q_k]$ by

$$q_k^{l+1} = \arg\max_{|q_k| \le \alpha(x)} -\int_{\Omega}(\mathrm{div} q_k - \lambda^{l+1} + h_k^{l+1} - \frac{\phi_k^l}{c})^2 \,\mathrm{d}x,$$

   which can be approximately solved by one step of projected gradient descend,

$$q_k^{l+1} = \Pi_{|q_k| \le \alpha(x)}(q_k^l + \beta_l \nabla(\operatorname{div} q_k^l - \lambda^{l+1} + h_k - \frac{\phi_k^l}{c})).$$

2. Update the Lagrangian multipliers

$$\phi_k^{l+1} = \phi_k^l - c(h_k^{l+1} + \operatorname{div} q_k^{l+1} - \lambda^{l+1}).$$

4. Numerical Experiments

In this section, we explore the performance of the Potts model (P) and ($\bar{\text{P}}$) with the new region force function for multi-phase image segmentation and semi-supervised data clustering, and test the efficiency of the numerical algorithms presented in Section 3. Both PDHG (Algorithm 1) and ADMM (Algorithm 2) are implemented in MATLAB® and executed on a laptop.

4.1. Multi-phase image segmentation

We first examine the performance of the proposed region force function listed in (14) on RGB image segmentation problems, and compare them with the widely used l_2 region force function

$$f_k(x) = |I(x) - c_k|^2, \tag{20}$$

and the one in (16). The test images are obtained from BSDS500 [28], see Figures 1a, 2a, 3a, and 4a. They are discretized on a Cartesian grid with the weights on all the edges being equal to 1, and the image density centroids (i.e., c_k, $1 \le k \le K$) are computed by the `kmeans` algorithm in Matlab. We use un-smoothed images (i.e., $\sigma = 0$) to compute the edge detector in (2). The values of K and the values of β, γ in each image segmentation test are listed in Table 1. The probability function $p_k(x)$ in (14) and (16) is computed via (18) with unit variance.

Table 1: Parameters used for the image segmentation tests in Figures 1b to 1d, 2b to 2d, 3b to 3d, and 4b to 4d.

	1b	1c	1d	2b	2c	2d
K	4	4	4	7	7	7
β	0.6	0.3	0.5	0.6	0.25	0.5
γ	50	70	70	55	75	60
	3b	3c	3d	4b	4c	4d
K	10	10	10	6	6	6
β	1.35	0.45	1.35	1.45	0.5	1.35
γ	55	100	55	45	55	55

We solve the Potts model (P) with the three different region force functions using PDHG (Algorithm 1) and ADMM (Algorithm 2), where we set $\beta_l = \gamma_l = 0.4$ in PDHG, and $\beta_l = 0.05$ and $c = 0.1$ in ADMM. In order

to monitor the convergence of the algorithm, we record two quantities: The primal energy

$$E_P([\phi_k]) = \int_\Omega \sum_{i=1}^{K} f_k \phi_k + \alpha |\nabla \phi_k| \, \mathrm{d}x, \tag{21}$$

and the dual energy

$$E_D([q_k]) = \int_\Omega \min_{k \in \{1,\ldots,K\}} (f_k + \mathrm{div} q_k) \, \mathrm{d}x. \tag{22}$$

