Journal of Computational and Applied Mathematics 343 (2018) 602–618

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Probability-one homotopy methods for constrained clustering

David R. Easterling [a,*], Layne T. Watson [b,c], Naren Ramakrishnan [b]

[a] University of Dayton Research Institute, 300 College Park, Dayton, OH 45469-0101, United States
[b] Department of Computer Science, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061-0106, United States
[c] Department of Mathematics and Aerospace and Ocean Engineering, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061-0106, United States

ARTICLE INFO

Article history:
Received 28 June 2017
Received in revised form 15 April 2018

Keywords:
Constrained clustering
Homotopy algorithm
Multiobjective

ABSTRACT

Many algorithms for constrained clustering have been developed in the literature that aim to balance vector quantization requirements of cluster prototypes against the discrete satisfaction requirements of constraint (must-link or must-not-link) sets. A significant amount of research has been devoted to designing new algorithms for constrained clustering and understanding when constraints help clustering. However, no method exists to systematically characterize solution sets as constraints are gently introduced and how to assist practitioners in choosing a sweet spot between vector quantization and constraint satisfaction. A homotopy method is presented that can smoothly track solutions from unconstrained to constrained formulations of clustering. Beginning the homotopy zero curve tracking where the solution is (fairly) well-understood, the curve can then be tracked into regions where there is only a qualitative understanding of the solution set, finding multiple local solutions along the way. Experiments demonstrate how the new homotopy method helps identify better tradeoffs and reveals insight into the structure of solution sets not obtainable using pointwise exploration of parameters.

1. Introduction

As machine learning permeates multiple fields of science and engineering, new objective functions are continually being proposed to suit the demands of new application domains. Multicriteria objective functions especially are becoming more prevalent in areas such as mixing labeled and unlabeled data [1–3], incorporating constraints [4–6], and transfer learning [7–10]. The difficulty of solving such problems, and even interpreting the solution in a meaningful way, is likewise a growing field of research. A mechanism for solving problems in one field may or may not be adaptable to solving problems in another, and the information yielded by the method may not contain everything needed by the modern researcher.

One such multiobjective formulation is in the area of constrained clustering. In constrained clustering [11], the goal is not just to obtain clusters that are local in their respective spaces but that also obey a discrete set of a priori must-link (ML) and cannot-link or must-not-link (MNL) constraints between points. More complicated constraint sets can be represented in simpler form by these ML and MNL constraints, as well; in particular, the conventional cluster hypothesis may be enforced through the use of ϵ- and δ-constraints [12]. Although there are many powerful constrained clustering algorithms published in the literature [13–18], there is currently a lack of a systematic mathematical theory to guide the design of formulations and understand the tradeoffs that invariably result as each algorithm attempts to serve two masters.

The fundamental problem in algorithm design for constrained clustering problems is the tradeoff between conventional clustering objectives and the requirements of the linking constraints. Broadly speaking, there have been two types of

* Corresponding author.
E-mail addresses: david.easterling@gmail.com, david.easterling@udri.udayton.edu (D.R. Easterling).

https://doi.org/10.1016/j.cam.2018.04.035
0377-0427/

algorithms designed to deal with this problem [19]. The first uses the constraints to learn a distance function. The second strictly enforces the constraints as the algorithm iterates to a useful solution. The primary problem motivating the development of these two algorithmic approaches is that determining the feasibility of a set of constraints that contains MNL and ML constraints both is an NP-complete problem, being equivalent to the graph coloring problem. When the existence of a feasible solution cannot be determined in polynomial time, the usual approach is to fall back on heuristics, with the hope that the resulting solution will be good enough. This dovetails nicely with one of the dominant viewpoints of big-data machine learning, where rigorous solutions are usually impractical due to the sheer amount of data involved, NP-complete or not, and as such heuristic approaches are the norm for reaching reasonable solutions in a reasonable time.

One traditional solution to such heuristically solved biobjective problems is to introduce a parameter λ that balances or weights competing considerations, in this case cluster locality versus constraint satisfaction. Although there are interesting theoretical insights into the complexity of constrained clustering problems [20], there is little existing theory available that can deal with (1) how to efficiently compute solutions parametrically as λ varies, (2) how to find and deal with multiple solutions for a fixed λ, and (3) how to canonically define the best choice of λ. Furthermore, using such λ as an independent variable often poses insurmountable problems to the researcher, as will be shown.

Homotopy methods are systematic approaches to characterize solution sets by smoothly tracking solutions from one formulation to another (in this case, from an unconstrained formulation to a constrained formulation). This can allow the effect of changing λ on the quality and nature of the solutions to be mathematically characterized. Smoothly tracking solutions as λ varies provides a holistic understanding of the interplay between the algorithm and a dataset. The resulting tradeoff curve can yield information about the nature of the problem and the probability of improvement offered by constraints.

Corduneanu and Jaakkola [21] used classical continuation to study how two diverse information sources should be combined in order to arrive at an integrated model. The first application of modern homotopy methods to machine learning was by Ji et al. [22], who showed that a general semisupervised formulation for hidden Markov models (HMMs) can be realized using a probability-one homotopy as well.

The key contributions here are:

(1) The *first* homotopy maps, which combine quadratic loss functions with discrete evaluations of constraint violations, for constrained clustering problems are presented. This is a nontrivial task since there are several discrete aspects to the constrained clustering problem (e.g., discrete assignments of points to clusters, discrete satisfactions or violations of constraints) that need to be accommodated in a traditional homotopy framework.
(2) The construction of homotopy maps typically requires careful problem specific tweaking to ensure convergence. The general map constructed here applies to any constrained clustering problem where a distance function is meaningful, similar to existing algorithms for this purpose.
(3) Use of the theory of nonlinear complementarity problem (NCP) functions and the Kreisselmeier–Steinhauser envelope function is new.
(4) Numerous experimental results demonstrating the scalability, viability, usefulness, superiority, and interpretability of the homotopy map approach to constrained clustering are presented, as well as results for a new map applied to ϵ- and δ-style constraints, which constrain intercluster and intracluster distances.

This paper is organized as follows. Section 2 is the mathematical background, with subsections devoted to specific building blocks to be applied in the next section. Section 3 shows the application of homotopy theory to constrained clustering applications, with a first map dedicated solely to constraint satisfaction and a second map that strives to maintain strong clustering as well, along with proofs of convergence for both maps. Section 4 shows the experimental results, and Section 5 concludes.

2. Mathematical background

Let superscripts denote vector indices and subscripts denote components of vectors and scalar indices unless otherwise indicated. Let all norms be 2-norms unless otherwise indicated and let all distances be Euclidean distances. Let $\mathbb{R}^n$ denote n-dimensional real Euclidean space and let $\mathbb{R}^{n\times m}$ be the set of real $n \times m$ matrices. Let the ith row of a matrix $A \in \mathbb{R}^{n\times m}$ be denoted by $A_{i\cdot}$ and the jth column by $A_{\cdot j}$. Finally, for a vector $x \in \mathbb{R}^n$, $x > 0$ means all $x_i > 0$, $x \geqq 0$ means all $x_i \geqq 0$, and $x \geq 0$ means $x \geqq 0$ but $x \neq 0$.

Given a set $\hat{X} = \{x^i \mid x^i \in \mathbb{R}^d, i = 1, 2, \ldots, k\}$ of k points (cluster representatives) in d dimensions, let $X = \text{vec}\left(x^1, x^2, \ldots, x^k\right) \in \mathbb{R}^{kd}$. Given a set $\hat{Y} = \{y^i \mid y^i \in \mathbb{R}^d, i = 1, 2, \ldots, n\}$ of n data points in d dimensions, let $Y = \text{vec}\left(y^1, y^2, \ldots, y^n\right) \in \mathbb{R}^{nd}$. Represent a constraint by the vector $c = (a, b, z, w) \in \mathbb{R}^{2d+2}$ of two data points $a, b \in \hat{Y}$, an identifier $z = \pm 1$, and a degree-of-belief weight $\mathbb{R} \ni w > 0$, where an identifier of $z = 1$ means that a and b are bound by a must-link constraint (i.e., must be in the same cluster) and an identifier of $z = -1$ means that a and b are bound by a cannot-link constraint (cannot be in the same cluster). Given a set $\hat{C} = \{c^i \mid c^i \in \mathbb{R}^{2d+2}, i = 1, 2, \ldots, q\}$ of q constraints, let $C = \text{vec}\left(c^1, c^2, \ldots, c^q\right) \in \mathbb{R}^{q(2d+2)}$.

2.1. Penalty function and constraints

For a data point $y \in \hat{Y}$ and two cluster prototypes $x^i, x^j \in \hat{X}$ define the comparator function $D : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$D(x^i, x^j, y) = \left(\max\left\{0, \|x^i - y\|^2 - \|x^j - y\|^2\right\}\right)^4.$$

Note that D is three times continuously differentiable, $D \geqq 0$, and $D(x^i, x^j, y) > 0$ if and only if the distance between y and x^i is larger than the distance between y and x^j.

Given $a, b \in \hat{Y}$, let the must-link function

$$F_m : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{kd} \to \mathbb{R}$$

be defined by

$$F_m(a, b, X) = \prod_{i=1}^{k} \left(\sum_{j=1, j\neq i}^{k} D(x^i, x^j, a) + D(x^i, x^j, b) \right)$$

and let the cannot-link function $F_c : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{kd} \to \mathbb{R}$ be defined by

$$F_c(a, b, X) = \sum_{i=1}^{k} \left(\prod_{j=1, j\neq i}^{k} D(x^j, x^i, a) D(x^j, x^i, b) \right).$$

Then the following observations are easily verified.

Observation 1. F_m and F_c are nonnegative and three times continuously differentiable.

Observation 2. For any must-link constraint

$$c = (a, b, 1, w) \in \hat{C},$$

the must-link function $F_m(a, b, X) = 0$ if and only if constraint c is satisfied.

Observation 3. For any cannot-link constraint

$$c = (a, b, -1, w) \in \hat{C},$$

the cannot-link function $F_c(a, b, X) = 0$ if and only if constraint c is satisfied.

Observation 4. The penalty function

$$F(C, X) = \sum_{\{i : z_i = 1\}} w_i F_m(a^i, b^i, X) + \sum_{\{i : z_i = -1\}} w_i F_c(a^i, b^i, X)$$

is zero if and only if all the constraints in $\hat{C}$ are satisfied.

