Statistics and Computing **12**: 115–123, 2002
© 2002 Kluwer Academic Publishers. *Manufactured in The Netherlands.*

Steepest descent algorithms in a space of measures

ILYA MOLCHANOV* and SERGEI ZUYEV†

**Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK*
ilya@stats.gla.ac.uk. (http://www.stats.gla.ac.uk/~ilya)
†*Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, UK*
sergei@stams.strath.ac.uk. (http://www.stams.strath.ac.uk/~sergei)

The paper describes descent type algorithms suitable for solving optimisation problems for functionals that depend on measures. We mention several examples of such problems that appear in optimal design, cluster analysis and optimisation of spatial distribution of coverage processes.

Keywords: gradient methods, steepest descent, *k*-means, P-means, optimal experimental design, Poisson process, Boolean model, Splus, R

1. Introduction

Functionals that depend on measures naturally appear in many areas of science. In this paper we give a few examples mainly from statistics and probability including construction of optimal designs, finding mixture distribution and optimising functionals of Poisson point processes whose distributions are determined by the corresponding intensity measure. Such point processes naturally appear in approximation problems for sets and functions (McClure and Vitale 1975, Schneider 1988) and optimising complex networks (Okabe *et al.* 2000) with nodes or resource allocations determined by a measure.

Given a functional $f(\mu)$ of a measure, in most cases it is not feasible to find an analytical expression for a measure μ that minimises it over a given family of measures. This calls for use of numerical algorithms producing a sequence of measures whose accumulating point solves the underlying minimisation problem. This paper develops algorithms of steepest descent type applicable for a general differentiable functional f and for families of measures that satisfy general linear constraints. Note that *linear* programming problems in the space of measures for a linear functional f was considered in Kellerer (1988) and Lai and Wu (1994).

Specific features of gradient algorithms for optimisation on the family $\mathbb{M}$ of (non-negative) measures are explained by the fact that $\mathbb{M}$ does not form a linear space, but is a cone in the linear space $\tilde{\mathbb{M}}$ of all signed measures. This immediately renders inadmissible many descent directions that lead out of $\mathbb{M}$. A kind of projected (or conditional) gradient method is therefore required that should also be suitable to deal with typical constraints that are imposed on measures, like keeping the total mass fixed or preserving the centre of gravity (expectation), etc.

The structure of the paper is as follows. In Section 2 we describe a general optimisation on measures framework and formulate a first-order necessary condition for extremum. Section 3 contains motivation examples of optimisation problems on measures on which our numeric algorithms will be tested later in Section 7. Section 4 provides characterisation of the steepest descent direction leading to the algorithms described in Sections 5 and 6.

2. Optimisation in the space of measures

Let $f(\mu)$ be a numerical functional defined on measures $\mu \in \mathbb{M}$, where $\mathbb{M}$ is the family of all non-negative finite measures on a Polish space X with its Borel σ-algebra. In many cases f can be naturally extended to a functional on the space $\tilde{\mathbb{M}}$ of all signed measures with bounded total variation. The Jordan decomposition of a signed measure $\mu \in \tilde{\mathbb{M}}$ is denoted by

$$\mu = \mu^+ - \mu^-,$$

and $\|\mu\| = \mu^+(X) + \mu^-(X)$ is the total variation of μ. In numerical implementation X becomes a grid in a Euclidean space $\mathbb{R}^d$ and $\mu \in \mathbb{M}$ is a non-negative array indexed by the grid's nodes.

0960-3174 © 2002 Kluwer Academic Publishers

source: https://doi.org/10.7892/boris.85359 | downloaded: 27.7.2024

Consider the following optimisation problem:

$$f(\mu) \to \inf, \quad \mu \in \mathbb{M}, \quad H(\mu) \in C, \tag{2.1}$$

where C is a closed convex subset of $\mathbb{R}^k$, $f : \tilde{\mathbb{M}} \mapsto \mathbb{R}$ and $H : \tilde{\mathbb{M}} \mapsto \mathbb{R}^k$ are Fréchet differentiable functions. Remember that f is said to be Fréchet differentiable (Hille and Phillips 1957) if

$$f(\mu + \eta) = Df(\mu)[\eta] + o(\|\eta\|) \quad \text{as } \|\eta\| \to 0.$$

Here $Df(\mu)[\eta]$ is a bounded linear functional on $\eta \in \tilde{\mathbb{M}}$ which determines the directional derivative of f in the direction of η. This directional derivative is equal to

$$Df(\mu)[\eta] = \lim_{t \downarrow 0} t^{-1}(f(\mu + t\eta) - f(\mu)).$$

Theorem 2.1 below focuses on a common case of finitely many differentiable constraints of the equality and inequality types:

$$\begin{cases} H_i(\mu) = 0, & i = 1, \dots, m; \\ H_j(\mu) \le 0, & j = m+1, \dots, k. \end{cases} \tag{2.2}$$

These constraints correspond to (2.1) with $H = (H_1, \dots, H_k) : \tilde{\mathbb{M}} \mapsto \mathbb{R}^k$ and $C = \{0\}^m \times (-\infty, 0]^{k-m}$.

Most differentiable functionals of measures met in practice have derivatives which can be represented in the integral form. In this paper we will assume that there exist measurable real-valued functions $d_f(x, \mu)$ and $h_i(x, \mu)$, $i = 1, \dots, k$, such that for all $\eta \in \tilde{\mathbb{M}}$

$$\begin{aligned} Df(\mu)[\eta] &= \int d_f(x, \mu)\eta(dx) \quad \text{and} \\ DH(\mu)[\eta] &= \int h(x, \mu)\eta(dx), \end{aligned} \tag{2.3}$$

where $h = (h_1, \dots, h_k)$. Note that all integrals are over X unless specified otherwise. The function $d_f(x, \mu)$ is called the gradient.