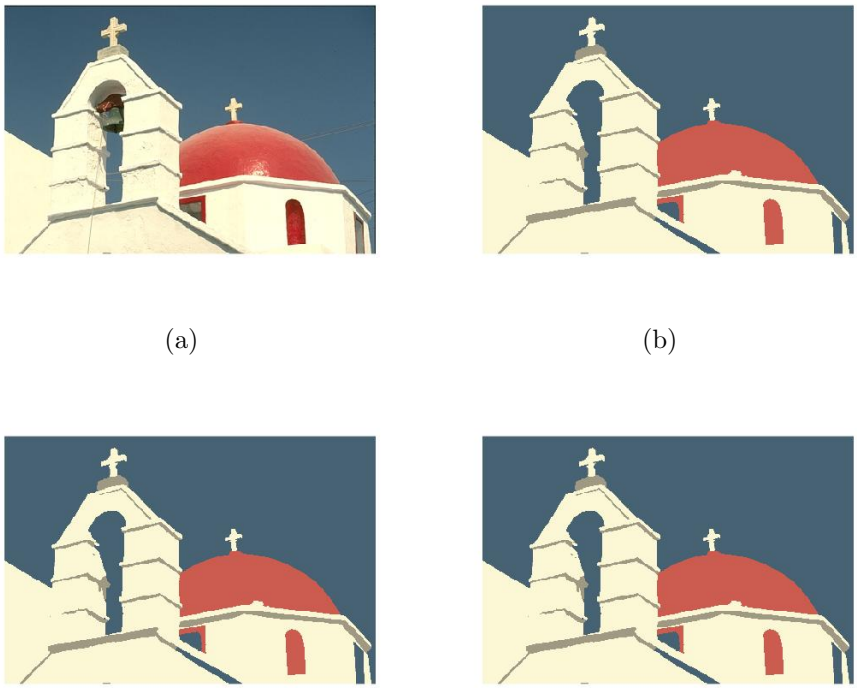
Because the primal variables $[\phi_k]$ from Algorithm 2 could violate the simplex constraint, the duality gap is not always great than or equal to zero. However, as the algorithm converges, the duality gap is approaching zero. Therefore, both PDHG and ADMM are terminated if either of the following two conditions is satisfied: a) a maximum of 2500 iterations is reached; b) the relative absolute duality gap is smaller than ϵ,

$$\frac{|E_P - E_D|}{|E_P|} \leq \epsilon, \tag{23}$$

where $\epsilon = 10^{-5}$ in the experiments for image segmentation.

The segmentation results obtained from PDHG and ADMM are visually close, so we only present the ones obtained from PDHG in Figures 1 to 4. These figures show that for image segmentation the new region force function (14) proposed in this paper is as effective as the L_2 fidelity and the one proposed in [45]. We direct the interested reader to [16] for indirect comparisons of the segmentation results with other approaches.

We present the computational results of PDHG and ADMM in Table 2, which shows both methods can solve the segmentation with very good efficiency. For these test images, we find that PDHG requires many fewer iterations and less computation time than ADMM to converge to the moderate accuracy. However, the computing time could also be in favor of the ADMM method for some other images. The table also shows an interesting feature about the three region force functions. Typically, it requires the least number of iterations and computation time for PDHG and ADMM to compute the solution of the Potts model with the region force function (14), while it requires the most number of iterations and computation time to compute the solution of the Potts model with the region force function (20).



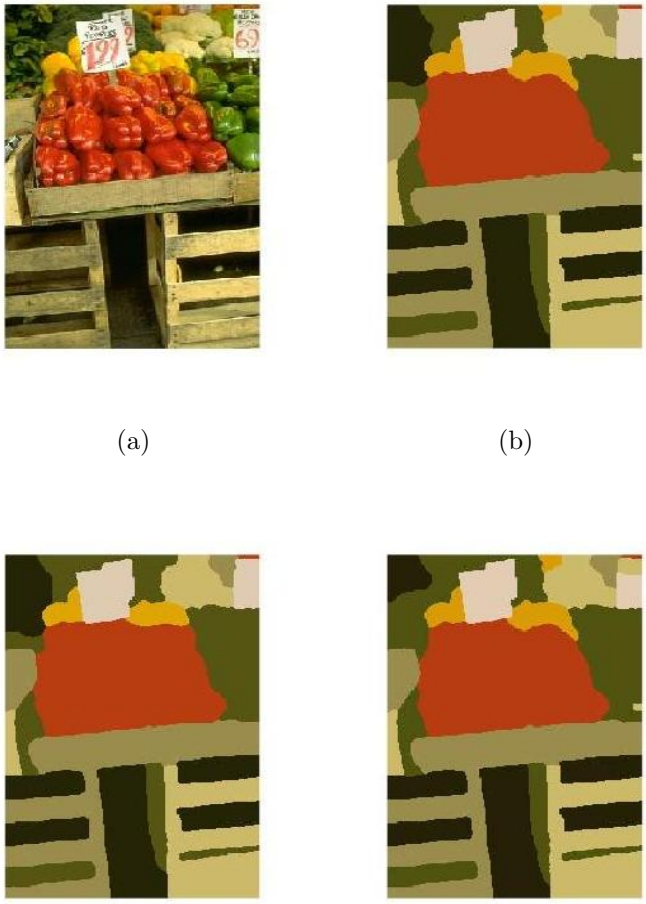
(a) (b)

(c) (d)

Figure 1: (a): Original image; (b), (c), (d): Segmentation results obtained from the Potts model with the region force functions (14), (16), and (20), respectively.

