ALGORITHMS FOR MINIMIZING DIFFERENCES OF CONVEX FUNCTIONS AND APPLICATIONS

Mau Nam Nguyen

(joint work with D. Giles and R. B. Rector)

Fariborz Maseeh Department of Mathematics and Statistics
Portland State University

SIAM Conference on Imaging Sciences
Albuquerque, New Mexico, May 2016

Outline

1. Minimizing Differences of Convex Functions-The DCA
2. Nesterov's Smoothing Technique via Convex Analysis
3. The DCA and Nesterov's Smoothing Technique for Weighted Fermat-Torricelli Problems
4. The DCA and Nesterov's Smoothing Technique for Multifacility Location

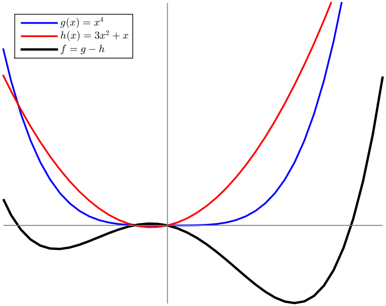
Differences of Convex Functions

Consider the problem

$$\text{minimize} \qquad f(x) := g(x) - h(x)\ , \qquad x \in \mathbb{R}^n$$

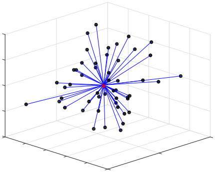
where $g\colon \mathbb{R}^n \to \mathbb{R}$ and $h\colon \mathbb{R}^n \to \mathbb{R}$ are convex.

We call $g - h$ a **DC decomposition** of f.



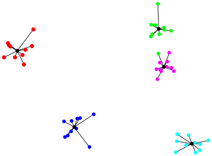
Examples of DC Programming

Fermat-Torricelli



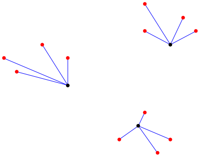
$f(x) = \sum_{i=1}^{m} c_i \|x - a_i\|$

Clustering



$f(x_1, \ldots, x_\ell) = \sum_{i=1}^{m} \min_{\ell=1}^{k} \|x_\ell - a_i\|^2$

Multifacility Location



$f(x_1, \ldots, x_\ell) = \sum_{i=1}^{m} \min_{\ell=1}^{k} \|x_\ell - a_i\|$

Subgradients and Fenchel Conjugates of Convex Functions

Definition

Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a convex function and let $\bar{x} \in \mathrm{dom}(f) := \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. A **subgradient** of f at $\bar{x}$ is any $v \in \mathbb{R}^n$ such that

$$\langle v, x - \bar{x}\rangle \leq f(x) - f(\bar{x}) \text{ for all } x \in \mathbb{R}^n.$$

The **subdifferential** $\partial f(\bar{x})$ of f at $\bar{x}$ is the set of all subgradients of f at $\bar{x}$.

Definition

Let $f\colon \mathbb{R}^n \to (-\infty, \infty]$ be a function. The **Fenchel conjugate** of f is defined by

$$f^*(x) = \sup_{u \in \mathbb{R}^n} \{\langle x, u\rangle - g(u)\}, \; x \in \mathbb{R}^n.$$

DC Algorithm-The DCA

Consider the problem

$$\text{minimize} \qquad f(x) := g(x) - h(x) \,, \qquad x \in \mathbb{R}^n$$

where $g\colon \mathbb{R}^n \to \mathbb{R}$ and $h\colon \mathbb{R}^n \to \mathbb{R}$ are convex.

The DCA[1]**.**

INPUT: $x_1 \in \mathbb{R}^n$, $N \in \mathbb{N}$
for $k = 1, \ldots, N$ **do**
  Find $y_k \in \partial h(x_k)$
  Find $x_{k+1} \in \partial g^*(y_k)$
end for
OUTPUT: x_{N+1}

[1] P.D. Tao, L.T.H. An, A d.c. optimization algorithm for solving the trust-region subproblem, SIAM J. Optim. 8 (1998), 476–505.