In addition, it is always assumed that the solution of Problem (2.1) is regular, i.e. it satisfies Robinson's regularity condition (Robinson 1976, Theorem 1) which guarantees the existence and boundedness of the Lagrange multipliers (see, Zowe and Kurcyusz 1979). In case of constraints (2.2) satisfying (2.3), Robinson's regularity condition means the linear independence of the functions $h_1(\cdot, \mu), \dots, h_m(\cdot, \mu)$ and the existence of $\eta \in \tilde{\mathbb{M}}$ such that

$$\begin{cases} \int h_i(x, \mu)\eta(dx) = 0 & \text{for all } i = 1, \dots, m, \\ \int h_j(x, \mu)\eta(dx) < 0 & \text{for all } j \in \{m+1, \dots, k\} \\ & \text{verifying } H_j(\mu) = 0. \end{cases} \tag{2.4}$$

Without the inequality constraints (2.4) trivially holds with η taken to be the zero measure. The following theorem is a particular case of (Molchanov and Zuyev 2000a, Theorem 4.1).

Theorem 2.1. *Let μ be a regular local minimum of f subject to* (2.2). *Then there exists $u = (u_1, \dots, u_k)$ with $u_j \le 0$ if $H_j(\mu) = 0$ and $u_j = 0$ if $H_j(\mu) < 0$ for $j \in \{m+1, \dots, k\}$, such that*

$$\begin{cases} d_f(x, \mu) = u\, h(x, \mu)^\top & \mu\text{-almost everywhere}, \\ d_f(x, \mu) \ge u\, h(x, \mu)^\top & \text{for all } x \in X. \end{cases} \tag{2.5}$$

Example 2.1. An important example concerns minimisation of $f(\mu)$ assuming that the total mass of μ is fixed at a value a. In the above framework, this corresponds to the case when $k = 1$ and the only constraint is $H_1(\mu) = \int \mu(dx) - a = 0$. Then (2.5) turns into the following necessary condition for a minimum in the fixed total mass problem:

$$\begin{cases} d_f(x, \mu) = u & \mu\text{-almost everywhere}, \\ d_f(x, \mu) \ge u & \text{for all } x \in X. \end{cases} \tag{2.6}$$

3. Examples of optimisation problems

3.1. *Optimal experimental design*

The basic problem in the theory of linear optimal design concerns the best choice of design (or observation) points x_i in the following regression model:

$$y_i = \sum_{j=1}^{k} \beta_j f_j(x_i) + \varepsilon_i = f(x_i)\beta^\top + \varepsilon_i, \quad i = 1, \dots, n, \tag{3.1}$$

where x belongs to a *design space* X, $\beta = (\beta_1, \dots, \beta_k)$ is a vector of unknown parameters, $f(x) = (f_1(x), \dots, f_k(x))$ is a row of linearly independent functions on X and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. uncorrelated centred errors with the variance σ^2. The theory of experimental design addresses the problem of choosing the observation points $x_1, \dots, x_n$ in order to achieve better properties of the least squares estimator $\hat{\beta}$.

Let $\mu(dx)$ be a probability distribution on X describing the frequency of taking x as an observation point. Then the covariance matrix $\|\mathbf{cov}(\hat{\beta}_i, \hat{\beta}_j)\|$ equals $\sigma^2 M(\mu)^{-1}$, where

$$M(\mu) = \int f(x)^\top f(x)\, \mu(dx)$$

is the *information matrix*, see, e.g., Atkinson and Donev (1992) for details. The measure μ that maximises $\det M(\mu)$ or, equivalently, minimises

$$f(\mu) = -\log \det M(\mu)$$

is called D-*optimal design measure* (taking logarithm makes it a convex minimisation problem). The gradient $d_f(x, \mu) = -f(x)M^{-1}(\mu)f^\top(x)$ then becomes the standardised variance of the predicted response at point x (Atkinson and Donev 1992), also called the sensitivity function (Fedorov and Hackl 1997). Typically, the only constraint is that μ is a probability measure, while further constraints can be naturally incorporated if such need arises. For example, let $X = \mathbb{R}^d$ and e_i, $i = 1, \dots, d$ be the unit coordinate vectors. Assume that along with the constraint on the total mass $\mu(X) = 1$ the expectation of μ, which is a vector $m = \int x\mu(dx)$, is fixed. These constraints can be written

as $H(\mu) = (m, 0)$, where $H(\mu)$ is a $(d+1)$-dimensional vector function with the components $H_i(\mu) = \int \langle x, e_i \rangle \mu(dx)$ for $i = 1, \ldots, d$ and $H_{d+1}(\mu) = \mu(X) - 1$. Clearly, $DH(\mu)[\eta] = \int h(x)\eta(dx)$ for $h(x) = (x, 1)$, $x \in \mathbb{R}^d$.

3.2. *Mixtures*

Let $\{p_x(\cdot) : x \in X\}$ be a parametric family of probability densities with respect to some σ-finite measure on X (in usual for the literature on mixtures notation, θ replaces x and Θ replaces X). Define the function

$$p_\mu(y) = \int p_x(y)\mu(dx)$$

to be the mixture density corresponding to mixing distribution μ. Given a random sample $y_1, \ldots, y_n$, the objective will be to estimate the mixing distribution μ that maximises the log-likelihood

$$f(\mu) = \sum_{i=1}^{n} \log p_\mu(y_i).$$

This is a concave objective function in a maximisation problem. The gradient function (that also can be interpreted as a score function) can be easily calculated as

$$d_f(x, \mu) = \sum_{i=1}^{n} \frac{p_x(y_i)}{\int p_x(y)\mu(dx)}.$$

A rich source of examples of functionals on measures is provided by expectations of functionals on the corresponding Poisson processes. Two of such examples are considered below. Recall that a finite measure μ on X gives rise to a Poisson point process on X with μ being its intensity measure, so that the number of points in any set $K \subset X$ is Poisson with mean $\mu(K)$ and the numbers of points in disjoint sets are independent.

3.3. *P-means*

Given a set of observation points $Y = \{y_i\}$, the *k-means* is the set of k points $Z = \{z_1, \ldots, z_k\}$ minimising

$$\sum_{y_i \in Y} \rho(y_i, Z)^\beta$$

for $\beta = 2$, where $\rho(y_i, Z) = \min\{\|y_i - z_j\| : z_j \in Z\}$. The chosen norm is Euclidean, but, in general, it may be any other norm on the space that contains observation points. If $\beta = 1$, one obtains the k-medians of the sample (see, Hartigan 1975). The problem of finding k-means (or medians) is a *non-convex* optimisation problem and the algorithms proposed up to now for its solution do not guarantee to find the global minimum (see, Hartigan and Wong 1979). Typically, such algorithms require a reasonable initial guess of k-means that is subsequently improved using descent methods.