4.2. Semi-supervised data clustering

Next, we evaluate the performance of the new region force function on three benchmark semi-supervised clustering data sets: `Three-Circles`, `COIL`, and `MNIST`. `Three-Circles` is a synthetic data set which are constructed from three circles having an identical center. We first create three circles on the 2D plane, centered at (0,0) with radii 1, 2, and 3, and then sample 6000 points uniformly at random from these circles. The sampled points are embedded into $\mathbb{R}^{100}$ by padding 98 zeros to their end, followed by the perturbation of each coordinate with i.i.d Gaussian noise of mean 0 and variance 0.16. The `COIL` data set is downloaded from the supplementary ma-

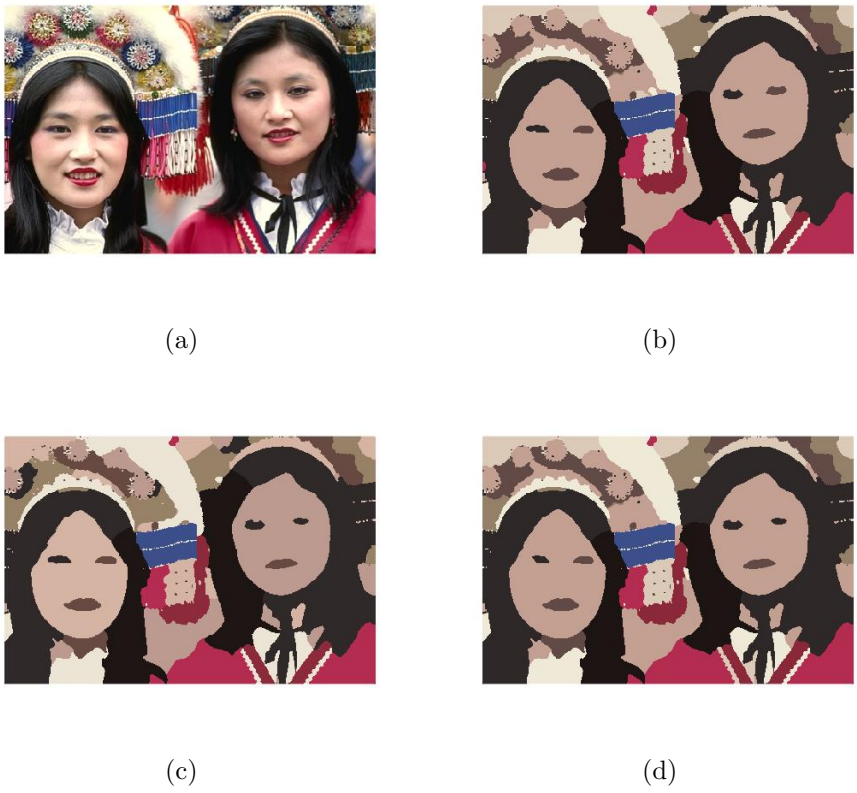


(a) (b)

(c) (d)

Figure 2: (a): Original image; (b), (c), (d): Segmentation results obtained from the Potts model with the region force functions (14), (16), and (20), respectively.

terial of [14] (http://olivier.chapelle.cc/ssl-book/benchmarks.html, originally from COIL-100 [34]). It contains 1500 natural images of 6 different objects taken from various angles. All the images are preprocessed