By Observation 4, if it is possible to satisfy all of the constraints with a convex clustering, then there exists a vector of cluster representatives $\mathcal{X}$ such that $F(C, \mathcal{X}) = 0$. This vector of cluster representatives represents a global minimum point of the function F at which $\nabla_X F(C, \mathcal{X}) = 0$. Unfortunately, if multiple $x^i \in \hat{X}$ coalesce, the resulting D values will result in zero even though there is no clear clustering arising from this case, and $\min_X F(C, X)$ has the trivial solutions $x^1 = \cdots = x^k$. Thus it is necessary to constrain the optimization problem $\min_X F(C, X)$ to prevent such a degenerate case from occurring. In addition, X should be bounded, as $\lim_{\|X\| \to \infty} F(C, X) = 0$ is possible.

First, consider the bounding constraint. A straightforward concave function $\Psi : \mathbb{R}^{kd} \to \mathbb{R}$ to achieve bounding is $\Psi(X) = B - \sum_{i=1}^{k} \|x^i\|^2 \geqq 0$, where $B \in \mathbb{R}$ is a given large constant. Second, to prevent the degenerate condition noted above, a set of constraints $g_i : \mathbb{R}^{kd} \to \mathbb{R}$ can be constructed as $g_i(X) = \epsilon_g - \|x^{i_1} - x^{i_2}\|^2 \leqq 0$, where $1 \leqq i \leqq \binom{k}{2}$, $x^{i_1}, x^{i_2} \in \hat{X}$ are different cluster representatives and $\epsilon_g > 0$ is a small constant. Note that these constraints are differentiable everywhere, and satisfy the reverse convex constraint qualification [23] at $\mathcal{X}$ if $\Psi(\mathcal{X}) > 0$ is inactive. If the active constraints at $\mathcal{X}$ satisfy a constraint qualification (e.g., Arrow–Hurwicz–Uzawa [24]), then the resulting optimization problem

$$\begin{aligned} &\min_X F(C, X) \\ &\text{subject to } -\Psi(X) \leqq 0, \\ &\qquad g_i(X) \leqq 0, \quad 1 \leqq i \leqq \binom{k}{2} \end{aligned} \tag{1}$$

satisfies the Karush–Kuhn–Tucker (KKT) necessary conditions at $\mathcal{X}$.

Now that the problem has been defined, it remains to show how homotopy methods can be used to effect a smooth transition from a traditional clustering to a clustering that solves the above optimization problem, yielding a useful tradeoff curve. First, however, several tools necessary to utilize the homotopy method are described.

2.2. Kreisselmeier–Steinhauser function

Since there are $\binom{k}{2}$ separation constraints in the optimization problem, an aggregation function that can reduce these to a single inequality constraint is of benefit. The Kreisselmeier–Steinhauser envelope function [25]

$$Z(X) = \frac{1}{\kappa} \ln \left[\sum_{i=1}^{\binom{k}{2}} \exp(\kappa g_i(X)) \right],$$

where $\kappa > 0$ is a regularization parameter, is a common aggregation function used in optimization to reduce the number of inequality constraints to one. Let $g_{\max}(X) = \max_{1 \leq i \leq \binom{k}{2}} g_i(X)$. Note that

$$g_{\max}(X) \leqq Z(X) \leqq g_{\max}(X) + \frac{\ln(k(k-1)/2)}{\kappa},$$

which means that if $Z(X) \leqq 0$ then $g_i(X) \leqq 0$ for all i. As an approximation function, however, there are some drawbacks. The selection of κ is important; a large κ will result in a small difference between the value of Z and $g_{\max}$, but may also cause some numerical difficulties. Furthermore, the feasible region defined by Z is generally smaller than the feasible region defined by $g_{\max}$, as should be obvious from the above inequalities. While each g_i is concave, Z is neither pseudoconcave nor pseudoconvex. Nevertheless, except for the degenerate case

$$\nabla\Psi(\mathcal{X})(\nabla Z(\mathcal{X}))^T = \|\nabla\Psi(\chi)\| \, \|\nabla Z(\chi)\|$$

the constraints $-\Psi(X) \leqq 0, Z(X) \leqq 0$ satisfy the Arrow–Hurwicz–Uzawa constraint qualification at $\mathcal{X}$. A KKT point $\bar{X}$ for

$$\begin{aligned} &\min_X F(C, X) \\ &\text{subject to } -\Psi(X) \leqq 0, \\ &\qquad\qquad Z(X) \leqq 0 \end{aligned} \tag{2}$$

is generally not a KKT point for (1). However, since ϵ_g is a fairly arbitrary value to separate cluster representatives, the distinction between formulations (1) and (2) is minimal. For a relatively small number of clusters, say $k < 10$, it is possible to select a large κ, say $\kappa = 100$, without encountering numerical difficulties summing the $\binom{k}{2}$ terms in $Z(X)$. Thus, for a moderate number of cluster representatives, $Z(X)$ is a practical way to combine the g_i constraints, reducing the number of dual variables and hence the dimension of the homotopy map.

2.3. Positively oriented nonlinear complementarity functions

A continuous function $\hat{\Psi} : R \times R \longrightarrow R$ is called an *NCP function* if $\hat{\Psi}(a, b) = 0 \iff 0 \leqq a \perp b \geqq 0$, and it is *positively oriented* if $\hat{\Psi}(a, b) \geqq 0 \iff a \geqq 0$ and $b \geqq 0$. The positively oriented NCP function of interest here, first introduced by Mangasarian [26], is

$$\hat{\Phi}(a, b) = -|a - b|^3 + a^3 + b^3.$$

Observe that $\hat{\Phi}$ is C^2; moreover, for $b > 0$, $\hat{\Phi}(\cdot, b)$ is strictly increasing and onto $\mathbb{R}$. NCP functions are used to represent the complementarity conditions within the KKT necessary conditions: for each inequality constraint $g_i \leqq 0$ with associated Lagrange multiplier μ_i, $\hat{\Phi}(-g_i, \mu_i) = 0 \iff -g_i\mu_i = 0, -g_i \geqq 0, \mu_i \geqq 0$. Thus, for an optimization problem (1) containing only inequality constraints, finding a KKT point is equivalent to solving the nonlinear system of equations

$$\begin{aligned} \nabla_X F(C, X) - \mu_0 \nabla\Psi(X) + \sum_{i=1}^{\binom{k}{2}} \mu_i \nabla g_i(X) &= 0, \\ \hat{\Phi}(\Psi, \mu_0) &= 0, \\ \hat{\Phi}(-g_i, \mu_i) = 0, \quad i = 1, \ldots, \binom{k}{2}. & \end{aligned}$$

2.4. Homotopy theory

Standard continuation methods [27,28] find a root $\bar{x}$ of a differentiable function $f(x)$ using a known root x_0 of a simple differentiable function $g(x)$ by solving

$$\rho(\lambda, x) = (1-\lambda)g(x) + \lambda f(x) = 0$$

as λ is increased from 0 to 1, starting with the known solution x_0 at $\lambda = 0$. λ is the continuation parameter, g is called the 'start' function, and f is called the 'target' function. Given a solution (λ, x_λ), standard local methods (such as Newton's method) are used to solve $\rho(\lambda + \delta\lambda, x) = 0$ for fixed small $\delta\lambda > 0$. This yields a series of solutions along a zero curve γ of $\rho(\lambda, x)$. However, there is no guarantee that a given starting function g will yield a zero of f, as the algorithm may fail at some intermediate $\tilde{\lambda}$ as continuation progresses.

Continuation can fail if the zero curve γ of ρ emanating from $(0, x_0)$ fails to exist past some $\tilde{\lambda} < 1$. γ can just stop at $\tilde{\lambda}$, turn back toward $\lambda = 0$ at $\tilde{\lambda}$, or go to infinity. γ may exist past $\tilde{\lambda}$, but bifurcate at $\tilde{\lambda}$, causing the local iteration to fail because $D_x\rho(\lambda, x)$ is singular at the bifurcation point $(\tilde{\lambda}, x_{\tilde{\lambda}})$.

Homotopy methods deal with bifurcation and turning points through a local parametrization of the zero curve $(\lambda, x) = (\lambda(t), x(t))$. Most importantly, homotopy methods treat λ as an independent variable, and do not increase λ monotonically from 0 to 1. The issues of nonexistence, bifurcation, and divergence to infinity are addressed by modern *probability-one homotopy* methods [27–29], which under certain conditions guarantee almost surely (in the probability measure theoretic sense) the existence of a smooth, nonbifurcating, bounded zero curve γ of a homotopy map $\rho_a(\lambda, x)$ that connects a start point $(0, x_0)$ to a point $(1, \bar{x})$, where $f(\bar{x}) = 0$.

These algorithms are implemented in FORTRAN 77 as HOMPACK [30], and extended in Fortran 90 as HOMPACK90 [31]. The following theorems about probability-one homotopy maps and the associated zero curves γ are central.

Theorem 1 (*Parametrized Sard's Theorem*)**.** *Let $U \subset \mathbb{R}^m$, $V \subset \mathbb{R}^n$ be nonempty open sets, $\rho : U \times [0, 1) \times V \to \mathbb{R}^n$ be a C^2 map, and define*

$$\rho_a(\lambda, x) = \rho(a, \lambda, x).$$

*If ρ is transversal to zero (*rank $D\rho = n$ *on $\rho^{-1}(0)$), then for almost all $a \in U$ the map ρ_a is also transversal to zero.*

Theorem 2. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ and $\rho : \mathbb{R}^m \times [0, 1) \times \mathbb{R}^n \to \mathbb{R}^n$ be C^2, and define $\rho_a(\lambda, x) = \rho(a, \lambda, x)$. Assume that*
(1) *ρ is transversal to zero;*
(2) *for each fixed $a \in \mathbb{R}^m$, $\rho_a(0, x) = 0$ has a unique solution x_a at which* rank $D_x\rho_a(0, x_a) = n$*;*
(3) *$\rho_a(1, x) = F(x)$;*
(4) *for each $a \in \mathbb{R}^m$, the connected component of the zero set $\rho_a^{-1}(0)$ containing $(0, x_a)$ is bounded.*
Then for almost all $a \in \mathbb{R}^m$ there exists a zero curve γ of $\rho_a(\lambda, x)$, emanating from $(0, x_a)$, along which the $n \times (n+1)$ Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, that does not intersect itself and is disjoint from any other zeros of ρ_a, and accumulates at a point $(1, \bar{x})$ for which $F(\bar{x}) = 0$. Furthermore, if rank $D\rho_a(1, \bar{x}) = n$*, then the curve γ connecting $(0, x_a)$ to $(1, \bar{x})$ has finite arc length.*