DC Algorithm-The DCA

Theorem

Let $g, h\colon \mathbb{R}^n \to (-\infty, \infty]$ be proper lower semicontinuous convex functions. Then $v \in \partial g^*(y)$ if and only if

$$v \in \operatorname{argmin}\left\{g(x) - \langle y, x\rangle \mid x \in \mathbb{R}^n\right\}.$$

Moreover, $w \in \partial h(x)$ if and only if

$$w \in \operatorname{argmin}\left\{h^*(y) - \langle y, x\rangle \mid y \in \mathbb{R}^n\right\}.$$

DC Algorithm-The DCA

The DCA.

INPUT: $x_1 \in \mathbb{R}^n$, $N \in \mathbb{N}$

for $k = 1, \ldots, N$ **do**

Find $y_k \in \partial h(x_k)$ or find y_k approximately by solving:

$$\text{minimize } \psi_k(y) := h^*(y) - \langle x_k, y \rangle, \; y \in \mathbb{R}^n.$$

Find $x_{k+1} \in \partial g^*(y_k)$ or find x_{k+1} approximately by solving:

$$\text{minimize } \phi_k(x) := g(x) - \langle x, y_k \rangle, \; x \in \mathbb{R}^n.$$

end for

OUTPUT: x_{N+1}

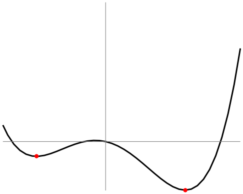
An Example of the DCA

$$\text{minimize } f(x) = x^4 - 3x^2 - x, x \in \mathbb{R}$$

Then $f(x) = g(x) - h(x)$, where $g(x) = x^4$ and $h(x) = 3x^2 + x$. We have

$$\partial h(x) = \{6x + 1\}, \partial g^*(y) = \left\{ \underset{x \in \mathbb{R}^n}{\text{argmin}} \left(x^4 - yx\right) \right\} = \left\{ \sqrt[3]{\frac{y}{4}} \right\}$$

$$x_{k+1} = \sqrt[3]{\frac{6x_k + 1}{4}}$$



DC Algorithm-The DCA

Definition

A function $h\colon \mathbb{R}^n \to (-\infty, \infty]$ is called γ-convex ($\gamma \geq 0$) if there exists $\gamma \geq 0$ such that the function defined by $k(x) := h(x) - \frac{\gamma}{2}\|x\|^2$, $x \in \mathbb{R}^n$, is convex. If there exists $\gamma > 0$ such that h is $\gamma-$convex, then h is called strongly convex with parameter γ.

Theorem

Consider the sequence $\{x_k\}$ generated by the DCA. Suppose that g is γ_1-convex and h is γ_2-convex. Then

$$f(x_k) - f(x_{k+1}) \geq \frac{\gamma_1 + \gamma_2}{2}\|x_{k+1} - x_k\|^2 \text{ for all } k \in \mathbb{N}.$$

DC Algorithm-The DCA

Definition

We say that an element $\bar{x} \in \mathbb{R}^n$ is a stationary point of the function $f = g - h$ if $\partial g(\bar{x}) \cap \partial h(\bar{x}) \neq \emptyset$. In the case where g and h are differentiable, $\bar{x}$ is a stationary point of f if and only if $\nabla f(\bar{x}) = \nabla g(\bar{x}) - \nabla h(\bar{x}) = 0$.

Theorem

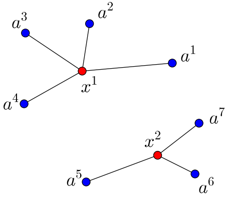
Consider sequence $\{x_k\}$ generated by the DCA. Then $\{f(x_k)\}$ is a decreasing sequence. Suppose further that f is bounded from below and that g is γ_1-convex and h is γ_2-convex with $\gamma_1 + \gamma_2 > 0$. If $\{x_k\}$ is bounded, then every subsequential limit of the sequence $\{x_k\}$ is a stationary point of f.

