Measures everywhere

Numerical optimisation

Sergei Zuyev

University of Strathclyde, Glasgow, U.K.

Numeric approach

Optimal μ can rarely be obtained explicitly.

Steepest descent: Move from μ to $\mu + \eta$, where η minimises $D(\psi(\mu))[\eta]$ over $\|\eta\| = \varepsilon$.

Difficulty: $\mu + \eta$ must also satisfy all the constraints. For a fixed mass problem this implies $\eta(X) = 0$, thus $\mu + \eta$ may not be a probability measure even for very small ε!

Not really steepest descent

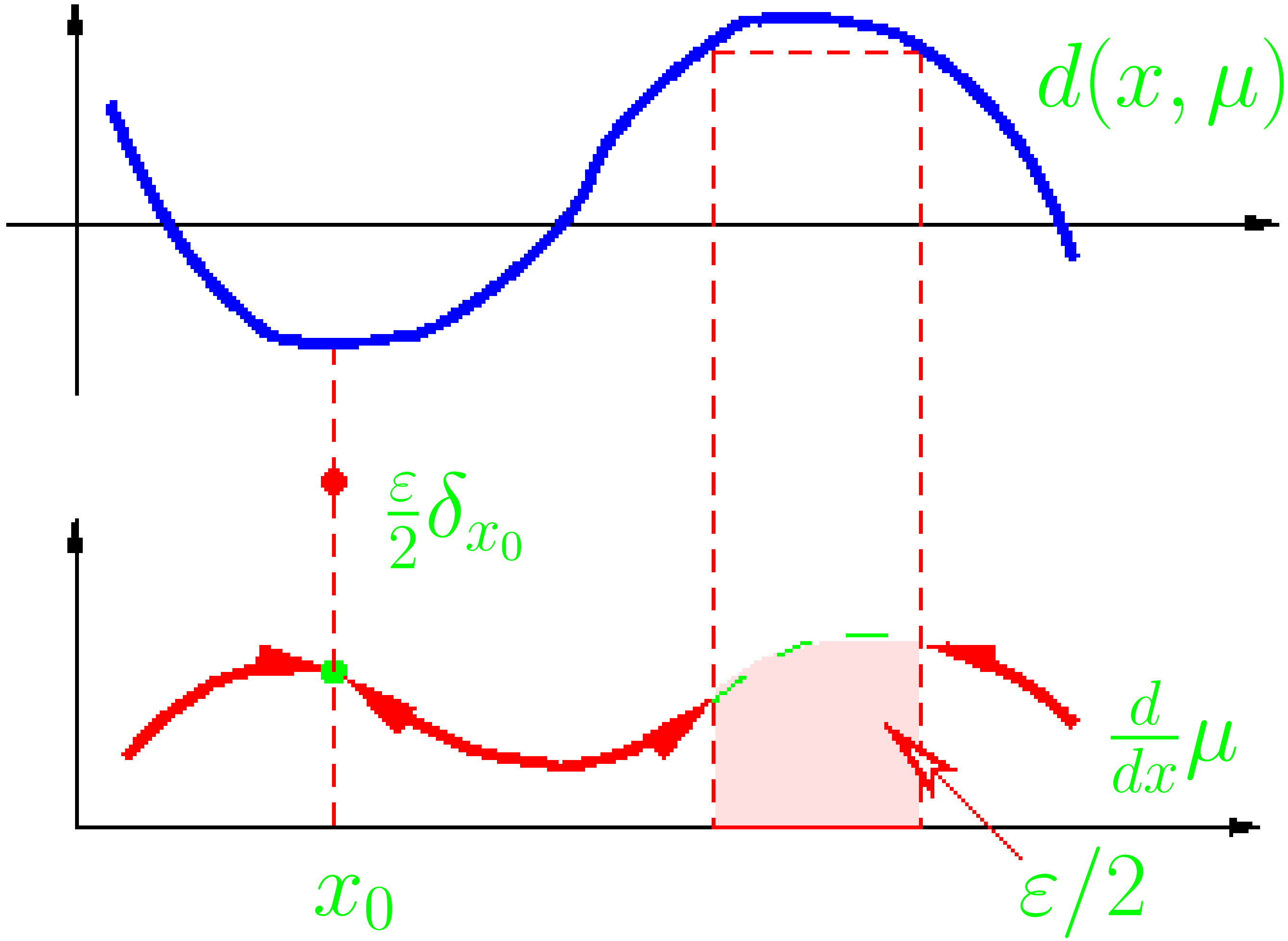
Common approach for probability measures: add ‘optimally’ a *positive* measure and rescale the result to unit mass. Specifically, move from μ to $(1-\varepsilon)\mu + \varepsilon\nu$, where $\nu \in \mathbb{M}_+$ minimises

$$\tilde{D}\psi(\mu)[\nu] = \lim_{t \downarrow 0} t^{-1}(\psi((1-t)\mu + t\nu) - \psi(\mu))\,.$$

But $\tilde{D}\psi(\mu)[\nu] = D\psi(\mu)[\nu - \mu]$. As a result:

- the direction given by $\tilde{D}$ is not the *true* steepest descent;
- convergence is slower and not evident.