Theorem 3. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be C^2, and suppose there exist r_0, $r > 0$, $r > r_0$, such that for any $a \in \mathbb{R}^n$ with $\|a\|_2 < r_0$, $x - a$ and $F(x)$ do not point in opposite directions on $\left\{x \in \mathbb{R}^n \mid \|x\|_2 = r\right\}$. Define $\rho : \mathbb{R}^n \times [0, 1) \times \mathbb{R}^n \to \mathbb{R}^n$ by*

$$\rho(a, \lambda, x) = (1-\lambda)(x - a) + \lambda F(x),$$

and let $\rho_a(\lambda, x) = \rho(a, \lambda, x)$. Then for almost all vectors $a \in \mathbb{R}^n$ with $\|a\|_2 < r_0$ there exists a zero curve γ of $\rho_a(\lambda, x)$, emanating from $(0, a)$, along which the $n \times (n+1)$ Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, that does not intersect itself and is disjoint from any other zeros of ρ_a, and accumulates at a point $(1, \bar{x})$ for which $F(\bar{x}) = 0$. Furthermore, if rank $D\rho_a(1, \bar{x}) = n$*, then the curve γ connecting $(0, a)$ to $(1, \bar{x})$ has finite arc length.*

Theorem 1 means that the set of points (λ, x) where $\rho_a(\lambda, x) = 0$ looks like the curves in Fig. 1 for almost all points $a \in U$. The hypotheses in Theorems 2 and 3 guarantee that the curve γ in Fig. 1 is the only curve emanating from $\lambda = 0$ and that γ must accumulate at $\lambda = 1$. Thus, a probability-one homotopy algorithm simply tracks the zero curve γ_a of ρ_a, which is guaranteed to reach a solution $\bar{x}$ of $F(x) = 0$ at $\lambda = 1$, with probability one (almost surely) so long as the hypotheses of Theorems 2 and 3 are met. (Theorem 3 is simply a special case of Theorem 2.)

In practice, the full rank of the Jacobian matrix $D_{\rho_a}(1, \bar{x})$ is not necessary, as the zero curve usually approaches a solution as $\lambda \to 1$ with finite arc length. This is especially true when applied to the semisupervised clustering problem, as the desired clustering (which satisfies the given constraints) is expected to be present at some point along γ before $\lambda = 1$. Homotopy maps that fulfill the theorems' assumptions are called *globally convergent probability-one homotopy maps*. Proving a map to be globally convergent with probability one reduces to proving that it meets the given assumptions. Given time to trace the finite arc length of the solution curve with a robust enough tracker, such curves will inevitably yield a useful solution. More valuable, tracing the curve yields the entire parametrized solution trace for analysis, generating a tradeoff curve that can be further analyzed.

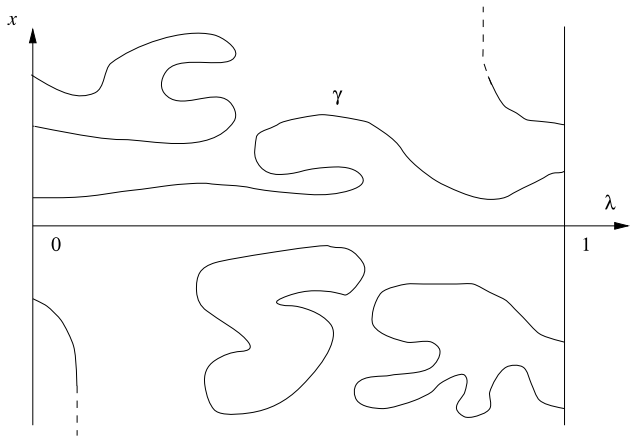

Fig. 1. The inverse image $\rho_a^{-1}(0)$ for ρ_a transversal to zero.

It is possible to modify the given "natural" homotopy map $(1-\lambda)g(x)+\lambda f(x)$ by manipulating the λ parameter to yield more useful maps. In particular, where $g(x)$ lacks a unique zero (such as the clustering problem case here), a unique zero x_0 at $\lambda = 0$ can be enforced by modifying the map to

$$\begin{aligned}\rho_a(\lambda, x) = &(1-\tanh(60\lambda))(x-x_0)\\ &+\tanh(60\lambda)\left[(1-\lambda)g(x)+\lambda f(x)\right],\end{aligned}$$

where tanh is the hyperbolic tangent function. $\tanh(60\lambda) \approx 1$ for $\lambda > 0.3$, to within 64-bit machine accuracy. Thus $\rho_a(\lambda, x) = 0$ has a unique solution $x = x_0$ at $\lambda = 0$, but for $\lambda > 0.3$ the map looks essentially like $(1-\lambda)g(x)+\lambda f(x)$. Semisupervised learning problems often have such "natural" start functions g with multiple zeros, making this a useful general trick in homotopy map generation. (The rigorous convergence theory for this map actually uses $\tanh\big(60\lambda/(1-\lambda)\big)$, which is computationally indistinguishable from $\tanh(60\lambda)$ for $\lambda > 0.3$.)

3. Clustering application

Let $k^0 \in \mathbb{R}^{kd}$ be some (presumably poor) solution to the unsupervised clustering problem for k clusters in d dimensions, generated by a traditional clustering approach, such as the K-Means algorithm. For the present discussion, consider each cluster assignment to be a "hard" assignment, that is, each data point is assigned to a single cluster determined by its distances from the cluster representatives, not assigned a probability of belonging to each cluster based on those distances.

It is worth noting at this juncture that disjunctive and conjunctive combinations of constraints can be represented by the penalty functions F_m and F_c defined in Section 2, which are of particular value when ϵ- and δ-constraints are considered. ϵ- and δ-constraints are constraints that act upon groups of instances. ϵ-constraints are constraints that dictate that any data point in a cluster must have another data point in that cluster within ϵ distance, or be the only data point in the cluster. δ-constraints are constraints that dictate that any datapoint in a cluster must be at least δ distance from every datapoint that resides in a different cluster. Both of these types of constraints can be represented as disjunctions and conjunctions of must-link constraints [12].

Let C^1 and C^2 be constraints (must-link, cannot-link, or combinatorial) and let F^1 and F^2 be the corresponding penalty functions. Then $C^3 = C^1 \vee C^2$ has the corresponding penalty function $F^3 = F^1F^2$. Similarly, $C^4 = C^1 \wedge C^2$ has the corresponding penalty function $F^4 = F^1 + F^2$. Observe that $F^3 = 0$ if and only if C^3 is satisfied, and $F^4 = 0$ if and only if C^4 is satisfied. Finally, observe that any number of must-link and cannot-link constraints can thus be combined in conjunctive normal form by summing products of these penalty functions. As such, these penalty functions can easily be adapted to represent penalty functions for ϵ- and δ-constraints.

By Observation 4 in Section 2, if it is possible to satisfy all of the constraints defined by C, then there exists a vector of cluster representatives $\mathcal{X}$ such that the penalty function $F(C, \mathcal{X}) = 0$. This vector of cluster representatives represents a global minimum point of the function F at which $\nabla_X F(C, \mathcal{X}) = 0$. This suggests the homotopy map (where $a = k^0$)

$$\breve{\rho}_a(\lambda, X) = (1-\lambda)(X-a) + \lambda\left(\nabla_X F(C, X)\right)^T.$$

This homotopy map is appealing: when $\lambda = 0$, the solution is simply the solution k^0 to the unsupervised clustering problem. When $\lambda = 1$, the solution, if one exists, represents a local minimum point (or stationary point) of the penalty function, which is based on the violation of constraints. This is not to say that the solution generated will satisfy all the constraints if such a solution is possible, as it is fairly easy to construct a degenerate set of constraints so that there is a local solution close to $X = k^0$. However, in practice this has not proven to be a problem.

$\check{\rho}_a$ is a probability-one homotopy map, but while it satisfies conditions (1), (2), and (3) in Theorem 2, it fails to satisfy condition (4), bounded γ. Furthermore, there is a trivial solution to all constraints at $x^1 = x^2 = \cdots = x^k$, where all cluster representatives are equal. Thus, modifications must be made to the above map to accommodate the constraints outlined earlier in Section 2.

3.1. First homotopy map

First, consider the bounding constraint

$$\Psi(X) = B - \sum_{i=1}^{n} \|x^i\|^2 \geqq 0,$$

and let $B > \|k^0\|^2$ be a given constant. The Lagrangian of the new bounded penalty function is $\hat{L}(X, \mu) = F(C, X) - \mu\Psi(X)$, and its derivative, replacing $\nabla_X F(C, X)$, is

$$\nabla_X \hat{L}(X, \mu) = \nabla_X F(C, X) - \mu \nabla_X \Psi(X).$$

This yields a new variable, the Lagrangian multiplier μ, which in turn adds a new function to the map (since the map must be from $\mathbb{R}^{p+1} \to \mathbb{R}^p$ for some p), along with the requirement that $\mu \geqq 0$, $\Psi \geqq 0$, and $\mu\Psi = 0$. This naturally leads to the use of the Mangasarian NCP function presented in [26] and modified in [32,33].

Define the function $\Phi : [0, 1] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} \Phi(\lambda, \mu, \Psi(X), h_0) = & \\ & - |\mu - \Psi(X)|^3 + \mu^3 \\ & + \Psi(X)^3 - (1 - \lambda)h_0 \end{aligned}$$

for some constant $h_0 > 0$. The constant term h_0 is designed to force the remaining terms to remain positive for $\lambda < 1$, which enforces the bounding of X for $\lambda < 1$, since $\Psi(X)$ must remain positive when $\Phi = 0$. When $\lambda = 1$, $\Phi(1, \mu, \Psi, h_0) = 0 \iff \mu \geqq 0, \Psi \geqq 0, \mu\Psi = 0$. The previous homotopy map $\check{\rho}_a$ is then modified to (where now $a = (k^0, h_0)$)

$$\hat{\rho}_a(\lambda, X, \mu) = \begin{pmatrix} (1 - \lambda)(X - k^0) + \lambda(\nabla_X \hat{L}(X, \mu))^T \\ \Phi(\lambda, \mu, \Psi(X), h_0) \end{pmatrix}.$$

Note that $\hat{\rho}_a(1, X, \mu) = 0$ is equivalent to the KKT conditions: $\nabla_X \hat{L} = 0$, $\Psi \geqq 0$, $\mu \geqq 0$, $\mu\Psi = 0$. Unfortunately, while the stated Φ enforces a lower bound on μ, it does not enforce an upper bound on μ; in fact, if $\Psi(X) \to 0$ as $\lambda \to \tilde{\lambda} < 1$, μ must potentially become arbitrarily large to compensate for this. While this map is better than the previous one in that it prevents cluster representatives from migrating arbitrarily far from the data set, it does not prevent μ from growing arbitrarily large, although in practice this is not a common occurrence.