The Poisson mean (or P-mean) is the 'randomised' variant of the k-means where Z is a realization of a Poisson process with intensity measure μ and the total mass k (so that the mean number of Poisson points is k). The P-mean of Y is a measure μ that minimises

$$\begin{aligned} f(\mu) &= \mathbf{E}_\mu \sum_{y_i \in Y} \rho(y_i, Z)^\beta \\ &= \sum_{y_i \in Y} \mathbf{E}_\mu \rho(y_i, Z)^\beta \\ &= \sum_{y_i \in Y} \int_0^{u^\beta} \exp\{-\mu(B_{t^{1/\beta}}(y_i))\}\, dt, \end{aligned} \tag{3.2}$$

where $\mathbf{E}_\mu$ is the expectation with respect to Poisson process' distribution with intensity measure μ, $B_r(x)$ is a ball of radius r centred at x, and u is a certain fixed (large) number that is used to replace $\rho(y_i, Z)$ if Z is empty. Now (3.2) is a *convex* functional of μ, so that its global minimum can be found using gradient descent algorithms. Apart from being of interest on its own, the P-mean μ can be used to sample k points from μ that may be subsequently taken as a sensible starting configuration for the standard k-means search algorithms. It follows from Molchanov and Zuyev (2000b, Theorem 2.1) that

$$d_f(x, \mu) = \mathbf{E}_\mu \left[\sum_{y_i \in Y} \rho(y_i, Z \cup \{x\})^\beta - \sum_{y_i \in Y} \rho(y_i, Z)^\beta \right],$$

meaning that the gradient is the expected difference between the values of the objective function on the configuration Z with the point $\{x\}$ added and on Z itself. Direct computations now show that the gradient of the functional (3.2) can be expressed as

$$d_f(x, \mu) = -\sum_{y_i \in Y} \int_{\rho(x, y_i)^\beta}^{u^\beta} \exp\{-\mu(B_{t^{1/\beta}}(y_i))\}\, dt.$$

The functionals of type (3.2) also arise in many other situations, for instance, in locational optimisation (Okabe *et al.* 2000), and telecommunications modelling (Molchanov and Zuyev 2000b).

3.4. *Maximisation of the covered volume in a Boolean model*

Let again $Z = \{z_1, z_2, \ldots\}$ be a Poisson point process in $X \subset \mathbb{R}^d$ with a finite intensity measure μ. If $B_r(x)$ is a ball of radius r centred at x, then

$$\Xi = \bigcup_{z_i \in Z} B_r(z_i)$$

is called a Boolean model (see, Molchanov 1997). The set $B_r(0)$ is called the typical grain. It is easy to see that

$$\mathbf{P}\{x \notin \Xi\} = \exp\{-\mu(B_r(x))\}.$$

By Fubini's theorem, the expected area left uncovered by Ξ can be written as

$$f(\mu) = \int_X \mathbf{P}\{x \notin \Xi\}\, dx = \int_X \exp\{-\mu(B_r(x))\}\, dx.$$

Minimisation of $f(\mu)$ corresponds to maximisation of the volume covered by Ξ. One readily obtains that the gradient of $f(\mu)$ is

$$d_f(x, \mu) = -\int_{B_r(x)} \exp\{-\mu(B_r(z))\}\, dz.$$

The gradient can be interpreted as the expected increment of the covered area if a ball centred at x is being added to the Boolean model.

Note that all arguments above can be easily generalised to a random typical grain with a rather general shape and size.

4. Gradient methods

Algorithms of the steepest descent type and their various modifications are widely known in the optimisation literature (see, e.g., Polak 1997). As before, it is assumed that $f(\mu)$ is Fréchet differentiable and its derivative is representable in the integral form (2.3). The most basic method of the gradient descent type suggests moving from μ_n (the approximate solution on step n) to $\mu_{n+1} = \mu_n + \eta_n$, where $\eta = \eta_n$ minimises

$$Df(\mu_n)[\eta] = \int d_f(x, \mu_n)\eta(dx) \tag{4.1}$$

over all measures $\eta \in \tilde{\mathbb{M}}$ with the total variation ε_n, the latter being the size of step n. At every step it is important to ensure that the $\mu_n + \eta_n$ is a non-negative measure that satisfies the imposed constraints.

Explicit determination of the steepest direction η_n is generally a difficult problem. Theorem 4.1 below describes the steepest descent direction for optimisation over positive measures under several linear equality constraints

$$H_i(\mu) = \int h_i(x)\mu(dx) = a_i, \quad i = 1, \dots, k, \tag{4.2}$$

where $a_1, \dots, a_k$ are given real numbers. In vector form, $H(\mu) = \int h(x)\mu(dx) = A$, where $H = (H_1, \dots, H_k)$, $h = (h_1, \dots, h_k)$ and $A = (a_1, \dots, a_k)$. Then $DH(\mu)[\eta] = \int h(x)\eta(dx)$.

For $\mu \in \mathbb{M}$, denote by Υ_μ the family of all signed measures $\eta \in \tilde{\mathbb{M}}$ such that $\mu + \eta \in \mathbb{M}$ and $\mu + \eta$ satisfies the constraints (4.2). The family Υ_μ represents admissible directions of descent. Let $\mu|_B$ denote the restriction of a measure μ onto a Borel set B, i.e.

$$\mu|_B(\cdot) = \mu(\cdot \cap B).$$

Further, ε denotes a positive constant that controls the size of the step, i.e. the total variation of η. Recall that vectors $w_1, \dots, w_{k+1}$ are called *affinely independent* if $w_2 - w_1, \dots, w_{k+1} - w_1$ are linearly independent.