True steepest descent



Theorem 1. *If the only constraint is $\mu(X)=a$, then the minimum of $D\psi(\mu)[\eta]$ over all $\|\eta\|\le\varepsilon$ such that $\mu+\eta>0$ is achieved on a signed measure η such that η^+ has total mass $\varepsilon/2$ and concentrated on the points of the global minima of the gradient function $d(x,\mu)$; and $\eta^- = \mu|_{M(t_\varepsilon)} + \varepsilon'\mu|_{M(s_\varepsilon)\setminus M(t_\varepsilon)}$, where*

$$M(p) = \{x \in X \,:\, d(x,\mu) \ge p\}\,, \qquad \textit{and}$$

$$t_\varepsilon = \inf\{p:\ \mu(M(p)) < \varepsilon/2\}\,, \tag{1}$$

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

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

Algorithm

Realised in `R/Splus` library `mefista`. Convergence follows from the conventional steepest descent theory.

Procedure `go.steep`

Data. Initial measure μ.

Step 0. Compute $y \leftarrow \psi(\mu)$.

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

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

Step 3. If $y_1 \leftarrow \psi(\mu_1) < y$, then $\mu \leftarrow \mu_1$; $y \leftarrow y_1$; and go to Step 2. Otherwise, go to Step 1.

Checking optimality

Procedure `is.optim`

Data. Measure μ, gradient function d, tolerance `tol`, tolerance of the support `supp.tol`[a].

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

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

---

[a] We may wish to ignore atoms of a very small mass

Taking a step

Procedure `take.step`

Data. Step size ε, measure μ, gradient function $d(x,\mu)$.

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

Step 1. Find the global minima of $d(x,\mu)$ and add the total mass $\varepsilon/2$ to one of these points or spread it somehow (e. g. uniformly) over these points.

Step 2. Find t_ε and s_ε from (1) and (2) and assign mass 0 to all the points of the set $M(t_\varepsilon)$, decrease the total mass of the point $M(s_\varepsilon) \setminus M(t_\varepsilon)$ by value $\varepsilon/2 - \mu(M(t_\varepsilon))$ and return the obtained measure.

Armijo method for the step size

It defines the new step size to be $\beta^m \varepsilon$, the integer m is such that

$$\psi(\mu + \eta_m) - \psi(\mu) \le \alpha \int d(x,\mu)\eta_m(dx)\,,$$

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

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

Comparison with rescaling method

- It is a *true* steepest descent. All the convergence results and properties are inherited from a general descent theory.

- Faster to run.

Example: *cubic regression through the origin*.

$$y(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sigma dw(x)\,,\ x \in [0,1]\,.$$

Find D-optimal design measure $\mu(dx)$ minimising the generalised variance:

$$\det \| \mathbf{cov}(\widehat{\beta}_i, \widehat{\beta}_j) \| = \sigma^2 \det M^{-1}(\mu)\,,$$

where

$$M(\mu) = \int f(x)^\top f(x)\,\mu(dx)\,, \quad f(x) = (x, x^2, x^3)\,,$$

is the corresponding information matrix.