The DCA for Clustering

- Problem formulation[2]: Let a_i for $i = 1, ..., m$ be target points in $\mathbb{R}^n$.

$$\text{Minimize} \quad f(x_1, \ldots, x_\ell) := \sum_{i=1}^{m} \min\{\|x_l - a_i\|^2 \ : \ l = 1, ..., k\}$$

$$\text{over } x_l \in \mathbb{R}^n, \ l = 1, ..., k.$$



[2]L.T.H. An, M.T. Belghiti, P.D. Tao, A new efficient algorithm based on DC programming and DCA for clustering, J. Glob. Optim., 27 (2007), 503–608.

K-Mean Clustering

Let $x_1, x_2, \ldots, x_m$ be the data points and let $c_1, \ldots, c_k$ denote the centers.

- Randomly select k cluster centers.
- Assign each data point to the nearest center.
- Find the average of the data points assigned to each center.
- Repeat the second step with the obtained new centers in the third step until the centroids no longer move.

Although $k-$mean clustering is effective in many situations, it also has some disadvantages.

- The k-means algorithm does not necessarily find the optimal solution
- The algorithm is sensitive to the initial selected cluster centers

DCA for Clustering and K-Mean[3]

- Both DCA1 and DCA2 are better than K-means: the objective values given by DCA1 and DCA2 are much smaller than that computed by K-means.
- DCA2 is the best among the three algorithms: it provides the best solution with the shortest time. DCA2 is very fast and can then handle large-scale problems.

$$f(x_1,\ldots,x_k)=\sum_{i=1}^{m}\min_{\ell=1,\ldots,k}\|x_\ell-a_i\|_1$$

This is a nonsmooth nonconvex program for which there are rarely efficient solution algorithms, especially in the large scale setting.

[3] L.T.H. An, L.H. Minh, P.D. Tao, New and efficient DCA based algorithms for minimum sum-of-squares clustering, Pattern Recognition, 47 (2014), 388–401.

Nesterov's Smoothing Technique via Convex Analysis

Consider the function

$$f_0(x) := \max\{\langle Ax, u\rangle - \phi(u) \mid u \in Q\},$$

where A is an $m \times n-$matrix, Q is a nonempty closed bounded convex subset of $\mathbb{R}^m$, and $\phi\colon \mathbb{R}^m \to \mathbb{R}$ is a convex function.
Define $\|A\| = \sup\{\|Ax\| \mid \|x\| \le 1\}$.
For $\mu > 0$, define

$$f_\mu(x) := \max\{\langle Ax, u\rangle - \phi(u) - \frac{\mu}{2}\|u - u_0\|^2 \mid u \in Q\}, u_0 \in Q.$$

Then f_μ is a C^1 function with $\ell-$Lipschitz gradient where $\ell = \frac{\|A\|^2}{\mu}$ and $\nabla f_\mu(x) = A^T u_\mu(x)$. Here $u_\mu(x) \in Q$ is the element for which the maximum is attained in the definition of $f_\mu(x)$. [4]

[4] Nesterov: Smooth minimization of non-smooth functions. Math.Program., Ser. A 103 , 127-152 (2005).

Nesterov's Smoothing Technique via Convex Analysis

$$f^*(x) = \sup\{\langle x, u\rangle - f(u) \mid u \in \mathbb{R}^n\}.$$

Theorem

If f is $\mu-$strongly convex, then f^* has a Lipschitz continuous gradient with modulus $\frac{1}{\mu}$. Moreover, $\nabla f^*(x) = u(x)$, where $u(x)$ is the unique element of $\mathbb{R}^n$ for which the maximum is attained in the definition of $f^*(x)$.[a]

[a] J. Hiriart-Urruty and C. Lemaréchal. Convex Analysis and Minimization Algorithms I & II. Springer, New York, 1993.

Nesterov's Smoothing Technique via Convex Analysis

Theorem

Let A be an $n \times m$-matrix. Suppose that $\varphi \colon \mathbb{R}^m \to \mathbb{R}$ is a strongly convex function with parameter $\mu > 0$. Then the function $\mu \colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) := \max\{\langle Ax, u\rangle - \varphi(u) \mid u \in Q\}$$

is differentiable with $\nabla f(x) = A^T v(x)$, where $v(x)$ is the unique element for which the maximum is attained in the definition of $f(x)$. The gradient is Lipschitz continuous with constant $\ell = \frac{\|A\|^2}{\mu}$.