Theorem 4.1. *The minimum of $Df(\mu)[\eta]$ over all $\eta \in \Upsilon_\mu$ with $\|\eta\| \le \varepsilon$ is achieved on a signed measure $\eta = \eta^+ - \eta^-$ such that η^+ has at most k atoms, $\eta^- = \sum_{i=1}^{k+1} t_i\, \mu \mid_{B_i}$ for some $0 \le t_i \le 1$ with $t_1 + \dots + t_{k+1} = 1$ and some measurable sets B_i such that vectors $H(\mu|_{B_i})$, $i = 1, \dots, k+1$, are affinely independent.*

Proof: Let η_0 minimise $Df(\mu)[\eta]$ with respect to $\eta \in \Upsilon_\mu$ and $\mu + \eta_0$ satisfy the required conditions (4.2). By Winkler (1988, Theorem 2.1), there exists a measure η^+ with at most k atoms such that

$$H(\eta^+) = H(\eta_0^+) \quad \text{and} \quad Df(\mu)[\eta^+] = Df(\mu)[\eta_0^+].$$

Next, the set $\mathbb{A}_\mu = \{\nu \in \mathbb{M} : \mu - \nu \in \mathbb{M}\}$ is a convex linearly compact subset of $\mathbb{M}$ and H is a linear map from $\mathbb{M}$ to $\mathbb{R}^k$. By Winkler (1988, Proposition 2.1) the extreme points of the set $\mathbb{G} = \mathbb{A}_\mu \cap H^{-1}(\eta_0^-)$ are contained in the convex combinations of the type $\sum_{i=1}^{k+1} t_i\, \mu|_{B_i}$ with affinely independent $H(\mu|_{B_i})$, $i = 1, \dots, k+1$. Therefore, the maximum of a linear functional $Df(\mu)[\nu]$ over all $\nu \in \mathbb{G}$ is attained at one of such measures which can then be taken to replace η_0^-. Therefore, $\eta^+ - \eta^-$ can be taken instead of η_0 without altering the value of the derivative $Df(\mu)[\eta] = Df(\mu)[\eta_0]$. □

Corollary 4.2. *If the only constraint is $\mu(X) = a$, then the minimum of $Df(\mu)[\eta]$ over all $\eta \in \Upsilon_\mu$ with $\|\eta\| \le \varepsilon$ is achieved on a signed measure η such that η^+ is the positive measure of total mass $\varepsilon/2$ concentrated on the points of the global minima of $d_f(x, \mu)$ and $\eta^- = \mu|_{M(t_\varepsilon)} + \delta\mu|_{M(s_\varepsilon)\setminus M(t_\varepsilon)}$, where*

$$M(p) = \{x \in X : d_f(x, \mu) \ge p\},$$

and

$$t_\varepsilon = \inf\{p : \mu(M(p)) < \varepsilon/2\}, \tag{4.3}$$

$$s_\varepsilon = \sup\{p : \mu(M(p)) \ge \varepsilon/2\}. \tag{4.4}$$

The factor δ is chosen in such a way that $\mu(M(t_\varepsilon)) + \delta\mu(M(s_\varepsilon) \setminus M(t_\varepsilon)) = \varepsilon/2$.

Proof: By Theorem 4.1, we can search the optimal η among all signed measures with η^+ concentrated at a single point x_0. The contribution of this positive part to the gradient of f at direction μ is provided by $(\varepsilon/2)d_f(x_0, \mu)$ and is minimised if x_0 is a point of the set of global minima of $d_f(x, \mu)$.

The negative part η^- should be chosen to maximise $q(\eta^-) = \int d_f(x, \mu)\eta^-(dx)$ among all measures η^- dominated by μ; note that Theorem 4.1 implies that $\eta^- = c\mu|_{B_1} + (1-c)\mu|_{B_2}$ is a convex combination of μ restricted onto two sets B_1 and B_2. A necessary condition for minimum of a functional defined on a family of dominated measures (Molchanov and Zuyev 2000a, Theorem 5.1(ii)) implies that $\eta = \mu|_B$, where B should consist

of the points where $d_f(x, B)$ is as large as possible. Therefore, if the decreasing function $\mu(M(p))$ takes the value $\varepsilon/2$ for some $p = t_\varepsilon$, then $t_\varepsilon = s_\varepsilon$ and $\eta^- = \mu|_{M(t_\varepsilon)}$ provides the maximum. If $\mu(M(p))$ is discontinuous at the point t_ε then $\mu|_{M(t_\varepsilon)}$ has total mass smaller than $\varepsilon/2$ and the rest of the mass is provided by $\delta\mu|_{M(s_\varepsilon)\setminus M(t_\varepsilon)}$, leading to $\eta^- = \mu|_{M(t_\varepsilon)} + \delta\mu|_{M(s_\varepsilon)\setminus M(t_\varepsilon)}$. Both cases correspond to the form of η^- given by Theorem 4.1 with $B_0 = M(t_\varepsilon)$, $B_1 = M(s_\varepsilon)\setminus M(t_\varepsilon)$ and $t_0 = 1,\ t_1 = 0$ in the first case and $t_0 = (1-\delta),\ t_1 = \delta$ in the second. □

It is interesting to note that, without constraint $\|\eta\| \le \varepsilon$ on the total variation norm of the increment measure, the steepest direction $\eta \in \Upsilon_\mu$ that preserves the total mass a is the measure $\eta = a\delta_{x_0} - \mu$, where x_0 is a global minimum point of $d_f(x, \mu)$.

5. Steepest descent algorithm for fixed total mass problem

Corollary 4.2 that describes the steepest descent direction in the minimisation problem with a fixed total mass, gives rise to the following algorithm.

Procedure `go.steep`

Data. Initial measure μ.

Step 0. Compute $f \leftarrow f(\mu)$.

Step 1. Compute $g \leftarrow d_f(x, \mu)$. If `is.optim` $(\mu,\ g)$, stop. Otherwise, choose the step size ε.

Step 2. Compute $\mu_1 \leftarrow$ `take.step`(ε, μ, g).

Step 3. If $f_1 \leftarrow f(\mu_1) < f$, then $\mu \leftarrow \mu_1$; $f \leftarrow f_1$; and go to Step 2. Otherwise, go to Step 1.

The necessary condition for the optimum (2.6) used as a stopping rule in Step 1 means that the function $g(x) = d_f(x, \mu)$ is constant on the support of an optimal μ and takes its minimal value there. Although the support of μ on the *discrete* space X is the set $S = \{x \in X : \mu(x) > 0\}$, in practice one may wish to ignore the atoms of mass less than a certain small threshold `supp.tol`. The boolean procedure `is.optim` has the following structure.