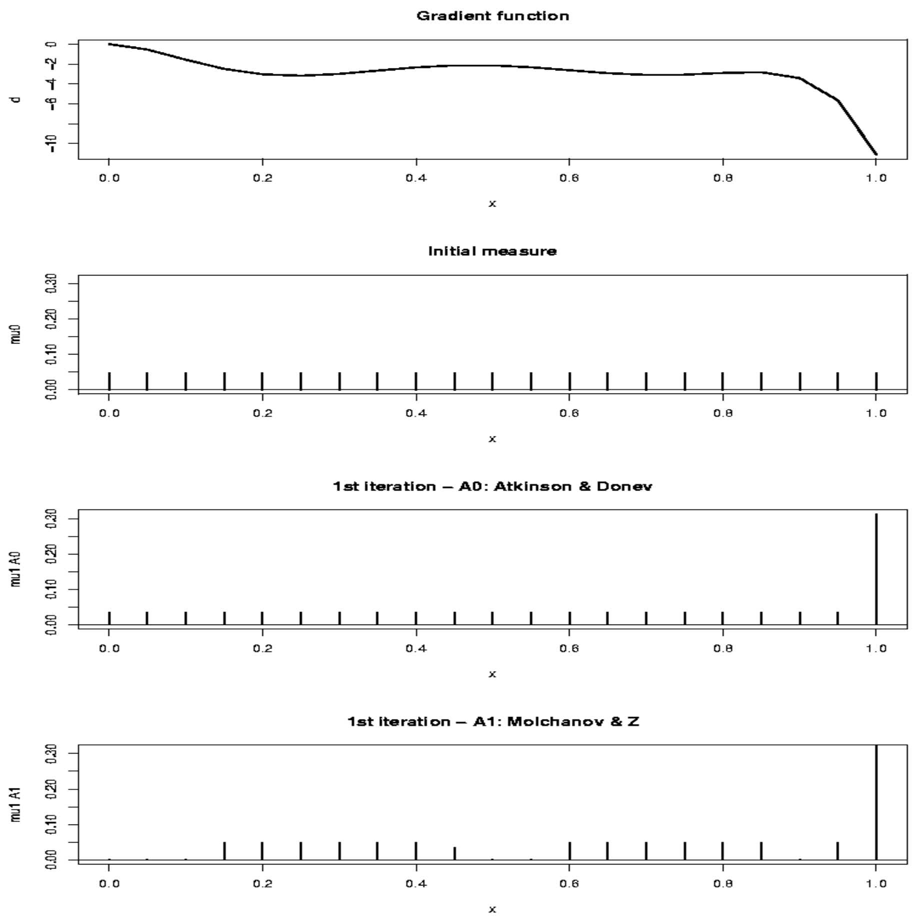


Figure 1: The first interation in $A0$ – classical renormalisation algorithm (as described in Atkinson & Donev) and $A1$ – true steepest descent algorithm

Optimisation under linear constraints

Consider the problem $\varphi(\mu) \to \inf\,,\ \mu \in \mathbb{M}_+$ under finite number of linear constraints:

$$H_i(\mu) = \int h_i(x)\mu(dx) = a_i\,, \quad i = 1, \dots, k\,, \tag{3}$$

where $a = (a_1, \dots, a_k)$ is a given vector.

Definition: 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.

General form of the increment measure

Theorem 2. *The minimum of* $D\psi(\mu)[\eta]$ *over all* $\eta \in T_{\mathbb{M}_+ \cap H^{-1}(a)}(\mu)$ *such that* $\|\eta\| \le \varepsilon$ *is achieved on a signed measure* $\eta = \eta^+ - \eta^-$*, where* η^+ *has at most* k *atoms and* $\eta^- = \sum_{i=1}^{k+1} t_i \, \mu|_{B_i}$ *for some* $0 \le t_i \le 1$ *with* $t_1 + \cdots + t_{k+1} = 1$ *and some measurable sets* B_i *such that vectors* $H(\mu|_{B_i})$*,*[a] $i = 1, \ldots, k+1$*, are affinely independent.*

Caution: Finding the optimal η here is equivalent to solving a Linear Programming Problem: not efficient. Need faster approximate solutions.

---

[a] $\mu|_B(\bullet) = \mu(\bullet \cap B)$ is the restriction of μ onto B.

In Theorem 1, $B_1 = M(t_\varepsilon),\ B_2 = M(s_\varepsilon) \setminus M(t_\varepsilon)$ and $t_1 = \varepsilon/2 - \varepsilon',\ t_2 = \varepsilon'$.

Realisation in library `medea`

Move from the current measure μ to $\mu + \eta$, where $\eta = \nu - \gamma\mu$ for some $\gamma > 0$ which has similar meaning to the step size.

Due to Theorem 2, the positive part $\nu = \eta^+ = \sum \delta_{x_i}$ of the steepest increment measure has at most k atoms. The masses $p_1, \ldots, p_k$ located at points $x_1, \ldots, x_k$ may be chosen so that to minimise the directional derivative $D\psi(\mu)[\eta]$. To satisfy the constraints $H(\mu + \nu - \gamma\mu) = a = (a_1, \ldots, a_k)$ we impose

$$H(\nu) = \sum_{j=1}^{k} p_j h(x_j) = \gamma a \,, \quad \text{or}$$

$$H(x_1, \ldots, x_k) p^\top = \gamma a^\top$$

with $p = (p_1, \ldots, p_k)$ and $H(x_1, \ldots, x_k) = [h_i(x_j)]_{i,j=1}^k$.