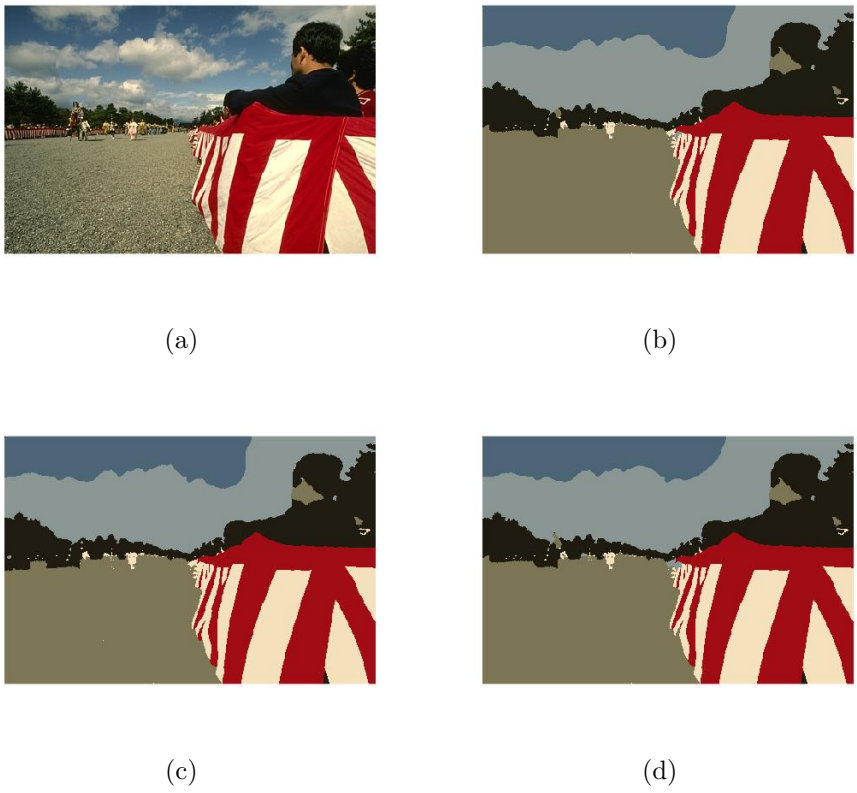
(a) (b)

(c) (d)

Figure 3: (a): Original image; (b), (c), (d): Segmentation results obtained from the Potts model with the region force functions (14), (16), and (20), respectively.

to the same size, and the labels for the images are also contained in the data set. `MNIST` is obtained from "The MNIST Database of Handwritten Digits" (`http://yann.lecun.com/exdb/mnist/`), which consists of 70,000 gray-scale images of labeled handwritten digits from 0 to 9, all scaled to the same size. The basic properties of the three data sets are listed in Table 3.

The graph G for each test data set is constructed as a s-nearest-neighbor (s-NN) graph under the l_2-metric. We make use of an implementation of the randomized kd-tree [38, 32], called `VLFeat` [40], to find the s-nearest neighbors of each data point. The Zelnik-Manor and Perona weight function in (8) is used to construct the affinity matrix, where the standard deviation



(a) (b)

(c) (d)

Figure 4: (a): Original image; (b), (c), (d): Segmentation results obtained from the Potts model with the region force functions (14), (16), and (20), respectively.

at a data point is estimated using the distance between the data point and its s-th nearest neighbor.

In the tests, a small fraction of data points are drawn uniformly at random from each data set and marked as labelled data, and then we apply the discrete Potts model ($\bar{\text{P}}$) to determine the labels for the rest of the data points. The probabilities $p_k(x_i)$ used to define the region force functions (14) and (16) are computed via (17) for $m = 1$ or $m = 2$. We choose $\alpha(x_i)$ to be a constant, denoted by α, in ($\bar{\text{P}}$) . The solution to the discrete Potts model is computed by PDHG (Algorithm 1) and ADMM (Algorithm 2). The algorithms are terminated using the same criteria as in Section 4.1 but

Table 2: Number of iterations and computational time of PDHG and ADMM for the image segmentation tests in Figures 1b to 1d, 2b to 2d, 3b to 3d, and 4b to 4d.