Procedure `is.optim`

Data. Measure μ, gradient function $g(x)$, tolerance `tol`, tolerance of the support `supp.tol`.

Step 1. Compute support S of μ up to tolerance `supp.tol`.

Step 2. If $\max_{x\in S} g(x) - \min g(x) <$ `tol` return `TRUE`, otherwise return `FALSE`.

The procedure `take.step` returns the updated measure $\mu+\eta$, where η is the steepest direction increment measure with total mass 0 and total variation ε given by Corollary 4.2.

Procedure `take.step`

Data. Step size ε, measure μ, gradient function $g(x)$.

Step 0. Assign to each point $x \in X$ the mass $\mu(\{x\})$.

Step 1. Find the points of global minimum of $g(x)$ and add the total mass $\varepsilon/2$ to one of these points or spread it uniformly (or in any other manner) over these points.

Step 2. Find t_ε and s_ε from (4.3) and (4.4) and assign mass 0 to all the points of the set $M(t_\varepsilon)$, decrease the total mass of the points from $M(s_\varepsilon)\setminus M(t_\varepsilon)$ by value $\varepsilon/2 - \mu(M(t_\varepsilon))$ and return the obtained measure.

The described algorithm surely leads to a global minimum when applied to convex objective functions. In the general case it may stuck in a local minimum, the feature common for gradient algorithms applied in the context of global optimisation.

There are many possible methods suitable to choose the step size ε in Step 1 of procedure `go.steep`. Many aspects can be taken into consideration: the previous step size and/or difference between the supremum and infimum of $d_f(x, \mu)$ over the support of μ.

We mention two common methods widely used for general gradient descent algorithms (see, Polak 1997). The first one takes into account the number of steps taken along the previously computed gradient. Namely, if k is the number of times Step 2 in `go.steep` was taken before passing to Step 1, then the new value of the step size is decreased by a factor $0 < \beta < 1$, if $k = 0$; and taken to be $k\varepsilon$, if $k \ge 1$.

The second one is Armijo method (see, Polak 1997, Section 1.3.2). It defines the new step size to be $\beta^m\varepsilon$, the integer m is such that

$$f(\mu + \eta_m) - f(\mu) \le \alpha \int d_f(x, \mu)\eta_m(dx),$$

$$f(\mu + \eta_{m-1}) - f(\mu) > \alpha \int d_f(x, \mu)\eta_{m-1}(dx),$$

where $0 < \alpha < 1$ and η_m is the steepest descent measure with the total variation $\beta^m\varepsilon$ described in Corollary 4.2.

Both variants are realised in `Splus/R` library `mefista` (for MEasures with FIxed mass STeepest Ascent/descent) that can be obtained from the author's web-pages.

6. Descent algorithm for optimisation with many constraints

Although the steepest direction for optimisation with many linear constraints (4.2) is characterised in Theorem 4.1, its practical determination becomes a difficult problem. Indeed, it is easy to see that for a discrete space X (used in numerical methods) minimisation of $Df(\mu)[\eta]$ over all signed measures $\eta \in \Upsilon_\mu$ with $\|\eta\| = \varepsilon$ is a linear programming problem of dimension equal to the cardinality of X. Therefore, in the presence of many

constraints, it might be computationally more efficient to use an approximation to the exact steepest direction.

One possible approach is to fix the negative component of the increment η at every step to be proportional to the current measure μ and to choose its positive part η^+ in an optimal way. Due to Theorem 4.1, the positive part $\nu = \eta^+$ of the steepest increment measure consists of at most k atoms. We, therefore, propose to move from the current measure μ to $\mu + \eta$, where $\eta = \nu - \gamma\mu$ for some $\gamma > 0$. The new measure is thus a renormalised measure $c(\mu+\nu')$ with $c = (1-\gamma)$ and $\nu' = c^{-1}\nu$. The locations $x_1, \dots, x_k$ and the corresponding masses $p_1, \dots, p_k$ of k atoms of ν are chosen to minimise the directional derivative $Df(\mu)[\eta]$ (or, equivalently, $Df(\mu)[\nu]$) and to maintain the imposed constraints (4.2). The value of γ characterises the size of the step, although not necessarily equals the total variation of η.

Since H is linear, the constraints $H(\mu + \nu - \gamma\mu) = A = (a_1, \dots, a_k)$ are satisfied if

$$H(\nu) = \sum_{j=1}^{k} p_j h(x_j) = \gamma A,$$

that can be written in a matrix form as

$$\mathbb{H}(x_1, \dots, x_k) p^\top = \gamma A^\top$$

with $p = (p_1, \dots, p_k)$ and $\mathbb{H}(x_1, \dots, x_k) = [h_i(x_j)]_{i,j=1}^k$. By the regularity condition (2.4) the matrix $\mathbb{H}$ is invertible implying that

$$p^\top = \gamma \mathbb{H}(x_1, \dots, x_k)^{-1} A^\top. \tag{6.1}$$

Since $\eta = \nu - \gamma\mu$, the directional derivative $Df(\mu)[\eta]$ is minimised if ν minimises

$$\begin{aligned} Df(\mu)[\nu] &= \sum_{j=1}^{k} p_j d_f(x_j, \mu) \\ &= \gamma d(x_1, \dots, x_k) \mathbb{H}(x_1, \dots, x_k)^{-1} A^\top, \end{aligned} \tag{6.2}$$

where $d(x_1, \dots, x_k) = (d_f(x_1, \mu), \dots, d_f(x_k, \mu))$. The right-hand side of (6.2) is a function of k variables $x_1, \dots, x_k$ that should be minimised to find the 'best' locations for the atoms. Their masses $p_1, \dots, p_k$ are then obtained from (6.1). Note that the minimisation here is restricted to only those k-tuples $x_1, \dots, x_k$ that provide $p^\top$ with all non-negative components.

In the case of a single constraint on the total mass the described approach turns into a conventional method widely used in the optimal design literature (see, e.g., Wynn 1970). In this case $Df(\mu)[\nu]$ is minimised for ν having a single atom placed at a point of global minimum of $d_f(\cdot, \mu)$. Since this descent differs from the *steepest* descent given in Corollary 4.2, an additional analysis is necessary to ensure that the algorithm of the described kind does converge to the desired solution (see, e.g., Wu and Wynn 1978).