To avoid the trivial solution $x^1 = \cdots = x^k$, an alternative to the constraint functions $g_i(X)$ and $Z(X)$ described in Section 2 is to define the function $G : \mathbb{R}^{kd} \to \mathbb{R}$ by

$$G(X) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \max(0, \ell - \|x^i - x^j\|^2)^4, \quad x^i, x^j \in \hat{X}.$$

Then $G \geqq 0$, $G \in C^3$, and $G = 0$ unless two cluster representatives x^i and x^j are less than a distance $\sqrt{\ell}$ from each other, where $\ell > 0$ is a given regularization constant. This represents an equality constraint on the original problem. The updated Lagrangian becomes $\bar{L}(X, \mu, \nu) = F(C, X) - \mu\Psi(X) + \nu G(X)$. In turn, $\nabla_X \bar{L}(X, \mu, \nu) = \nabla_X F(C, X) - \mu \nabla_X \Psi(X) + \nu \nabla_X G(X)$. The additional function is much simpler here, as $G(X)$ is obviously bounded above by $k(k-1)\ell^4/2$ and below by 0. Let $G(X)$ serve as the final regularization function when $\lambda = 1$, thus fulfilling the equality constraint, and let ν be uniquely determined at $\lambda = 0$ by some initial $\mathbb{R} \ni \nu_0 > 0$. Since it can be assumed that $G(k^0) = 0$ for any reasonable ℓ, and $\Psi(k^0) > 0$, the final hard clustering map is (where now $a = (k^0, h_0, \nu_0)$)

$$\bar{\rho}_a(\lambda, X, \mu, \nu) = \begin{pmatrix} (1 - \lambda)(X - k^0) + \lambda(\nabla_X \bar{L}(X, \mu, \nu))^T \\ \Phi(\lambda, \mu, \Psi(X), h_0) \\ (1 - \lambda)(\nu - \nu_0) + G(X) \end{pmatrix}.$$

This map is also a probability-one homotopy map. Taking $a = \left(k^0, h_0, \nu_0\right)$ the map $\bar{\rho}(a, \lambda, X, \mu, \nu) = \bar{\rho}_a(\lambda, X, \mu, \nu)$ is transversal to zero $- D_a\bar{\rho} = (\lambda - 1)I$, a multiple of the identity matrix, hence $D\bar{\rho}$ has full rank. For $0 \leqq \lambda \leqq 1$ and $\Psi(k^0) > 0$, Φ is a strictly increasing function of μ (when the other variables are held constant), unbounded above, and therefore $\Phi(0, \mu, \Psi(k^0), h_0) = 0$ uniquely determines μ. Thus $\bar{\rho}_a = 0$ has a unique solution at $\lambda = 0$, and a straightforward calculation shows that $D_{(X,\mu,\nu)}\bar{\rho}_a(0, X, \mu, \nu)$ is invertible at this solution. It is also clear from the construction of $\bar{\rho}_a$ that $\bar{\rho}_a(1, X, \mu, \nu) = 0$ is equivalent to the KKT conditions for the problem of minimizing $F(X, C)$ subject to the bounding

constraint $\Psi \geqq 0$ and the regularization constraint $G = 0$. Therefore, $\bar{\rho}_a$ satisfies conditions (1), (2), and (3) of Theorem 2, but the bounded γ condition (4) is not satisfied without further assumptions. Conditions for the zero curve γ being bounded (and hence reaching a solution at $\lambda = 1$) are addressed in the next lemma.

Lemma 1. *Let $\Psi(k^0) > 0$, $G(k^0) = 0$, γ be a zero curve of $\bar{\rho}_a(\lambda, X, \mu, \nu)$ emanating from $\left(0, k^0, \mu_0, \nu_0\right)$ along which $D\bar{\rho}_a$ has full rank, and assume that ν is bounded along γ. Then γ itself is bounded.*

Proof. $\Phi(\lambda, \mu, \Psi(X), h_0) = 0$ along γ implies that, for $\lambda < 1$, $\mu > 0$ and $\Psi(X) > 0$, which in turn implies that X is bounded along γ. Since $0 \leqq \lambda \leqq 1$, and ν is assumed to also be bounded along γ, it suffices to prove that μ is bounded along γ. Assume otherwise, so there exists a sequence of points $\left(\lambda_i, X^i, \mu_i, \nu_i\right)$ on γ with $\mu_i \geqq 0$, $\mu_i \to \infty$. Passing to a subsequence if necessary, it may be assumed (by compactness) that $\left(\lambda_i, X^i, \nu_i\right) \to (\bar{\lambda}, \bar{X}, \bar{\nu})$.

Suppose $\Psi(\bar{X}) > 0$. Then because Φ is a strictly increasing function of μ and unbounded above [26,32,33] $\Phi\left(\bar{\lambda}, \mu_i, \Psi(\bar{X}), h_0\right) \to 0$ implies $\{\mu_i\}$ is bounded, a contradiction. Hence $\Psi(\bar{X}) = 0$.

Suppose then that $\Psi(\bar{X}) = B - \|\bar{X}\|^2 = 0$ and $\bar{\lambda} > 0$. In this case $\nabla\Psi(\bar{X}) = -2\bar{X} \neq 0$ and (from the first component of $\bar{\rho}_a = 0$)

$$\mu_i \nabla\Psi(\bar{X}) \to \frac{1-\bar{\lambda}}{\bar{\lambda}}\left(\bar{X} - k^0\right)^T + \nabla_X F(C, \bar{X}) + \bar{\nu}\nabla G(\bar{X})$$

$\Longrightarrow \{\mu_i\}$ is bounded, a contradiction.

The remaining case is $\Psi(\bar{X}) = 0$ and

$$\bar{\lambda} = 0 \Longrightarrow \nabla\Psi(\bar{X}) \neq 0$$

and

$$\lambda_i\mu_i\left(\nabla\Psi(\bar{X})\right)^T \to \bar{X} - k^0 \Longrightarrow w\left(-\nabla\Psi(\bar{X})\right)^T = k^0 - \bar{X}$$

for some $w \geqq 0$. Since $-\Psi(X)$ is convex, $-\nabla\Psi(\bar{X})(k^0 - \bar{X}) \leqq -\Psi(k^0) - (-\Psi(\bar{X})) < 0$. Then $0 \geqq (k^0 - \bar{X})^T\left(-\nabla\Psi(\bar{X})\right)^T w = (k^0 - \bar{X})^T(k^0 - \bar{X}) > 0$, a contradiction. Therefore μ is bounded along γ. •

Lemma 1 directly yields the next homotopy convergence theorem.

Theorem 4. *Using the notation of this section, define $\bar{\rho} : \mathbb{R}^{kd} \times (0, \infty) \times (0, \infty) \times [0, 1) \times \mathbb{R}^{kd} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{kd+2}$ by*

$$\bar{\rho}\left(k^0, h_0, \nu_0, \lambda, X, \mu, \nu\right) = \begin{pmatrix} (1-\lambda)(X - k^0) + \lambda(\nabla_X \bar{L}(X, \mu, \nu))^T \\ \Phi\left(\lambda, \mu, \Psi(X), h_0\right) \\ (1-\lambda)(\nu - \nu_0) + G(X) \end{pmatrix}.$$

Let $\Psi(k^0) > 0$, $G(k^0) = 0$, $a = \left(k^0, h_0, \nu_0\right)$ and

$$\bar{\rho}_a(\lambda, X, \mu, \nu) = \bar{\rho}\left(k^0, h_0, \nu_0, \lambda, X, \mu, \nu\right).$$

Then $\bar{\rho}$ is transversal to zero, and for almost all $a \in \mathbb{R}^{kd} \times (0, \infty) \times (0, \infty)$ there exists a zero curve γ of $\bar{\rho}_a$, emanating from $\left(0, k^0, \mu_0, \nu_0\right)$, along which the $(kd+2) \times (kd+3)$ Jacobian matrix $D\bar{\rho}_a$ has full rank, that does not intersect itself and is disjoint from any other zeros of $\bar{\rho}_a$. If ν is bounded along γ, then γ accumulates at a point $(1, \bar{X}, \bar{\mu}, \bar{\nu})$, where $(\bar{X}, \bar{\mu}, \bar{\nu})$ is a KKT point for the constrained clustering problem

$$\min_X F(C, X) \quad \text{subject to} \quad -\Psi(X) \leqq 0, \quad G(X) = 0.$$

Furthermore, if $\operatorname{rank} D\bar{\rho}_a(1, \bar{X}, \bar{\mu}, \bar{\nu}) = kd + 2$, *then the curve γ connecting $\left(0, k^0, \mu_0, \nu_0\right)$ to $(1, \bar{X}, \bar{\mu}, \bar{\nu})$ has finite arc length.*

3.2. K-means approximation

In order for the tradeoff curve to reflect an accurate picture of the differences between clustering based solely on the cluster hypothesis and clustering based on the satisfaction of cluster constraints, it is important that the start function $g(x)$ accurately represents the state of the clustering as determined by the cluster hypothesis and that $f(x)$ accurately represents the state of the clustering as determined by the clustering constraints. The latter case is handled by the optimization problem given above, but the former case requires a clustering formulation that will work in the context of a homotopy map.

The traditional K-Means clustering algorithm is the most popular clustering algorithm based on the cluster hypothesis available. However, the K-Means function $K : \mathbb{R}^{kd} \to \mathbb{R}$ to be minimized,

$$K(X) = \sum_{i=1}^{k} \sum_{y \in S_i} \|y - x^i\|^2,$$

where S_i is the set of all points in cluster i, with representative x^i, is not C^2, since cluster assignment is not differentiable. Ideally, besides being a fair approximation of the K-Means clustering algorithm, the approximation $\hat{K}(X) : \mathbb{R}^{kd} \to \mathbb{R}$ will have two additional qualities: It must be C^3, and $\nabla_X \hat{K}(X) : \mathbb{R}^{kd} \to \mathbb{R}^{kd}$ must be bounded in the feasible region.