Nesterov's Smoothing Technique via Convex Analysis

We have
$f(x) = \max\{\langle A^T x, \rangle - [\varphi(u) + \delta(u; Q)] \mid u \in \mathbb{R}^m\} = g^*(A^T x)$,
where $g(u) := \varphi(u) + \delta(u; Q)$.

By the chain rule, $\nabla f(x) = A^T \nabla g^*(A^T x) = A^T u(Ax) = A^T v(x)$.
We also have

$$\begin{aligned}\|\nabla f(x_1) - \nabla f(x_2)\| &= \|A^T u(Ax_1) - A^T u(Ax_2)\| \\ &\le \|A^T\| \|u(Ax_1) - u(Ax_2)\| \\ &\le \|A\| \frac{1}{\mu} \|Ax_1 - Ax_2\| \le \frac{\|A\|^2}{\mu} \|x_1 - x_2\|.\end{aligned}$$

The Minkowski Gauge

Let F be a nonempty closed bounded convex set in $\mathbb{R}^n$ that contains the origin in its interior. Define the *Minkowski gauge* associated with F by

$$\rho_F(x) := \inf\{t > 0 \mid x \in tF\}.$$

Note that if F is the closed unit ball in $\mathbb{R}^n$, then $\rho_F(x) = \|x\|$.

Given a nonempty bounded set K, the *support function* associated with K is given by

$$\sigma_K(x) := \sup\{\langle x, y\rangle \mid y \in K\}.$$

It follows from the definition of the Minkowski function that $\rho_F(x) = \sigma_{F^\circ}(x)$, where

$$F^\circ := \{y \in \mathbb{R}^n \mid \langle x, y\rangle \leq 1 \text{ for all } x \in F\}.$$

Weighted Fermat-Torricelli problem

Let $a_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$ and let $c_i \neq 0$ for $i = 1, \ldots, m$ be real numbers. In the remainder of this section, we study the following generalized version of the Fermat-Torricelli problem:

$$\text{minimize } f(x) := \sum_{i=1}^{m} c_i \rho_F(x - a_i), \ x \in \mathbb{R}^n.$$

The function f has the following obvious DC decomposition:

$$f(x) = \sum_{c_i > 0} c_i \rho_F(x - a_i) - \sum_{c_i < 0} (-c_i) \rho_F(x - a_i).$$

Let $I := \{i \mid c_i > 0\}$ and $J := \{i \mid c_i < 0\}$ with $\alpha_i = c_i$ if $i \in I$, and $\beta_i = -c_i$ if $i \in J$. Then

$$f(x) = \sum_{i \in I} \alpha_i \rho_F(x - a_i) - \sum_{j \in J} \beta_j \rho_F(x - a_j).$$

Weighted Fermat-Torricelli problem

Let $a_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$ and let $c_i \neq 0$ for $i = 1, \ldots, m$ be real numbers. In the remainder of this section, we study the following generalized version of the Fermat-Torricelli problem:

$$\text{minimize } f(x) := \sum_{i=1}^{m} c_i \rho_F(x - a_i), \ x \in \mathbb{R}^n.$$

The function f has the following obvious DC decomposition:

$$f(x) = \sum_{c_i > 0} c_i \rho_F(x - a_i) - \sum_{c_i < 0} (-c_i) \rho_F(x - a_i).$$

Let $I := \{i \mid c_i > 0\}$ and $J := \{i \mid c_i < 0\}$ with $\alpha_i = c_i$ if $i \in I$, and $\beta_i = -c_i$ if $i \in J$. Then

$$f(x) = \sum_{i \in I} \alpha_i \rho_F(x - a_i) - \sum_{j \in J} \beta_j \rho_F(x - a_j).$$