		1b	1c	1d	2b	2c	2d
ADMM	#iter	352	1033	2500	921	2500	2500
	time (s)	162.8	417.5	1144.2	761.3	2062.8	2043.5
PDHG	#iter	176	307	642	826	670	2500
	time (s)	58.9	108.9	224.1	492.3	432.1	1344.8
		3b	3c	3d	4b	4c	4d
ADMM	#iter	1911	2500	2500	2363	2500	2500
	time (s)	2231.6	2922.9	2930.5	1671.2	1751.8	1754.6
PDHG	#iter	1751	2500	2500	744	562	2500
	time (s)	1392.9	1905.1	1895.7	386.9	305.4	1151.3

Table 3: Basic properties of `Three-Circles`, `COIL`, and `MNIST`. The original data sets also contain labels for all the data points

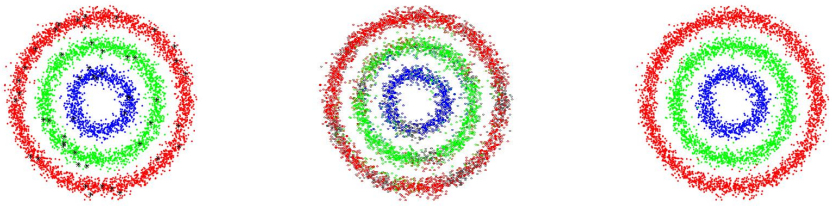
Data set	**Classes**	**Dimension**	**Points**
`Three Circles`	3	100	6000
`COIL`	6	241	1500
`MNIST`	10	784	70,000

with $\epsilon = 10^{-3}$. All the simulations are repeated 10 times.

Table 4: Parameters used in the tests where n is the number of labelled data points in each data set, s is the number of neighbors used to construct the graph, m is the value used in the computation of $p_k(x_i)$, and α is the TV weight in the Potts model. The last two rows include the parameters used in PDHG and ADMM.

	`Three Circles`	`COIL`	`MNIST`
n	50	100	350
s	10	5	10
m	2	1	2
α for (14)	3	5.5	5.5
α for (16)	0.5	1.5	1.5
PDHG	$\beta_l = \gamma_l = 0.4$	$\beta_l = \gamma_l = 0.4$	$\beta_l = 0.5l$, $\gamma_l = \frac{0.5}{(1+0.1l)}$
ADMM	$\beta_l = 0.05$, $c = 0.05$	$\beta_l = 0.05$, $c = 0.1$	$\beta_l = 0.05$, $c = 5$

We begin with comparing the performance of the region force functions (14) and (16). When the region force function (14) is used in the discrete



(a) (b) (c)

Figure 5: `Three-Circle` synthetic data: (a) all the data points with labelled data, (b) classification result computed from $p_k(x_i)$, and (c) classification result of the Potts model with (14).

Potts model, we add a small quantity $\delta = 10^{-3}$ to the logarithm and test

$$f_k(x) = -\log(p_k(x) + \delta) + \log(1 - p_k(x) + \delta)$$

to avoid numerical overflow. The values of the parameters used in our tests for the three different data sets are listed in Table 4.

The average classification accuracy (out of the ten random samples of the labelled data) of the Potts model with the region force functions (14) and (16) are listed in Tables 5, 6, and 7 for `Three Circles`, `COIL`, and `MNIST`, respectively. The three tables show that the average classification accuracy of (14) is slightly higher than that of (16), and overall they are equally effective for the tested semi-supervised clustering problems. We also plot the classification result of the Potts model with (14) for `Three Circles` in Figure 5 by projecting the data points onto the first two dimensions. The average computation time and average number of iterations of PDHG and ADMM in each test are also listed in the tables, which show that PDHG is typically faster than ADMM.

We further compare our approach, referred to as Potts-RF, with another two existing variational methods from the literature: multiclass total variation (MTV [8]) and multiclass-MBO [19]. The codes for MTV are downloaded from the author's website, while we reproduce the codes for multiclass MBO using the parameters suggested in [19]. We test three different numbers of labeled samples for each data set. The average classification accuracy of Potts-RF, MTV and multiclass-MBO is listed in Tables 8, 9, and 10 for `Three Circles`, `COIL`, and `MNIST`, respectively. Table 10 shows that the classification accuracy of Potts-RF is only about 0.5% lower than that of

Table 5: Average classification accuracy of the discrete Potts model with the region forces functions (14) and (16) for `Three-Circles`, as well as the average computational time of PDHG and ADMM. The number of labelled data points is 50 (0.83%).