The most common such approximation [34] for a given data set $\hat{Y}$ is

$$\hat{K}(X) = \sum_{i=1}^{|\hat{Y}|} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|y^i - x^j\|^2}}.$$

This continuously differentiable approximation arises from posing the original K-Means clustering problem as a sum of products of the weight or probability $P_{i,j}(X)$ that a data point $y^i \in \hat{Y}$ belongs to a particular cluster represented by $x^j \in \hat{X}$, measured here as

$$P_{i,j} = \frac{\frac{1}{\|y^i - x^j\|^2}}{\sum_{m=1}^{k} \frac{1}{\|y^i - x^m\|^2}},$$

and a measure $\tilde{D}_{i,j}$ of the distance that the data point resides from the cluster representative, taken here as $\tilde{D}_{i,j} = \|y^i - x^j\|^2$, a simple square of the Euclidean distance from x^j to y^i. Thus, points very close to their cluster representatives have high probability of belonging to that cluster, but bring a corresponding low value to the final optimization function, since such points are considered ideal. Summing the products yields the approximation $\hat{K}(X) = \sum_{i=1}^{|\hat{Y}|} \sum_{j=1}^{k} P_{i,j} \tilde{D}_{i,j}$ after cancellation. Note, however, that this cancellation may yield overflow if $\|y^i - x^j\| \approx 0$, in which case the summand for that index i is taken as zero. These singularities are removable, and $\hat{K}(X)$ is an entire function (in each of the components of X, viewed as a complex vector). Minimizing $\hat{K}(X)$ thus yields an approximation of the minimum of the original K-Means function. Furthermore, $\nabla_X \hat{K}(X)$ is bounded if X is bounded, and the components of $\nabla_X \hat{K}(X)$ are also entire functions in each of the components of (complex) X. Precisely,

$$\frac{\partial \hat{K}(X)}{\partial x_c^b} = \sum_{i=1}^{|\hat{Y}|} \frac{2k(x_c^b - y_c^i)}{\|y^i - x^b\|^4 \left(\sum_{j=1}^{k} \frac{1}{\|y^i - x^j\|^2}\right)^2}$$

is bounded if X is bounded, with the same removable singularities as $\hat{K}(X)$.

Note that while the problem has been represented here using the squared 2-norm as the measure of distance, points can be "spread out" by using higher order (even) p-norms raised to the pth power. In fact, any C^3 nonnegative function could be used as the distance measure between data points; the squared 2-norm is simply the most convenient one for the present purpose.

3.3. Second homotopy map

Generally inequality constraints are easier to deal with than equality constraints, so consider replacing the equality constraint $G(X) = 0$ used for the first homotopy map $\bar{\rho}_a$ by the inequality constraint $Z(X) \leqq 0$ discussed earlier. Keep the same bounding constraint $\Psi(X) \geqq 0$. Using the same modified NCP function Φ as before, the equation

$$\Phi(\lambda, \mu, \Psi(X), h_0) = 0$$

for $h_0 > 0$, $\Psi(k^0) > 0$, $\Phi(0, \mu_0, \Psi(k^0), h_0) = 0$, and $0 \leqq \lambda < 1$ forces $\Psi(X) > 0$ along the zero curve γ. Similarly the equation

$$\Phi(\lambda, \nu, -Z(X), h_1) = 0$$

for $h_1 > 0$, $Z(k^0) < 0$, $\Phi(0, \nu_0, -Z(k^0), h_1) = 0$, and $0 \leqq \lambda < 1$ forces $Z(X) < 0$ along γ. When $\lambda = 1$, these two equations enforce the KKT conditions for the constraints $-\Psi(X) \leqq 0, Z(X) \leqq 0$ and their Lagrange multipliers μ, ν, respectively.

The Lagrangian function associated with (2) is

$$\tilde{L}(X, \mu, \nu) = F(C, X) - \mu\Psi(X) + \nu Z(X),$$

and a KKT point $(\bar{X}, \bar{\mu}, \bar{\nu})$ for (2) satisfies

$$\begin{aligned} \nabla_X \tilde{L}(X, \mu, \nu) &= 0, \\ 0 \leqq \mu \perp \Psi(X) &\geqq 0, \\ 0 \leqq \nu \perp -Z(X) &\geqq 0. \end{aligned}$$

Furthermore, should $Z(\bar{X}) < 0$, the KKT point $(\bar{X}, \bar{\mu}, 0)$ for (2) yields a KKT point $(\bar{X}, \bar{\mu}, 0, \ldots, 0)$ for (1).

Finally, putting all the pieces together, the proposed constrained clustering homotopy map is

$$\tilde{\rho}_a(\lambda, X, \mu, \nu) = \begin{pmatrix} (1-\tanh(60\lambda))(X-k^0) + \tanh(60\lambda)\varphi(\lambda, X, \mu, \nu) \\ \Phi\,(\lambda, \mu, \Psi(X), h_0) \\ \Phi\,(\lambda, \nu, -Z(X), h_1) \end{pmatrix},$$

where

$$\varphi(\lambda, X, \mu, \nu) = \left((1-\lambda)\nabla_X \hat{K}(X) + \lambda \nabla_X \tilde{L}(X, \mu, \nu)\right)^T,$$

$a = (k^0, h_0, h_1)$ and k^0 is any point for which $\Psi(k^0) > 0$, $Z(k^0) < 0$, and $\nabla_X \hat{K}(k^0) \approx 0$, e.g., a K-Means solution (locally) minimizing $K(X)$.

The $X - k^0$ term in the above construction arises because $\nabla_X \hat{K}(X) = 0$ has multiple possible solutions. At the very least, permutations of the cluster representatives in X will yield identical values for $\nabla_X \hat{K}$. The $(X - k^0)$ term ensures that $\tilde{\rho}_a(0, X, \mu, \nu) = 0$ has a unique solution as required by Theorem 2. Since $h_0 > 0$, $\Phi(0, \mu, \Psi(k^0), h_0) = 0$ uniquely determines $\mu = \mu_0 > 0$, and similarly $h_1 > 0$, $\Phi(0, \nu, -Z(k^0), h_1) = 0$ uniquely determines $\nu = \nu_0 > 0$. The tanh terms exist to prevent the erroneous computation of singular Jacobians close to $\lambda = 0$.

Computationally, as mentioned earlier, $\tanh(60\lambda) = 1$ in 64-bit arithmetic for $\lambda > 0.3$, and thus, for $\lambda > 0.3$, this map functions identically to

$$\begin{pmatrix} \left((1-\lambda)\nabla_X \hat{K}(X) + \lambda \nabla_X \tilde{L}(X, \mu, \nu)\right)^T \\ \Phi\,(\lambda, \mu, \Psi(X), h_0) \\ \Phi\,(\lambda, \nu, -Z(X), h_1) \end{pmatrix}.$$

3.4. Convergence proof

(1) Taking $a = \left(k^0, h_0, h_1\right)$ the map $\tilde{\rho}(a, \lambda, X, \mu, \nu) = \tilde{\rho}_a(\lambda, X, \mu, \nu)$ is transversal to zero: Observe that

$$D_a\tilde{\rho}(a, \lambda, X, \mu, \nu) = \operatorname{diag}\left(-(1-\tanh(60\lambda))I, -(1-\lambda), -(1-\lambda)\right),$$

which has rank $kd + 2$ for $0 \leqq \lambda < 1$.
(2) $\tilde{\rho}_a = 0$ has a unique solution at $\lambda = 0$: For $0 \leqq \lambda \leqq 1$ and $\Psi(k^0) > 0$, $\Phi(\lambda, \mu, \Psi(k^0), h_0)$ is a strictly increasing function of μ, unbounded above, and therefore $\Phi_0(0, \mu, k^0, h_0) = 0$ uniquely determines $\mu = \mu_0$. Similarly, for $0 \leqq \lambda \leqq 1$ and $Z(k^0) < 0$, $\Phi(\lambda, \nu, -Z(k^0), h_1)$ is a strictly increasing function of ν, unbounded above, and therefore $\Phi(0, \nu, -Z(k^0), h_1) = 0$ uniquely determines $\nu = \nu_0$. A straightforward calculation shows that $D_{(X,\mu,\nu)}\tilde{\rho}_a(0, k^0, \mu_0, \nu_0)$ is invertible.
(3) It is clear from the construction of $\tilde{\rho}_a$ that

$$\tilde{\rho}_a(1, X, \mu, \nu) = 0$$

is equivalent to the KKT necessary conditions for the problem (2).

Therefore, $\tilde{\rho}_a$ satisfies conditions (1), (2), and (3) of Theorem 2, but the bounded γ condition (4) is not satisfied without further assumptions. Conditions for the zero curve γ being bounded (and hence reaching a solution at $\lambda = 1$) are addressed in the next lemma.

Lemma 2. *Let $\Psi(k^0) > 0$, $Z(k^0) < 0$, γ be a zero curve of $\tilde{\rho}_a(\lambda, X, \mu, \nu)$ emanating from $\left(0, k^0, \mu_0, \nu_0\right)$ along which $D\tilde{\rho}_a$ has full rank, and assume that ν is bounded along γ. Then γ itself is bounded for $0 \leqq \lambda \leqq 1$.*

Proof. For $0 \leqq \lambda < 1$, $\Phi(\lambda, \mu, \Psi(X), h_0) = 0$ along γ implies that $\mu > 0$ and $\Psi(X) > 0$, which in turn implies that X is bounded along γ. Since $0 \leqq \lambda \leqq 1$, and ν is assumed to also be bounded along γ, it suffices to prove that μ is bounded along γ. Assume otherwise, so there exists a sequence of points $\left(\lambda_i, X^i, \mu_i, \nu_i\right)$ on γ with $\mu_i \geqq 0$, $\mu_i \to \infty$. Passing to a subsequence if necessary, it may be assumed (by compactness) that $\left(\lambda_i, X^i, \nu_i\right) \to (\bar{\lambda}, \bar{X}, \bar{\nu})$.

Suppose $\Psi(\bar{X}) > 0$. Then because Φ is strictly increasing in μ and unbounded above [26,32,33] $\Phi\left(\bar{\lambda}, \mu_i, \Psi(\bar{X}), h_0\right) \to 0$ implies $\{\mu_i\}$ is bounded, a contradiction. Hence $\Psi(\bar{X}) = 0$.

Suppose then that $\Psi(\bar{X}) = B - \|\bar{X}\|^2 = 0$ and $\bar{\lambda} > 0$. In this case $\nabla_X \Psi(\bar{X}) = -2\bar{X}^T \neq 0$ and (from the first component of $\tilde{\rho}_a = 0$) $\mu_i \nabla_X \Psi(\bar{X}) = -2\mu_i \bar{X}^T \to \frac{1-\tanh(60\bar{\lambda})}{\bar{\lambda}\tanh(60\bar{\lambda})}\left(\bar{X} - k^0\right)^T + \frac{1-\bar{\lambda}}{\bar{\lambda}}\nabla_X \hat{K}(\bar{X}) + \nabla_X F(C, \bar{X}) + \bar{\nu}\nabla_X Z(\bar{X}) \Longrightarrow \{\mu_i\}$ is bounded, a contradiction.