Weighted Fermat-Torricelli problem

Theorem

Let $\gamma_1 := \sup\{r > 0 \mid B(0; r) \subset F\}$ *and* $\gamma_2 := \inf\{r > 0 \mid F \subset B(0; r)\}$. *Suppose that*

$$\gamma_1 \sum_{i \in I} \alpha_i > \gamma_2 \sum_{j \in J} \beta_j.$$

Then the function f *and its approximation* f_μ *have absolute minima.*

Smoothing the Minkowski Gauge

Given any $a \in \mathbb{R}^n$ and $\mu > 0$, a Nesterov smoothing approximation of $\varphi(x) := \rho_F(x-a)$ has the representation

$$\varphi_\mu(x) = \frac{1}{2\mu}\|x-a\|^2 - \frac{\mu}{2}\big[d(\frac{x-a}{\mu};F^\circ)\big]^2.$$

Moreover, $\nabla\varphi_\mu(x) = P(\frac{x-a}{\mu};F^\circ)$ and

$$\varphi_\mu(x) \le \varphi(x) \le \varphi_\mu(x) + \frac{\mu}{2}\|F^\circ\|^2,$$

where $\|F^\circ\| := \sup\{\|u\| \mid u \in F\}$.

Smoothing the Minkowski Gauge

Theorem

Given any $\mu > 0$, an approximation of the function f is the following DC function:

$$f_\mu(x) := g_\mu(x) - h_\mu(x),\ x \in \mathbb{R}^n,$$

where

$$g_\mu(x) := \sum_{i \in I} \frac{\alpha_i}{2\mu}\|x - a_i\|^2,$$

$$h_\mu(x) := \sum_{i \in I} \frac{\mu\alpha_i}{2}\Big[d(\frac{x - a_i}{\mu}; F^\circ)\Big]^2 + \sum_{j \in J} \beta_j \rho_F(x - a^j).$$

Moreover, $f_\mu(x) \leq f(x) \leq f_\mu(x) + \frac{\mu\|F^\circ\|^2}{2}\sum_{i \in I}\alpha_i$ for all $x \in \mathbb{R}^n$.

The DCA for Weighted Fermat-Torricelli Problems

Theorem

Given any $\mu > 0$, an approximation of the function f is the following DC function:

$$f_\mu(x) := g_\mu(x) - h_\mu(x),\ x \in \mathbb{R}^n,$$

where

$$g_\mu(x) := \sum_{i \in I} \frac{\alpha_i}{2\mu} \|x - a_i\|^2,$$

$$h_\mu(x) := \sum_{i \in I} \frac{\mu \alpha_i}{2} \Big[d\Big(\frac{x - a_i}{\mu}; F^\circ\Big)\Big]^2 + \sum_{j \in J} \beta_j \rho_F(x - a^j).$$

Moreover, $f_\mu(x) \le f(x) \le f_\mu(x) + \frac{\mu \|F^\circ\|^2}{2} \sum_{i \in I} \alpha_i$ *for all* $x \in \mathbb{R}^n$*.*

The DCA for Weighted Fermat-Torricelli Problems

Algorithm 3.

INPUTS: $\mu > 0$, $x_1 \in \mathbb{R}^n$, $N \in N$, F, $a^1, \ldots, a^m \in \mathbb{R}^n$, $c_1, \ldots, c_m \in \mathbb{R}$.

for $k = 1, \ldots, N$ **do**

Find $y_k = u_k + v_k$, where

$$u_k := \sum_{i \in I} \alpha_i \left[\frac{x_k - a_i}{\mu} - P\left(\frac{x_k - a_i}{\mu}; F^\circ\right)\right],$$

$$v_k \in \sum_{j \in J} \beta_j \partial \rho_F(x_k - a^j).$$

Find $x_{k+1} = \frac{y_k + \sum_{i \in I} \alpha_i a_i / \mu}{\sum_{i \in I} \alpha_i / \mu}$.

OUTPUT: x_{N+1}.

Multifacility Location

We now consider the multifacility location problem: given m points $a_1, \ldots, a_m \in \mathbb{R}^n$,

$$\text{minimize} \qquad f(x_1, \ldots, x_k) = \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \rho_F(x_\ell - a_i).$$

where F is a nonempty, closed and bound convex set containing the origin, and $\rho_F(x) = \inf_{x \in tF} \{t > 0\}$ is the Minkowski gauge.