region force	algorithm	accuracy (%)	iterations (ave.)	cpu time (s)
(14)	PDHG	98.2	162.8	2.94
(14)	ADMM	98.2	76.3	2.78
(16)	PDHG	97.9	65.6	1.45
(16)	ADMM	97.9	71.5	2.70

Table 6: Average classification accuracy of the discrete Potts model with the region forces functions (14) and (16) for `COIL`, as well as the average computational time of PDHG and ADMM. The number of labelled data points is 100 (6.7%).

region force	algorithm	accuracy (%)	iterations (ave.)	cpu time (s)
(14)	PDHG	90.90	307.6	2.16
(14)	ADMM	90.90	163.1	1.67
(16)	PDHG	90.25	306.6	2.09
(16)	ADMM	90.25	475.9	4.21

Table 7: Average classification accuracy of the discrete Potts model with the region forces functions (14) and (16) for `MNIST`, as well as the average computational time of PDHG and ADMM. The number of labelled data points is 350 (0.5%).

region force	algorithm	accuracy (%)	iterations (ave.)	cpu time (s)
(14)	PDHG	97.3	110.1	81.7
(14)	ADMM	97.3	385.8	2097
(16)	PDHG	97.2	203.8	161.3
(16)	ADMM	97.2	381.6	2063

MTV for `MNIST`, while Tables 8 and 9 show that the classification accuracy of Potts-RF is larger than that of MTV and multi-class MBO for the other two data sets. In addition, our approach is much easier to be implemented than MTV and multi-class MBO which requires either complicated initialization or computation of the eigenvectors of a large matrix. For the sake of completeness, we also include the classification accuracy computed from the initial probabilities $p_k(x_i)$ in the tables.

Table 8: Average classification accuracy (%) of Potts-RF, MTV and multiclass-MBO on `Three-Circles` for three different numbers of labeled samples.

l	0.83%	1.25%	1.67%
$p_k(x_i)$	75.92 ± 2.43	83.42 ± 2.45	89.91 ± 1.27
Potts-RF(14)	98.19 ± 3.58	99.35 ± 0.07	99.49 ± 0.05
MTV	75.44 ± 8.34	79.19 ± 4.96	79.47 ± 2.03
multiclass-MBO	66.15 ± 5.98	81.13 ± 5.53	90.02 ± 3.31

Table 9: Average classification accuracy (%) of Potts-RF, MTV and multiclass-MBO on `COIL` for three different numbers of labeled samples.

	3.3%	6.7%	10%
$p_k(x_i)$	46.64 ± 1.79	58.90 ± 2.41	66.46 ± 2.31
Potts-RF(14)	81.8 ± 4.9	90.9 ± 2.0	92.9 ± 0.9
MTV	78.4 ± 4.00	89.73 ± 1.5	92.20 ± 1.3
multiclass-MBO	70.53 ± 3.46	82.03 ± 3.90	89.09 ± 2.06

Table 10: Average classification accuracy (%) of Potts-RF, MTV and multiclass-MBO on `MNIST` for three different numbers of labeled samples.

	0.25%	0.5%	1%
$p_k(x_i)$	18.24 ± 3.34	24.67 ± 0.99	35.85 ± 0.77
Potts-RF(14)	97.15 ± 0.13	97.28 ± 0.09	97.32 ± 0.09
MTV	97.62 ± 0.03	97.63 ± 0.03	97.65 ± 0.01
multiclass-MBO	73.0 ± 3.91	90.1 ± 3.24	94.9 ± 2.78

5. Conclusion and Future Direction

We introduce a novel region force function into the Potts model and thus provide a uniformly effective framework for multi-phase image segmentation and semi-supervised data clustering. The new region force function is computed as the negative log-likelihood function of the indicator function under the Bernoulli distribution. The probability that an image pixel or a data point belongs to a given class is estimated based on the mixed Gaussian density model for image segmentation and based on the diffusion distance for semi-supervised data clustering.

Two numerical algorithms PDHG and ADMM are presented to compute the solution of the Potts model. Those two algorithms are developed from two different dual formulations of the Potts model. Extensive numerical experiments have been conducted on benchmark problems in image segmentation and semi-supervised data clustering and show that our approach is as effective as other existing variational methods in the literature.