The descent algorithm below is based on the arguments above and have been programmed in the SPLUS and R languages. The corresponding library `medea` (for MEasure DEscent/Ascent) can be obtained from the authors' web-pages.

Procedure `go.renorm`

Data. Initial measure μ.

Step 0. Compute $f \leftarrow f(\mu)$ and for each k-tuple $(x_1, \dots, x_k)$ compute $\mathbb{H}(x_1, \dots, x_k)^{-1} A^\top$.

Step 1. Compute $g \leftarrow d_f(x, \mu)$. If `is.optim.renorm` (μ, g), stop. Otherwise, choose the step size ε.

Step 2. Compute $\mu_1 \leftarrow$ `take.step.renorm` (ε, μ, g).

Step 3. If $f_1 \leftarrow f(\mu_1) < f$, then $\mu \leftarrow \mu_1$; $f \leftarrow f_1$; and go to Step 2. Otherwise, go to Step 1.

The most resource consuming part here is the pre-calculation of the inverse of the matrix $\mathbb{H}$ in Step 0 since there are $O(|X|^k)$ values to store, where $|X|$ is the cardinality of X. This eliminates repeating the same operations at each step. However, if the calculation time is not critical but the memory requirements are, this part may be omitted with $\mathbb{H}(x_1, \dots, x_k)^{-1} A^\top$ calculated each time its value is required.

By (2.5) the function $d_f(x, \mu)$ at the optimal μ should be a linear combination

$$L(x) = L(x, u_1, \dots, u_k) = \sum_{i=1}^{k} u_k h_i(x)$$

of the functions $h_i(x)$ defining the constraints (4.2) on the support of μ and not less then this combination in other points. The corresponding estimates $\hat{u}_1, \dots, \hat{u}_k$ can be found by minimising

$$\sum_{x \in \operatorname{supp} \mu} (L(x, u_1, \dots, u_k) - d_f(x, \mu))^2 \tag{6.3}$$

and the value of (6.3) used in the stopping rule for `go.renorm` algorithm.

Procedure `is.optim.renorm`

Data. Measure μ, gradient function $g(x)$, tolerance `tol`, tolerance of the support `supp.tol`.

Step 1. Compute support S of μ up to tolerance `supp.tol`.

Step 2. Find the vector $\hat{L}$ of $u_1, \dots, u_k$ minimising (6.3).

Step 3. Calculate $\hat{g} \leftarrow (g - \hat{L} - \min(g - \hat{L}))$. If $\sum_{x\,:\,\hat{g} > \texttt{tol}} \mu(x) <$ `supp.tol` return `TRUE`, otherwise return `FALSE`.

Step 2 above can be easily realised in `Splus/R` with the help of `glm` procedure. If C denotes the matrix of constraints $C = [h_j(x_i)]_{x_i \in X, 1 \le j \le k}$ calculated at all the points of X, then $u_1, \dots, u_k$ are the regression coefficients returned by `glm` after fitting the model `g~C-1` (to use the columns of C as regressors with no intercept fitted) over all the rows that correspond to $S \subset X$.

The procedure `take.step.renorm` returns the updated renormalised measure $\mu + \nu - \gamma\mu$, where ν contains at most k atoms found to minimise (6.2).

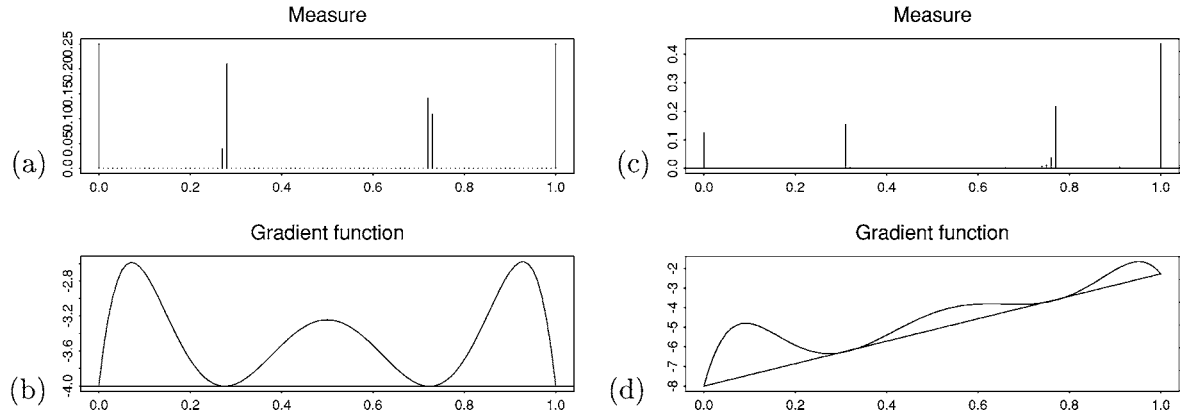

Fig. 1. *(a) the optimal design measure and (b) the corresponding gradient $d_f(x, \mu)$ as functions of x; (c) and (d) the corresponding results for the problem with the fixed barycentre*

Procedure `take.step.renorm`

Data. Step size γ, measure μ, gradient function $g(x)$, the k-tuples $Q(x_1, \ldots, x_k) = \mathbb{H}(x_1, \ldots, x_k)^{-1} A^\top$ computed for each k-tuple $(x_1, \ldots, x_k) \in X$.

Step 0. Assign to each point $x \in X$ the mass $\mu(\{x\})$.

Step 1. Find points $x_1, \ldots, x_k$ minimising $\sum_{i=1}^{k} g(x_i)\ Q_i(x_1, \ldots, x_k)$, where Q_i is the i-th component of Q, over all k-tuples $x_1, \ldots, x_k$ such that $Q(x_1, \ldots, x_k)$ contains only non-negative components.

Step 2. Find $p = (p_1, \ldots, p_k) \leftarrow \gamma Q(x_1, \ldots, x_k)$.