When F is the closed unit ball B, the problem becomes

$$\text{minimize} \qquad f(x_1, \ldots, x_k) = \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \|x_\ell - a_i\|.$$

It can be shown that a globally optimal solution exists.

DC Decomposition

$$f(x_1, \ldots, x_k) = \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \|x_\ell - a_i\|$$

We will utilize the fact that

$$f(x_1, \ldots, x_k) = \sum_{i=1}^{m} \left[\sum_{\ell=1}^{k} \|x_\ell - a_i\| - \max_{r=1,\ldots,k} \sum_{\substack{\ell=1 \\ \ell \neq r}}^{k} \|x_\ell - a_i\| \right]$$

$$= \sum_{i=1}^{m} \left(\sum_{\ell=1}^{k} \|x_\ell - a_i\| \right) - \sum_{i=1}^{m} \left(\max_{r=1,\ldots,k} \sum_{\substack{\ell=1 \\ \ell \neq r}}^{k} \|x_\ell - a_i\| \right).$$

DC Decomposition

We obtain the μ-smoothing approximation $f_\mu = g_\mu - h_\mu$, where

$$g_\mu(x_1, \ldots, x_k) = \frac{1}{2\mu} \sum_{i=1}^{m} \sum_{\ell=1}^{k} \|x_\ell - a_i\|^2$$

$$h_\mu(x_1, \ldots, x_k) = \frac{\mu}{2} \sum_{i=1}^{m} \sum_{\ell=1}^{k} \left[d\left(\frac{x_\ell - a_i}{\mu}; B \right) \right]^2 + \sum_{i=1}^{m} \max_{r=1,\ldots,k} \sum_{\substack{\ell=1 \\ \ell \neq r}}^{k} \left(\frac{1}{2\mu} \|x_\ell - a_i\|^2 - \frac{\mu}{2} \left[d\left(\frac{x_\ell - a_i}{\mu}; B \right) \right]^2 \right)$$

To implement DCA, we need ∂g_μ^* and $\partial h_\mu \ldots$

∂g_μ

Using the Frobenius norm in a space of matrices, we express g_μ as

$$G_\mu(X) = \frac{m}{2\mu}\|X\|^2 - \frac{1}{\mu}\langle X, B\rangle + \frac{k}{2\mu}\|A\|^2,$$

with the inner product $\langle A, B\rangle = \sum_{\ell}^{k}\sum_{j}^{n} a_{\ell j}b_{\ell j}$,
X is the $k \times n$ matrix with rows $x_1, \ldots, x_k$,
A is $m \times n$ with rows $a_1, \ldots, a_m$, and
B is $k \times n$ whose every row is the sum $a_1 + \cdots + a_m$.

$\nabla G_\mu(X) = \frac{m}{\mu}X - \frac{1}{\mu}B$
$\nabla G^*_\mu(Y) = \frac{1}{m}(B + \mu Y)$

$\boxed{X \in \partial G^*(Y) \text{ iff } Y \in \partial G(X)}$

∂h_μ

$$h_\mu(x_1,\ldots,x_k) = \frac{\mu}{2}\sum_{i=1}^{m}\sum_{\ell=1}^{k}\left[d\left(\frac{x_\ell - a_i}{\mu}; B\right)\right]^2$$

$$+\sum_{i=1}^{m}\max_{r=1,\ldots,k}\sum_{\substack{\ell=1\\ \ell\neq r}}^{k}\left(\frac{1}{2\mu}\|x_\ell - a_i\|^2 - \frac{\mu}{2}\left[d\left(\frac{x_\ell - a_i}{\mu}; B\right)\right]^2\right)$$

∂h_μ

$$\nabla H_1(X) = \begin{bmatrix} \sum_i \frac{x_1 - a_i}{\mu} - P_B\left(\frac{x_1 - a_i}{\mu}\right) \\ \vdots \\ \sum_i \frac{x_k - a_i}{\mu} - P_B\left(\frac{x_k - a_i}{\mu}\right) \end{bmatrix}$$