The remaining case is $\Psi(\bar{X}) = 0$ and $\bar{\lambda} = 0 \Longrightarrow \nabla_X \Psi(\bar{X}) = -2\bar{X}^T \neq 0$ and

$$\tanh(60\lambda_i)\lambda_i\mu_i\left(\nabla_X \Psi(\bar{X})\right)^T \to \bar{X} - k^0$$

$\implies w\left(-\nabla_X \Psi(\bar{X})\right)^T = k^0 - \bar{X} \neq 0$ for some $w > 0$, since $0 = \Psi(\bar{X}) \neq \Psi(k^0) > 0$. Since $-\Psi(X)$ is convex, $-\nabla_X \Psi(\bar{X})(k^0 - \bar{X}) \leqq -\Psi(k^0) - (-\Psi(\bar{X})) < 0$. Then $0 > (k^0 - \bar{X})^T \left(-\nabla_X \Psi(\bar{X})\right)^T w = (k^0 - \bar{X})^T (k^0 - \bar{X}) > 0$, a contradiction. Therefore μ is bounded along γ for $0 \leqq \lambda \leqq 1$. •

Note that ν and μ can both go to infinity as $\lambda \to 1$ and the h_0 and h_1 terms vanish from $\Phi(\lambda, \mu, \Psi(X), h_0)$ and $\Phi(\lambda, \nu, -Z(X), h_1)$, which corresponds to two or more mean prototypes approaching each other as both approach the boundary of the region. This indicates multiple active constraints. The solution $\bar{X}$ approached as $\lambda \to 1$ will still yield a KKT point for (2), although the Lagrange multipliers will not be available (however, they may be easily verified to be greater than zero in such a case).

Note also that ν may be unbounded at some point before $\lambda = 1$, causing the method to fail. However, in practice, this deficiency is rare enough that it has yet to be reproduced in any non-degenerate test case, making the assumption in Lemma 2 reasonable.

Lemma 2 and the earlier discussion of $\tilde{\rho}$ directly yield the next homotopy convergence theorem.

Theorem 5. *Using the notation of this section, define* $\tilde{\rho} : \mathbb{R}^{kd} \times (0, \infty) \times (0, \infty) \times [0, 1) \times \mathbb{R}^{kd} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{kd+2}$ *by*

$$\tilde{\rho}\left(k^0, h_0, h_1, \lambda, X, \mu, \nu\right) = \begin{pmatrix} (1 - \tanh(60\lambda))(X - k^0) + \tanh(60\lambda)\varphi(\lambda, X, \mu, \nu) \\ \Phi\,(\lambda, \mu, \Psi(X), h_0) \\ \Phi\,(\lambda, \nu, -Z(X), h_1) \end{pmatrix}$$

where

$$\varphi(\lambda, X, \mu, \nu) = \left((1 - \lambda)\nabla_X K(X) + \lambda \nabla_X \tilde{L}(X, \mu, \nu)\right)^T.$$

Let $\Psi(k^0) > 0, Z(k^0) < 0, a = \left(k^0, h_0, h_1\right)$, *and*

$$\tilde{\rho}_a(\lambda, X, \mu, \nu) = \tilde{\rho}\left(k^0, h_0, h_1, \lambda, X, \mu, \nu\right).$$

Then for almost all $a \in \mathbb{R}^{kd} \times (0, \infty) \times (0, \infty)$ *there exists a zero curve* γ *of* $\tilde{\rho}_a$, *emanating from* $\left(0, k^0, \mu_0, \nu_0\right)$, *along which the* $(kd + 2) \times (kd + 3)$ *Jacobian matrix* $D\tilde{\rho}_a$ *has full rank, that does not intersect itself and is disjoint from any other zeros of* $\tilde{\rho}_a$. *If* ν *is bounded along* γ , *then* γ *accumulates at a point* $(1, \bar{X}, \bar{\mu}, \bar{\nu})$, *where* $(\bar{X}, \bar{\mu}, \bar{\nu})$ *is a KKT point for the constrained clustering problem*

$$\min_X F(C, X) \quad \textit{subject to} \quad -\Psi(X) \leqq 0, \quad Z(X) \leqq 0.$$

Furthermore, if $\operatorname{rank} D\tilde{\rho}_a(1, \bar{X}, \bar{\mu}, \bar{\nu}) = kd + 2$, *then the curve* γ *connecting* $\left(0, k^0, \mu_0, \nu_0\right)$ *to* $(1, \bar{X}, \bar{\mu}, \bar{\nu})$ *has finite arc length.*

4. Experimental results

4.1. Random Constraints

Experiments to discover the effectiveness of the homotopy tracking algorithm with the proposed homotopy map, as compared to popular existing constrained clustering algorithms, are presented here. The constraints used involve combinations of ML and MNL constraints (solving problems involving solely ML constraints are fairly straightforward polynomial time graph problems).

The existence of MNL constraints in the constraint sets is crucial to understanding the complexity of the test problems. Davidson et al. [35] state that as a rough rule of thumb a set of constraints can be understood as fundamentally "difficult" for these iterative K-Means approaches if any single datapoint appears in k or more MNL constraints. As such, for each dataset presented here, both an "easy" and a "difficult" set of constraints were generated. The "easy" constraint set involves one hundred constraints such that no datapoint appears more than $k - 1$ times in a mix of ML and MNL constraints. The "difficult" constraint set, also one hundred constraints, involves predominately MNL constraints, and guarantees that at least one datapoint is involved in k MNL constraints. In both cases, the generated constraints were completely random, with no a priori knowledge about how well the generated constraints would guide the algorithms to a correct solution.

The datasets involved are all taken from the UCI machine learning dataset repository [36]. They represent a balanced selection of moderately easy clustering problems without constraints, and should demonstrate some of the key differences between the homotopy algorithms utilizing the maps $\bar{\rho}$ and $\tilde{\rho}$ developed here and the K-Means algorithms used previously. The datasets are "Liver Disorders" (liver), "Pima Indians Diabetes" (pima), "Steel Plates Faults" (faults), "Wine" (wine), "Iris" (iris), "Ionosphere" (iono), "Glass Identification" (glass), and "PAMAP2 Physical Activity Monitoring" (pamap). The datasets "faults" and "pamap" were modified in the following manner: The first three classification categories of the dataset "faults" were treated as additional data, and the last classification category was used for classification. The dataset "pamap" was

Table 1
Dataset summary.

	No. Instances	No. Features	No. Categories
liver	345	6	2
pima	768	7	2
faults	1 941	31	2
wine	178	12	3
iris	150	4	3
iono	351	34	2
glass	214	10	6
pamap	175 498	53	12

modified by eliminating all data points of the "0" classification (as recommended by the contributors) and any data point with a "NaN" data value. See Table 1 for the relevant details for each dataset.

The K-Means algorithms used for the comparison are those presented by Bilenko et al. [37]: metric pairwise constrained K-Means (MPCK-Means), metric learning K-Means without pairwise constraints (MK-Means), and pairwise constrained K-Means without metric learning (PCK-Means). The standard K-Means result is also presented, to be used as a baseline. These algorithms were chosen for several reasons. First, K-Means is by far the most popular clustering algorithm, if only because of its intuitive approach and ease of programming; thus, K-Means algorithms modified for constrained clustering are the most likely to be consulted by researchers who are interested in constrained clustering problems. Second, these constrained K-Means algorithms minimize a summed penalty function based on the distance from the cluster centroids to the data points assigned to that cluster. While this penalty function may be discrete, it is still similar enough to the penalty function presented here to make comparisons between these algorithms and the homotopy approach feasible.

It is worth noting immediately that three things set the homotopy algorithms apart from the K-Means algorithms presented here. First, for the K-Means algorithms, the ordering of the constraints plays a nonnegligible role in the quality of the final result, meaning that finding the best result theoretically involves searching every permutation of a given constraint set (which is not computationally feasible). For a homotopy algorithm, the ordering of the constraints is unimportant. Second, not only are problems involving concentrations of MNL constraints involving the same datapoint not qualitatively more "difficult" for homotopy algorithms, but, since distances only need to be calculated once per iteration, problems involving concentrations of datapoints are computationally less intense than problems where the constraints are more diverse, at least until each datapoint is involved in at least one constraint. Finally, the homotopy algorithms, like the K-Means algorithm, is limited to convex clusterings, which for some datasets can be potentially debilitating. In contrast, the adapted K-Means algorithms presented here distinguish between cluster assignment and cluster centroids, which allows for nonconvex clusterings.

Ten experiments are conducted for each algorithm on each dataset, and the tables report the minimum, median, and maximum of the stated index for the ten experiments. Table 2 shows the adjusted Rand index [38] of each dataset measured against the proper classification, for each algorithm discussed here, for 100 "easy" (in the sense of not NP-hard) constraints. Table 3 shows the same data for 100 "hard" (as in NP-hard) constraints. The exception is the "pamap" dataset, which used 250 of each constraint to allow for more differentiation (due to the massive size of the dataset). Note that these two tables show the highest adjusted Rand index found along the homotopy method's reported trace, excluding the K-Means solution starting point. For both of these tables the simpler homotopy map $\bar{\rho}$ was used. Asterisks indicate the highest adjusted Rand index found for the median results, with multiple asterisks indicating that several algorithms arrived at the same partitioning.

4.2. ϵ- and δ-constraints

One reasonable question that arises in semisupervised learning is what kinds of information would need to be present in a clustering problem that could not be represented by the datapoints themselves. One answer makes reference to the cluster hypothesis itself.

The cluster hypothesis states that if two datapoints are close to each other (for a vague notion of closeness), then they should belong to the same cluster; if they are far apart, they should belong to different clusters. The constraints that formalize this statement are the previously mentioned ϵ- and δ-constraints, which can be represented as disjunctions and conjunctions of the classic "must-link" and "must-not-link" constraints [35]. This presents a useful way of generating meaningful constraints to apply to the datasets at hand. The number of such constraints can grow quite large as the number of datapoints in the set increases (depending on the values assigned to ϵ and δ), but the entire set of constraints need not be brought to bear for the solution to show improvement. One advantage of such constraints is that they do not depend on the "real" clustering, which is to say the classification, of a given dataset, which is often unknown in practice. Thus, applying these constraints to test problems can yield tests of improvement in whatever measure of cluster hypothesis satisfaction is desired (of which there are many).