Thus

$$p^\top = \gamma H(x_1, \ldots, x_k)^{-1} a^\top \,. \tag{4}$$

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

$$D\psi(\mu)[\nu] = \sum_{j=1}^{k} p_j d(x_j, \mu)$$

$$= \gamma d(x_1, \ldots, x_k) H(x_1, \ldots, x_k)^{-1} a^\top \,,$$

where $d(x_1, \ldots, x_k) = (d(x_1, \mu), \ldots, d(x_k, \mu))$ are the values of the gradient function of ψ at the support points of ν.

Realisation in library `medea`

Procedure `go.steep`

Data. Initial measure μ.

Step 0. Compute $y \leftarrow \psi(\mu)$ and for each k-tuple $(x_1, \ldots, x_k)$ compute $H(x_1, \ldots, x_k)^{-1} a^\top$.

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

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

Step 3. If $y_1 \leftarrow \psi(\mu_1) < f$, update μ, y and go to Step 2. Otherwise, Step 1.

Numerical examples

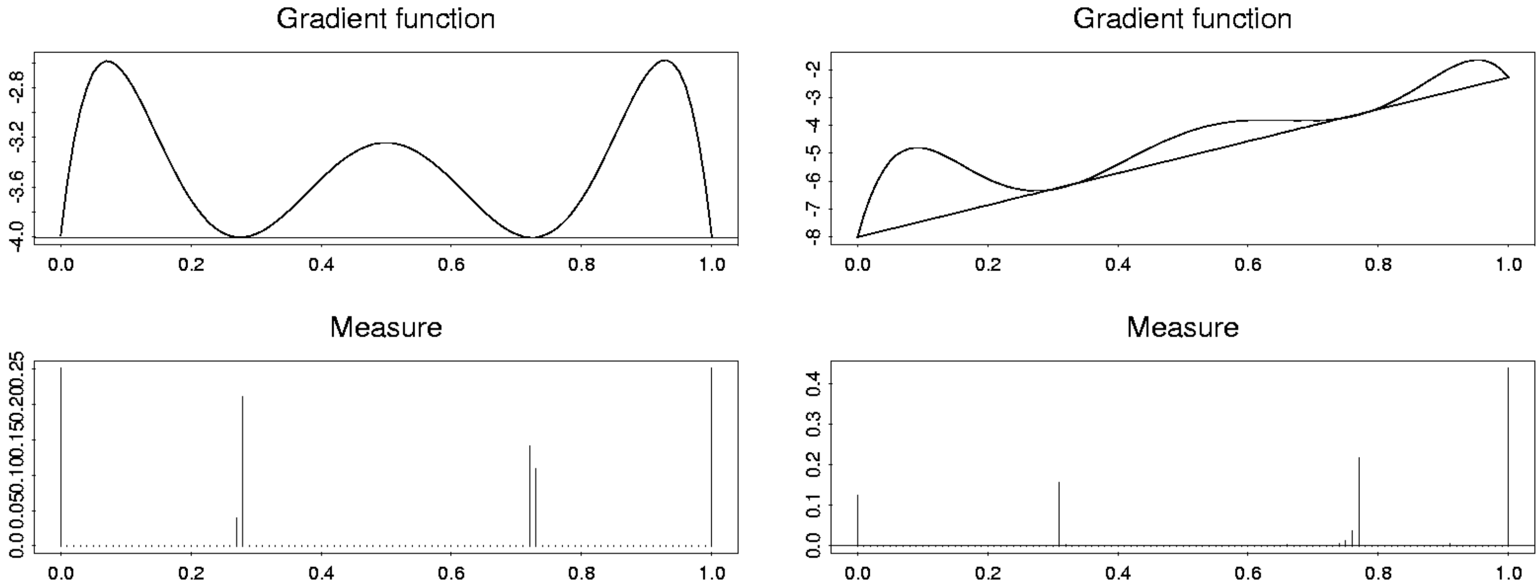


Figure 2: Optimal design measure in cubic regression through origin and with fixed barycentre = 0.7

References

- I. Molchanov and S. Zuyev. Steepest descent algorithms in space of measures. Statistics and Computing, **12**, 2002, 115–123.

❏ http://www.stams.strath.ac.uk/~sergei