For each $i = 1, \ldots, m$, there is some R_i such that the R-excluded sum is maximal. If we call this sum F_{R_i}, then ∇F_{R_i} is a $k \times n$ matrix whose $\ell^{\text{th}} \neq R$ row is $P_B\left(\frac{x_\ell - a_i}{\mu}\right)$, and R^{th} row is $\mathbf{0}$.

$$V \in \partial H_2(X) \text{ iff } V = \sum_{i=1}^{m} F_{R_i}$$

Multifacility Location Algorithm

INPUT: $X_1 \in \operatorname{dom} g$, $N \in \mathbb{N}$

for $k = 1, \ldots, N$ **do**

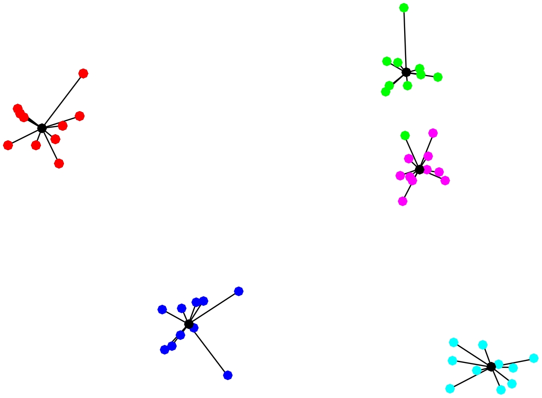
Compute $Y_k = \nabla H_1(X_k) + V_k$

Compute $X_{k+1} = \frac{1}{m}(B + \mu Y_k)$

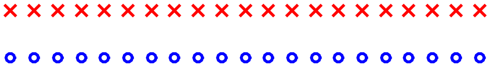
end for

OUTPUT: x_{N+1}

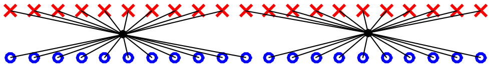
Clustering



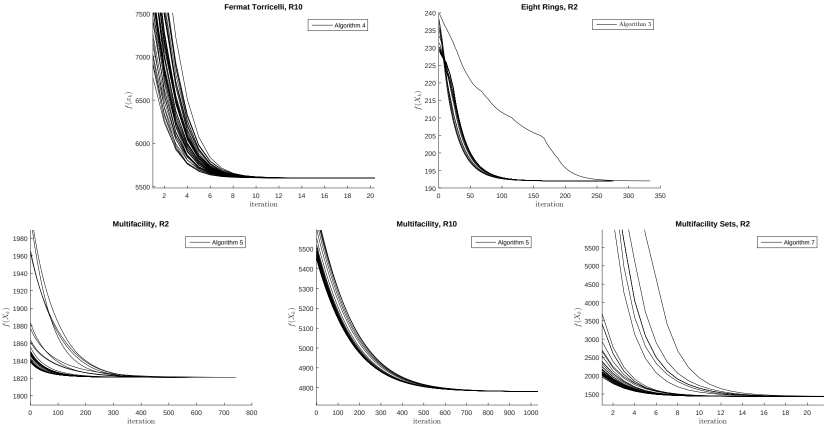
Clustering



Clustering



Results


Fermat Torricelli, R10
Algorithm 4
$f(x_k)$
iteration
Eight Rings, R2
Algorithm 5
$f(X_k)$
iteration
Multifacility, R2
Algorithm 5
$f(X_k)$
iteration
Multifacility, R10
Algorithm 5
$f(X_k)$
iteration
Multifacility Sets, R2
Algorithm 7
$f(X_k)$
iteration

References

[1] An, Belghiti, Tao: A new efficient algorithm based on DC programming and DCA for clustering. J. Glob. Optim., 27 (2007), 503–608.

[2] Nam, Rector, Giles: Minimizing Differences of Convex Functions and Applications to Facility Location and clustering. *arXiv*:1511.07595 (2015).

[3] Nesterov: Smooth minimization of non-smooth functions. *Math.Program., Ser. A* **103** , 127-152 (2005).

[4] Rockafellar: Convex Analysis, Princeton University Press, Princeton, NJ, 1970.