100 random constraints were generated using ϵ- and δ-constraints based on reasonable values for the given data sets. Since the adjusted Rand index is useless in this context, the Davies–Bouldin index (DBI) [39] was used instead. The DBI is a nonnegative measure of conformity to the cluster hypothesis; a lower DBI indicates closer conformity to the clustering

Table 2
Adjusted Rand index, "easy" constraints.

	K-means	MK-means	PCK-means	MPCK-means	Homotopy $\bar{\rho}$
	−0.0064	−0.0036	−0.0042	−0.0045	−0.0040
liver	−0.0064	−0.0036*	−0.0042	−0.0045	−0.0040
	−0.0064	−0.0036	−0.0042	−0.0045	−0.0040
	0.0744	0.0720	0.0510	0.0164	0.1322
pima	0.0744	0.0870	0.0510	0.0164	0.1322*
	0.0744	0.1040	0.0510	0.0164	0.1322
	0.1358	−0.1028	0.1109	−0.0863	0.1127
faults	0.1358*	−0.1028	0.1133	−0.0849	0.1127
	0.1358	−0.1028	0.1159	−0.0837	0.1127
	0.3711	0.7549	0.3420	0.6211	0.4377
wine	0.3711	0.7692*	0.3420	0.6211	0.4377
	0.3394	0.8309	0.3420	0.6211	0.4377
	0.4225	0.4290	0.5195	0.5234	0.8841
iris	0.7163	0.8857*	0.5195	0.5234	0.8841
	0.7302	0.8857	0.5195	0.5234	0.8841
	0.1728	0.1776	0.1122	0.1122	0.2450
iono	0.1776	0.1776	0.1122	0.1122	0.2450*
	0.1776	0.1776	0.1122	0.1122	0.2450
	0.1790	0.2023	0.1720	0.1824	0.1967
glass	0.1790	0.2023*	0.1720	0.1824	0.1967
	0.2258	0.2482	0.1720	0.1824	0.2258
	0.6457	0.3046	0.5560	0.2695	0.6457
pamap	0.6457*	0.3046	0.5560	0.2695	0.6457*
	0.6457	0.3046	0.5560	0.2695	0.6457

Table 3
Adjusted Rand index, "hard" constraints.

	K-means	MK-means	PCK-means	MPCK-means	Homotopy $\bar{\rho}$
	−0.0064	−0.0046	−0.0109	−0.0077	0.0400
liver	−0.0064	−0.0046	−0.0080	−0.0052	0.0400*
	−0.0064	−0.0046	0.0102	0.0253	0.0400
	0.0744	0.0722	0.0652	0.0114	0.0775
pima	0.0744	0.0722	0.0696	0.0340	0.0775*
	0.0744	0.0722	0.0696	0.0422	0.0775
	0.1358	−0.1028	0.1324	−0.0839	0.1119
faults	0.1358	−0.1028	0.1420*	−0.0839	0.1119
	0.1358	−0.1028	0.1459	−0.0832	0.1119
	0.3394	0.7692	0.3265	0.6731	0.8666
wine	0.3711	0.7840	0.3818	0.7031	0.8666*
	0.3711	0.8170	0.4451	0.8636	0.8666
	0.4225	0.4290	0.5127	0.5411	0.9216
iris	0.7163	0.8857	0.6779	0.8017	0.9216*
	0.7302	0.8857	0.8015	0.9222	0.9216
	0.1728	0.1776	0.1049	0.1122	0.1943
iono	0.1776	0.1776	0.1413	0.1122	0.1943*
	0.1776	0.1776	0.1413	0.1122	0.1943
	0.0162	0.0000	0.1849	0.1560	0.0162
glass	0.1790	0.2023	0.2422*	0.1741	0.1790
	0.2258	0.2482	0.2608	0.2102	0.2258
	0.6457	0.2929	0.6454	0.2700	0.6513
pamap	0.6457	0.2929	0.6454	0.2700	0.6513*
	0.6457	0.2929	0.6454	0.2700	0.6513

hypothesis. The "pamap" dataset was not used due to the difficulty in generating meaningful differences in the clusterings with this sort of constraint. Table 4 shows these results. The (more computationally expensive) homotopy map $\tilde{\rho}$ was used here to demonstrate the utility of incorporating the K-Means approximation function into the map. In these experiments the best cluster found by the homotopy algorithm was also uniformly the last one.

Table 4
Davies–Bouldin index, ϵ- and δ-constraints.

	K-means	MK-means	PCK-means	MPCK-means	Homotopy $\tilde{\rho}$
	1.7349	2.3067	1.7679	1.2516	0.8706
liver	1.7349	1.8801	1.4568	1.2516	0.8706*
	1.7349	1.6682	1.3542	1.2516	0.8706
	1.9995	0.9883	0.8762	0.8681	0.8094
pima	1.5653	1.9403	1.0585	1.4436	0.8601*
	1.5387	1.9316	1.0585	1.4436	0.8601
	0.9392	0.9883	0.8762	0.8681	0.8094
faults	0.9392	0.9652	0.8762	0.8681	0.8094*
	0.9392	0.9652	0.8637	0.8681	0.8094
	1.5126	1.6650	0.8185	1.5393	0.6604
wine	1.5126	1.5507	0.6542	1.4515	0.6097*
	1.5126	1.4506	0.6101	1.3447	0.4948
	0.7373	1.5023	1.4662	0.9612	0.9379
iris	0.7373	0.9455	0.8877	0.7175	0.6453*
	0.7373	0.7445	0.7041	0.6585	0.5776
	2.0706	2.0512	1.6898	1.6898	1.6188
iono	2.0706	1.8936	1.8936	1.6898	1.6188*
	2.0706	1.8936	1.8919	1.6898	1.6188
	3.4599	1.8348	1.0414	1.8284	2.2789
glass	2.2910	1.4204	1.0414*	1.2820	1.2204
	1.7415	1.0621	1.0414	1.0038	0.2403

5. Conclusion

For the experiments that use the true classifications of these datasets to show the validity of the "easy" constraints (Table 2), the homotopy method performed well for the pima, iris, and iono datasets, and performed well for the pamap dataset when compared to the other constrained clustering metrics for the "easy" constraint set, although none of them managed to improve on the K-Means starting points for that problem. The metric learning without pairwise constraints algorithm (M-Kmeans) performed about as well, performing the best in the liver, wine, iris, and glass datasets. It is worth noting that for the two largest datasets in terms of instance numbers (faults and pamap), no constrained clustering algorithm approached the actual classification better than the straightforward K-Means solution. This is not entirely surprising, as a dense population of datapoints means that any rearrangement of clustering, even one based on constraints known to be correct, is going to inevitably cause a transfer of instances from a cluster where they satisfy the clustering hypothesis to one where they do not, with an expected degradation of quality in the resulting partition.

The "hard" dataset saw a much better showing by the homotopy method, which performed the best for the liver, pima, wine, iris, iono, and pamap datasets. The faults and glass datasets were, in this case, best captured by the pairwise constrained K-Means algorithm. This demonstrates in particular the power of the homotopy map when faced with this kind of "hard" problem. The ordering of the constraints is quite important to the other algorithms, especially when a large number of MNL constraints are employed, as is the case here. While these constraints were randomly generated, they were randomly generated to satisfy the "hardness" of the constraint set, which meant that the first $k+1$ constraints of each list were MNL constraints all involving the same datapoint. This is exactly the kind of ordering guaranteed to give the other algorithms the most difficulty in satisfying the constraints. The homotopy algorithm, on the other hand, is unbiased by constraint ordering, and largely indifferent to constraint type.

The adjusted Rand index is a good tool for a posteriori judgment of clusters, but semisupervised clustering problems do not have the classification a priori. The tools of the researcher are (usually) limited to intercluster and intracluster distances, with limited extra information not presented as a dimension of the clustering. For this sort of situation, the homotopy method shines in the ϵ- and δ-constraint experiments (Table 4), due to the use of the K-Means approximation $\hat{K}(X)$ in the homotopy formulation. The net effect of this approximation is to cause the homotopy method to account for local minima of the K-Means approximation as λ increases. Assuming that $\Psi(\hat{X}) > 0$ and $Z(\hat{X}) < 0$, with a small ν and μ, which is almost always the case in practice, an $\hat{X}$ at some $\hat{\lambda} < 1$ that would satisfy $\nabla_X F(C, \hat{X}) \approx 0$, but violate $\nabla_X \hat{K}(\hat{X}) \approx 0$, would not lie along γ. Thus, γ contains those solutions that do not strongly violate the clustering hypothesis as arc length increases along γ, resulting in the end point at $\lambda = 1$ being generally favorable to the cluster hypothesis.

All of this is an aside to the true purpose of the development of the homotopy map. The Dunn index [40] is a reasonable tool for measuring the validity of multiple clusterings of the same dataset, although it is one that cannot handle nonconvex clusters (hence it was not used in the ϵ- and δ-experiments). Figs. 2–5 show the utility of the homotopy map $\bar{\rho}$ without reference to the "correct" clustering, simulating the constraints that a researcher may reasonably discover on their own. Note that generally, the Dunn index improves over the original K-Means clustering as constraints are satisfied. Since satisfying the given set of constraints improves the quality of the discovered partitions, the improvement to the Dunn index (or other

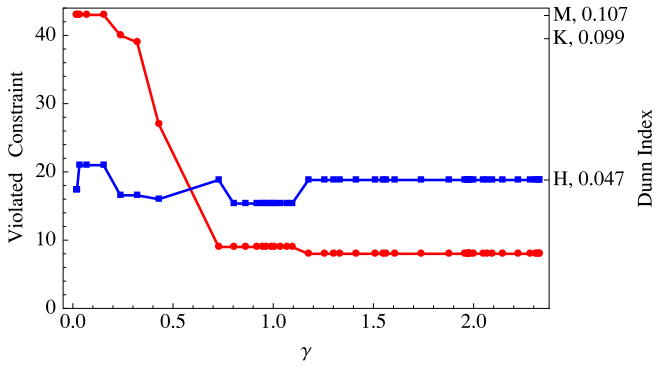

Fig. 2. The iris dataset with "easy" constraints. The Dunn index is tracked against the arc length of γ in blue, while the satisfied constraints are tracked in red. The Dunn Indices for the final homotopy ($\bar{\rho}$) clustering ("H"), MK-Means clustering ("M"), and K-Means clustering ("K") are also shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

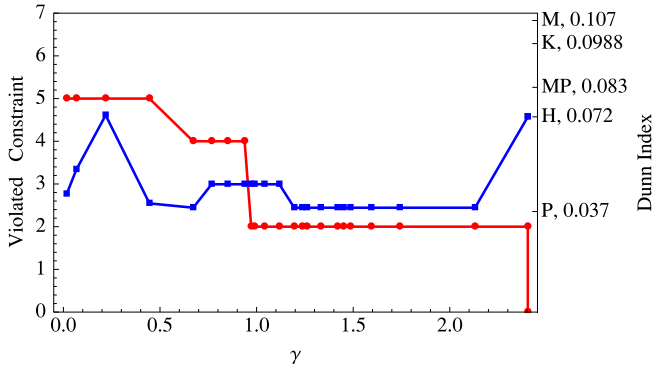

Fig. 3. The iris dataset with "hard" constraints. The Dunn Indices for the final homotopy ($\bar{\rho}$) clustering ("H"), MK-Means clustering ("M"), PK-Means clustering ("P"), MPK-Means clustering ("MP"), and K-Means clustering ("K") are also shown.