Step 3. Assign the value $\mu'(\{x_i\}) \leftarrow (1 - \gamma)\mu(\{x_i\}) + p_i$ to every x_i, $1 \le i \le k$. Assign $\mu'(\{x\}) \leftarrow (1 - \gamma)\mu(\{x\})$ to all other $x \in X$. Return values $\mu'(\{x\})$, $x \in X$.

It is clear that the initial measure must satisfy all the imposed constraints. But even finding such a measure in the presence of many constraints could be itself a time consuming task. Therefore, if the initial measure is not given, the procedure `go.renorm` in `medea` library will find one and start the descent from it.

7. Numerical examples

7.1. *Optimal design*

Consider the polynomial regression problem of order 4 with $X = [0, 1]$. For computational purposes, we discretise X with mesh size 0.01. Starting with the initial measure uniformly distributed over the grid points, `go.steep` algorithm arrives at the measure shown in Fig. 1(a). The gradient function shown in Fig. 1(b) clearly satisfies the optimality criterion 2.6 in the fixed total mass problem as the gradient is minimal and constant on the support of the obtained measure.

Consider now the same problem with an additional constraint being the fixed barycentre, $\int x\mu(dx) = 0.7$. The `go.renorm` algorithm yields the results shown in Fig. 1(c) and (d) that satisfies the optimality condition given by Theorem 2.1 as the gradient is affine at the support of the optimal measure. The initial measure was chosen automatically by the programme to satisfy the imposed constraints.

Detailed discussion of our algorithm in comparison with other descent algorithms suggested in the optimal design literature can be found in Molchanov and Zuyev (2000c). Further results related to constrained optimal design can be found, e.g., in Fedorov and Hackl (1997) and Pukelsheim (1983).

7.2. *Mixtures*

A descent algorithm for maximum likelihood estimation of the mixing distribution was described in Lindsay (1983). However, it uses only an approximation to the steepest descent direction by choosing the negative increment on each step proportional to the current value of the measure. Consider the problem of fitting a mixture of normal distributions with a known standard deviation to a given set of sample points.

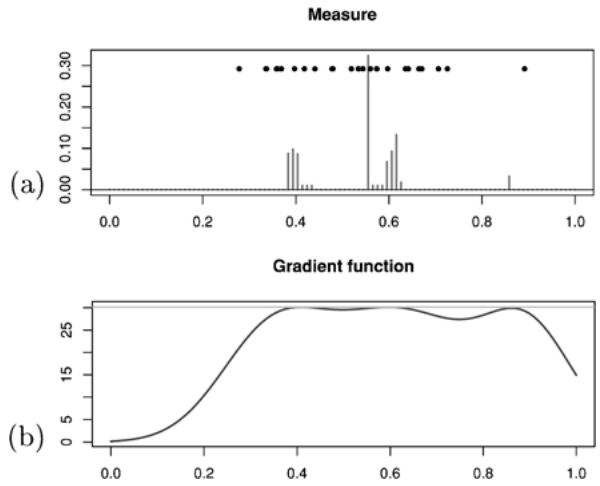

Fig. 2. *(a) the optimal mixing distribution. The observed points are shown as dots on the top line. (b) The corresponding gradient $d_f(x, \mu)$ as functions of x*

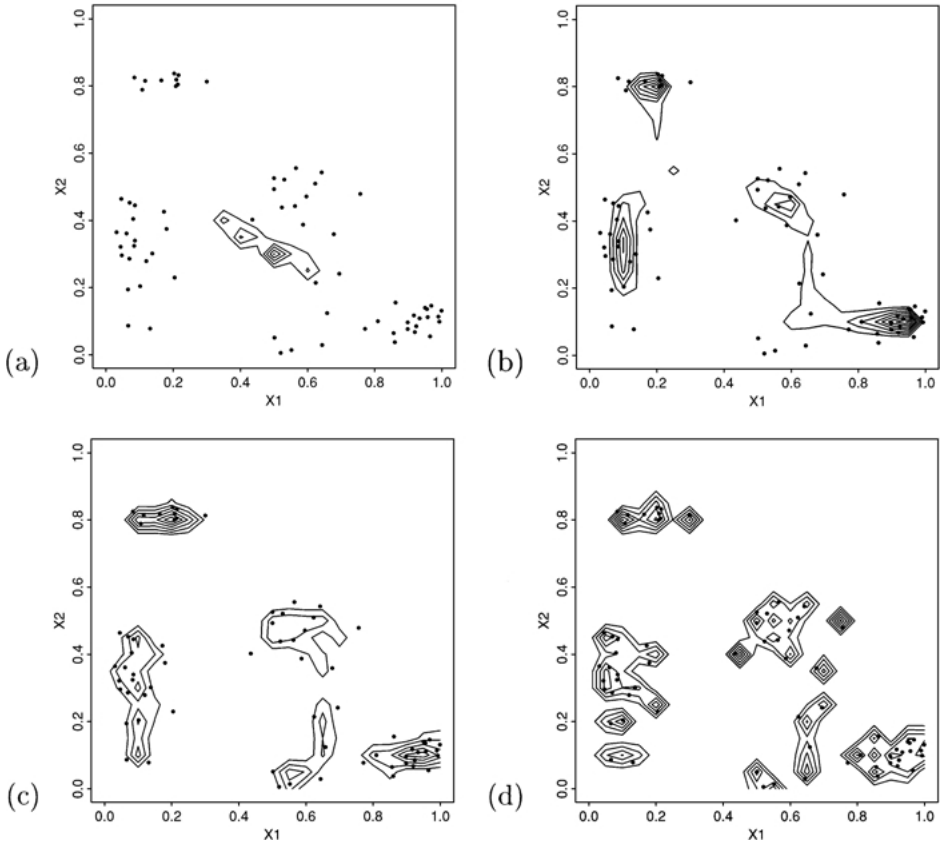

Fig. 3. *A sample of points superimposed with contour plots of the corresponding P-mean for the total mass a; (a) a = 1; (b) a = 5; (c) a = 25; (d) a = 100*

Let $p_x(\cdot)$ be the normal density with mean $x \in X = [0, 1]$ and the standard deviation σ. We will fit the optimal mixing distribution for a sample 30 observations, among them one third comes from the normal distribution with mean 0.4 and two other thirds from the normal distributions with mean 0.6. The standard deviation of all distributions is $\sigma = 0.1$. Figure 2 shows the optimal mixing distribution and the mixing measure for this case. In theory, the number of atoms in the optimal measure could be as large as the number of observation points. The obtained solution μ, however, has 15 atoms which are seen to concentrate around the true positions at 0.4 and 0.6. The total mass of μ in the neighbourhood of 0.4 is 0.3017 and 0.666 in the neighbourhood of 0.6. Observe also an artefact atom of mass 0.0323 appearing in 0.859 due to an outlier observation point at 0.892.