In this paper, the probabilities used in the computation of the region force function are fixed. For future work, we suggest updating the probabilities adaptively in the numerical algorithms, for example based on the maximum likelihood estimation. We also intend to apply data driven ideas to design new region force functions for different applications.

References

[1] Bae, E., Lellmann, J., and Tai, X.-C. (2013). Convex relaxations for a generalized chan-vese model. In *EMMCVPR 2013, LNCS 8081*, A. Heyden et al., Eds. Springer-Verlag Berlin Heidelberg 2013, 223–236.

[2] Bae, E. and Merkurjev, E. Convex variational methods for multiclass data segmentation on graphs.

[3] Bae, E. and Tai, X.-C. (2015). Efficient global minimization methods for image segmentation models with four regions. *J Math Imaging Vis 51*, 71–97.

[4] Bae, E., Yuan, J., and Tai, X.-C. (2011). Global minimization for continuous multiphase partitioning problems using a dual approach. *International journal of computer vision* **92**, 1, 112–129.

[5] Bonettini, S. and Ruggiero, V. (2012). On the convergence of primal dual hybrid gradient algorithms for total variation image restoration. *Journal of Mathematical Imaging and Vision* **44**, 3 (Jan.), 236–253.

[6] Boykov, Y. and Kolmogorov, V. (2003). Computing geodesics and minimal surfaces via graph cuts. In *ICCV 2003*. 26–33.

[7] Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 1222–1239.

[8] Bresson, X., Laurent, T., Uminsky, D., and von Brecht, J. (2013). Multiclass total variation clustering. In *Advances in Neural Information Processing Systems*. 1421–1429.

[9] Bresson, X., Tai, X.-C., Chan, T. F., and Szlam, A. (2013). Multiclass Transductive Learning Based on L_1 Relaxations of Cheeger Cut and Mumford-Shah-Potts Model. *Journal of Mathematical Imaging and Vision* **49**, 1 (Aug.), 191–201.

[10] Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to image. *J Math Imaging Vis 40*, 120–145.

[11] Chan, T. and Vese, L. (2001a). Active contours without edges. *IEEE Transactions on Image Processing* **10**, 2 (Feb.), 266–277.

[12] Chan, T. F., Esedoglu, S., and Nikolova, M. (2006). Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM J. Appl. Math.* **66**, 5, 1632–1648.

[13] Chan, T. F. and Vese, L. A. (2001b). Active contours without edges. *IEEE Transactions on Image Processing* **10**, 2, 266–277.

[14] Chapelle, O., Schölkopf, B., and Zien, A., Eds. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA. http://www.kyb.tuebingen.mpg.de/ssl-book.

[15] Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 21 (May), 7426–7431.

[16] Dubrovina, A., Rosman, G., , and Kimmel, R. (2015). Multi-region active contours with a single level set function. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 8, 1585–1601.

[17] Ekeland, I. and Teman, R. (1999). *Convex analysis and variational problems*. SIAM.

[18] Esser, E., Zhang, X., and Chan, T. F. (2010). A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Science* **3**, 4, 1015–1046.

[19] Garcia-Cardona, C., Merkurjev, E., Bertozzi, A., Flenner, A., and Percus, A. (2014). Multiclass Data Segmentation Using Diffuse Interface Methods on Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**, 8 (Aug.), 1600–1613.

[20] Gilboa, G. and Osher, S. (2008). Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation* **7**, 3, 1005–1028.

[21] Hu, H., Sunu, J., and Bertozzi, A. L. (2015). Multi-class graph Mumford-Shah model for plume detection using the MBO scheme. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 209–222.

[22] Kass, M., Witkin, A., , and Terzopoulos, D. (1988). Snakes: active contour models. In *IJCV*. Vol. **1**. 321–331.

[23] Kolmogorov, V. and Boykov, Y. (2005). What metrics can be approximated by geo-cuts, or global optimization of length/area and flux. In *ICCV 2005*. 564–571.

[24] Kolmogorov, V. and Zabih, R. (2002). Multi-camera scene reconstruction via graph cuts. In *European Conference on Computer Vision*. 82–96.

[25] Kolmogorov, V. and Zabih, R. (2004a). What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 65–81.

[26] Kolmogorov, V. and Zabih, R. (2004b). What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 65–91.