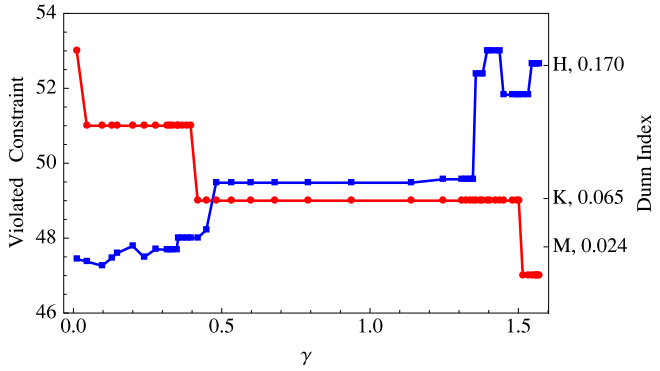

Fig. 4. The liver dataset with "easy" constraints.

cluster metric) can be viewed as providing a standard to judge imposed constraints. In this case, of course, the constraints are all known to be valid. In addition, these figures make it easy to show how valid constraints can guide the homotopy algorithm through regions of poor clustering to establish better partitions. For example, in Fig. 3 it is easy to see that several regions of poor clustering are encountered as arc length increases along γ, but the final clustering in this case, with all constraints satisfied, happens to be the best one. Of course, this need not be the case with these problems, but it would appear to speak to the reasonable nature of these constraints.

The new homotopy approach for constrained clustering problems uses state-of-the-art mathematical software to characterize multicriteria problems in constrained clustering. Just as in other applications of homotopy methods to science and engineering, the application of homotopy methods to machine learning problems can usher in greater understanding

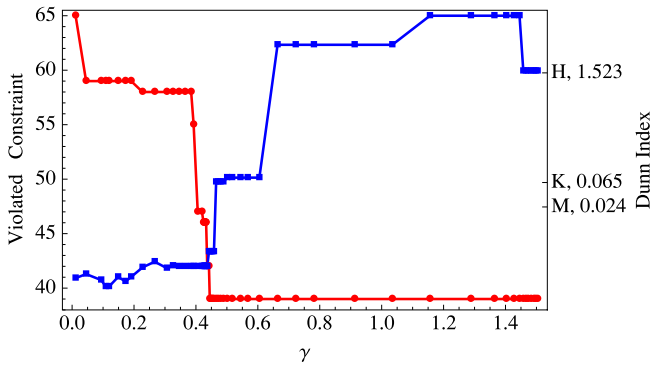

Fig. 5. The liver dataset with "hard" constraints.

of solution sets and the value of constraints. Besides the strong mathematical foundations and rigorous formalisms brought to classical machine learning problems, this homotopy approach has the potential to greatly reduce the ad hoc nature of methodological experimentation that is prevalent in practice. The approach given here not only helps extract better patterns from data, but also helps formally understand the internal workings of machine learning techniques. Future work includes homotopy maps for other multicriteria machine learning problems such as information bottleneck, time series segmentation, and transfer learning.

Acknowledgments

This work was supported in part by AFRL Grant FA8650-09-2-3938, NSF Grant DGE-154362, and NSF Grant CNS-1565314.

References

[1] M.F. Balcan, A. Blum, A discriminative model for semi-supervised learning, J. ACM 57 (3) (2010) 19:1–19:46.
[2] O. Chapelle, B. Scholkopf, A. Zien, Semi-Supervised Learning, first ed., MIT Press, Cambridge, MA, 2008.
[3] K. Sinha, M. Belkin, The value of labeled and unlabeled examples when the model is imperfect, Adv. Neural Inf. Process. Syst. 20 (2008).
[4] A. Demiriz, K. Bennett, P. Bradley, Chapter 9: Using assignment constraints to avoid empty clusters in K-Means clustering, in: K. Wagstaff, I. Davidson, S. Basu (Eds.), Constrained Clustering: Advances in Algorithms, Theory, and Applications, first ed., Chapman & Hall/CRC, Boca Raton, FL, 2008.
[5] X. Wang, I. Davidson, Flexible constrained spectral clustering, in: KDD'10, pp. 563–572, 2010.
[6] H. Yang, J. Callan, A metric-based framework for automatic taxonomy induction, in: ACL'09, Vol. 1, pp. 271–279, 2009.
[7] P. Luo, F. Zhang, H. Xiong, Y. Xiong, Q. He, Transfer learning from multiple source domains via consensus regularization, in: CIKM'08, pp. 103–112, 2008.
[8] M.E. Taylor, G. Kuhlmann, P. Stone, Autonomous transfer for reinforcement learning, in: AAMAS'08, Vol. 1, pp. 283–290, 2008.
[9] D. Zhang, J. He, Y. Liu, L. Si, R. Lawrence, Multi-view transfer learning with a large margin approach, in: KDD'11, pp. 1208–1216, 2011.
[10] Q. Yang, Y. Chen, G.R. Xue, W. Dai, Y. Yu, Heterogeneous transfer learning for image clustering via the social web, in: ACL'09, pp. 1–9, 2009.
[11] S. Basu, I. Davidson, K. Wagstaff (Eds.), Constrained Clustering: Advances in Algorithms, Theory, and Applications, first ed., Chapman and Hall, Boca Raton, FL, 2008.
[12] I. Davidson, S.S. Ravi, Clustering with constraints: feasibility issues and the K-Means algorithm, in: SDM'05, pp. 201–211, 2005.
[13] B.R. Dai, C.R. Lin, M.S. Chen, Constrained data clustering by depth control and progressive constraint relaxation, VLDB J. 16 (2007) 201–217.
[14] Y. Sato, M. Iwayama, Interactive constrained clustering for patent document set, in: PaIR'09, pp. 17–20, 2009.
[15] M.S. Hossain, S. Tadepalli, L.T. Watson, I. Davidson, R. Helm, N. Ramakrishnan, Unifying dependent clustering and disparate clustering for non-homogeneous data, in: KDD'10, pp. 593–602, 2010.
[16] M.S. Baghshah, S.B. Shouraki, Learning low-rank kernel matrices for constrained clustering, Neurocomputing 74 (12–13) (2011) 2201–2211.
[17] P. He, X. Xu, L. Chen, Constrained clustering with local constraint propagation, in: ECCV'12, pp. 223–232, 2012.
[18] J. Sese, Y. Kurokawa, M. Monden, K. Kato, S. Moroshita, Constrained clusters of gene expression profiles with pathological features, Bioinformatics 20 (17) (2004) 3137–3145.
[19] I. Davidson, S.S. Ravi, The complexity of non-hierarchical clustering with instance and cluster level constraints, Data Min. Knowl. Discov. 14 (2007) 25–61.
[20] I. Davidson, Two approaches to understanding when constraints help clustering, in: KDD'12, pp. 1312–1320, 2012.
[21] A. Corduneanu, T. Jaakkola, Continuation methods for mixing heterogeneous sources, in: UAI'02, pp. 111–118, 2002.
[22] S. Ji, L.T. Watson, L. Carin, Semisupervised learning of hidden Markov models via a homotopy method, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 275–287.
[23] O. Mangasarian, Nonlinear Programming, McGrawHill, New York, 1969.
[24] K.J. Arrow, L. Hurwicz, H. Uzawa, Constraint qualifications in maximization problems, Nav. Res. Logist. 8 (1961) 175–191.
[25] G. Kreisselmeier, R. Steinhauser, Systematic control design by optimizing a vector performance index, in: International Federation of Active Controls Symposium on Computer-Aided Design of Control Systems, Zurich, Switzerland, 1979.
[26] O. Mangasarian, Equivalence of the complementarity problem to a system of nonlinear equations, SIAM J. Appl. Math. 31 (1) (1976) 89–92.
[27] L.T. Watson, A globally convergent algorithm for computing fixed points of C^2 maps, Appl. Math. Comput. 5 (1979) 297–311.
[28] L.T. Watson, Probability-one homotopies in computational science, J. Comput. Appl. Math. 140 (2002) 785–807.
[29] S.N. Chow, J. Mallet-Paret, J.A. Yorke, Finding zeros of maps: homotopy methods that are constructive with probability-one, Math. Comp. 32 (1978) 887–899.

[30] L.T. Watson, S. Billups, A.P. Morgan, Algorithm 652: HOMPACK: a suite of codes for globally convergent homotopy algorithms, ACM Trans. Math. Softw. 13 (1987) 281–310.
[31] L.T. Watson, M. Sosonkina, R.C. Melville, A. Morgan, H. Walker, Algorithm 777: HOMPACK90: a suite of Fortran 90 codes for globally convergent homotopy algorithms, ACM Trans. Math. Softw. 23 (1997) 514–549.
[32] L.T. Watson, Solving the nonlinear complimentarity problem by a homotopy method, SIAM J. Control Optim. 17 (1979) 36–46.
[33] L.T. Watson, Theory of globally convergent probability-one homotopies for nonlinear programming, SIAM J. Optim. 11 (3) (2000) 761–780.
[34] R. Phillips, A Probabilistic Classification Algorithm with Soft Classification Output (Ph.D. thesis), Virginia Tech, Blacksburg, VA, 2009.
[35] I. Davidson, S.S. Ravi, Identifying and generating easy sets of constraints for clustering, in: AAAI'06, pp. 336–341, 2006.
[36] K. Bache, M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2013 [http://archive.ics.uci.edu/ml].
[37] M. Bilenko, S. Basu, R.J. Mooney, Integrating constraints and metric learning in semi-supervised clustering, in: ICML'04, pp. 11–18, 2004.
[38] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Amer. Statist. Assoc. 66 (336) (1961) 846–850.
[39] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[40] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (3) (1973) 32–37.