7.3. *P-means*

Consider a sample of $n = 69$ points in the unit square $X = [0, 1]^2$ shown in Fig. 3. These points have been sampled from 5 clusters,

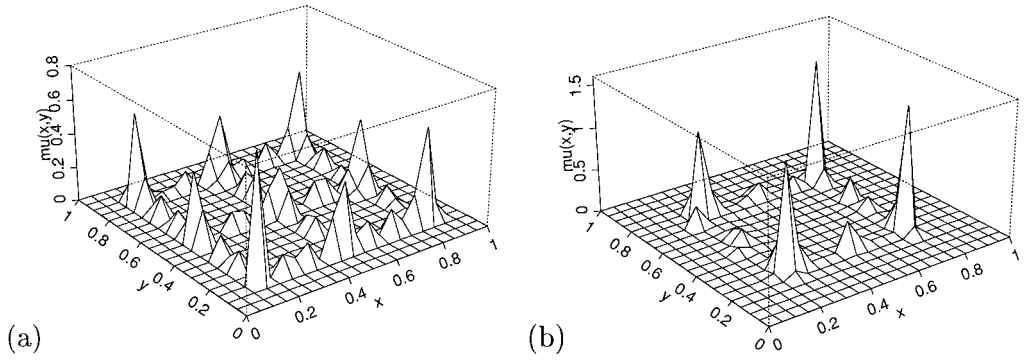

Fig. 4. *Optimal measures in the coverage problem; (a) radius r = 0.2 fixed; (b) exponentially distributed radius with mean 0.2*

each having bivariate normal distribution. As one can see, the P-mean μ with a small total mass a results in a fewer peaks of the measure density than the number of points, while if a increases substantially, the pattern of peaks of the optimal μ tends to concentrate on the point sample. It has been observed that the P-mean with the total mass equal to about the half of the sample size grasps the clustering pattern in the best way, the fact supported by Fig. 3(c).

7.4. *Maximisation of the covered volume in a Boolean model*

Let X be the unit square $[0, 1]^2$ discretised by a square grid with mesh size 0.05. The total mass of μ is fixed to 10. Figure 4 shows perspective plots of optimal measures on the grid for the Boolean model with a fixed radius $r = 0.2$ and exponentially distributed radius with mean 0.2. The initial measure was taken uniform. In both cases the optimal measure μ has a number of peaks, more explicit in the case of the exponentially distributed radii.

Other examples of functionals on Boolean models and corresponding optimisation problems are considered in Molchanov, Chiu and Zuyev (2000).

Acknowledgment

We are grateful to the referee whose suggestions led to the inclusion of the mixtures example and improvements to the presentation.

References

Atkinson A.C. and Donev A.N. 1992. Optimum Experimental Designs. Clarendon Press, Oxford.

Fedorov V.V. and Hackl P. 1997. Model-Oriented Design of Experiments, vol. 125 of Lect. Notes Statist. Springer, New York.

Hartigan J.A. 1975. Clustering Algorithms. Wiley, New York.

Hartigan J.A. and Wong M.A. 1979. A k-means clustering algorithm. Appl. Statist. 28: 100–108.

Hille E. and Phillips R.S. 1957. Functional Analysis and Semigroups, vol. XXXI of AMS Colloquium Publications. American Mathematical Society, Providence.

Kellerer H.G. 1988. Measure theoretic versions of linear programming. Math. Z. 198: 367–400.

Lai H.C. and Wu S.Y. 1994. Linear programming in measure spaces. Optimization 29: 141–156.

Lindsay B.G. 1983. The geometry of mixture likelihoods: A general theory. Ann. Statist. 11: 86–94.

McClure D.E. and Vitale R.A. 1975. Polygonal approximation of plane convex bodies. J. Math. Anal. Appl. 51: 326–358.

Molchanov I.S. 1997. Statistics of the Boolean Model for Practitioners and Mathematicians. Wiley, Chichester.

Molchanov I.S., Chiu S.N., and Zuyev S.A. 2000. Design of inhomogeneous materials with given structural properties. Phys. Rev. E 62: 4544–4552.

Molchanov I.S. and Zuyev S.A. 2000a. Tangent sets in the space of measures: With applications to variational calculus. J. Math. Anal. Appl. 249: 539–552.

Molchanov I.S. and Zuyev S.A. 2000b. Variational analysis of functionals of a poisson process. Math. Oper. Res. 25: 485–508.

Molchanov I.S. and Zuyev S.A. 2000c. Variational calculus in space of measures and optimal design. In: Atkinson A., Bogacka B., and Zhigljavsky A. (Eds.), Optimum Design 2000: Prospects for the New Millennium. Kluwer, Dordrecht, pp. 79–90.

Okabe A., Boots B., Sugihara K., and Chiu S.N. 2000. Spatial Tessellations—Concepts and Applications of Voronoi Diagrams. 2nd edn. Wiley, Chichester.

Polak E. 1997. Optimization: Algorithms and Consistent Approximations. Springer, New York.

Pukelsheim F. 1983. Optimal Design of Experiments. Wiley, New York.

Robinson S.M. 1976. First order conditions for general nonlinear optimization. SIAM J. Appl. Math. 30: 597–607.

Schneider R. 1988. Random approximations of convex sets. J. Microscopy 151: 211–227.

Winkler G. 1988. Extreme points of moment sets. Math. Oper. Res. 13: 581–587.

Wu C.-F. and Wynn H.P. 1978. The convergence of general step-length algorithms for regular optimum design criteria. Ann. Statist. 6: 1273–1285.

Wynn H.P. 1970. The sequential generation of D-optimum experimental designs. Ann. Math. Statist. 41: 1655–1664.

Zowe J. and Kurcyusz S. 1979. Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. 5: 49–62.