[27] Lézoray, O., Elmoataz, A., and Ta, V. T. (2012). Nonlocal PDEs on graphs for active contours models with applications to image segmentation and data clustering. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 873–876.

[28] Martin, D. R., Fowlkes, C. C., and Malik, J. (2004). learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 5, 530–549.

[29] Merkurjev, E., Bae, E., Bertozzi, A. L., and Tai, X.-C. (2015). Global binary optimization on graphs for classification of high-dimensional data. *Journal of Mathematical Imaging and Vision* **52**, 3, 414–435.

[30] Merkurjev, E., Kostic, T., and Bertozzi, A. L. (2013). An MBO scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences* **6**, 4, 1903–1930.

[31] Merriman, B., Bence, J. K., and Osher, S. (1992). *Diffusion generated motion by mean curvature*. Department of Mathematics, University of California, Los Angeles.

[32] Muja, M. and Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1) 2*, 331–340.

[33] Mumford, D. and Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics* **XLII**, 5, 577–685.

[34] Nene, S. A., Nayar, S. K., and Murase, H. (1996). Columbia object image library (coil-100). *Technical Report CUCS-006-96*.

[35] Paragios, N., Chen, Y., and Faugeras, O. (2005). *Handbook of Mathematical Models in Computer Vision*. Springer New York.

[36] Potts, R. B. (1952). Some generalized order-disorder transformations. In *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. **48**. 106–109.

[37] Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Kernel methods in computational biology*. MIT press.

[38] Silpa-Anan, C. and Hartley, R. (2008). Optimised KD-trees for fast image descriptor matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[39] Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* **24**, 4, 35–43.

[40] Vedaldi, A. and Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.

[41] Vogiatzis, G., Esteban, C. H., Torr, P. H., and Cipolla, R. (2007). Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI* **29**, 12, 2241–2246.

[42] Wei, K., Tai, X.-C., Chan, T. F., and Leung, S. (2015). Primal-dual method for continuous max-flow approaches. In *Computational Vision and Medical Image Processing V: Proceedings of the 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing (VipIMAGE 2015, Tenerife, Spain, October 19-21, 2015)*. CRC Press, 17.

[43] Wu, Q., Ng, M. K., and Ye, Y. (2013). Markov-miml: A markov chain-based multi-instance multi-label learning algorithm. *Knowledge and information systems* **37**, 1, 83–104.

[44] Wu, Q., Ng, M. K., Ye, Y., Li, X., Shi, R., and Li, Y. (2014). Multi-label collective classification via markov chain based learning method. *Knowledge-Based Systems 63*, 1–14.

[45] Yin, K. and Tai, X.-C. (2017). An Effective Region Force for Some Variational Models for Learning and Clustering. *Journal of Scientific Computing*, 1–22.

[46] Yuan, J., Bae, E., Tai, X.-C., and Boykov, Y. (2010). A continuous max-flow approach to potts model. In *Computer Vision–ECCV 2010*. Springer, 379–392.

[47] Yuan, J., Bae, E., Tai, X.-C., and Boykov, Y. (2014a). A spatially continuous max-flow and min-cut framework for binary labeling problems. *Numerische Mathematik* **126**, 3, 559–587.

[48] Yuan, J., Bae, E., Tai, X.-C., and Boykov, Y. (2014b). A study on continuous max-flow and min-cut approaches. *Numerische Mathemetik* **126**, 3, 559–587.

[49] Zelnik-Manor, L. and Perona, P. (2004). Self-tuning spectral clustering. In *Advances in neural information processing systems*. 1601–1608.

[50] Zhu, M. and Chan, T. (2008). An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 08–34.

Ke Wei,
Department of Mathematics
University of California at Davis
California, USA
E-mail address: kewei@math.ucdavis.edu

Ke Yin,
Center for Mathematical Sciences
Huazhong University of Science and Technology
Wuhan, China
E-mail address: kyin@hust.edu.cn

Xue-Cheng Tai
Department of Mathematics,
University of Bergen, Postboks 7800, 5020, Norway.
E-mail address: tai@math.uib.no

Tony F. Chan
Office of President
Hong Kong University of Science and Technology
Hong Kong, China
E-mail address: tonyfchan@ust.hk