Univariate Fans of Majorizers

Jan de Leeuw

First created on August 09, 2018. Last update on November 26, 2021

Abstract

dodo

Contents

```
## Warning in definition_nums(name = "majorization", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "support", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "strict", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "algorithm", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "scheme", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "algorithm2", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "fan", display = FALSE): No caption supplied.

## Warning in definition_nums(name = "common", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "parfan", display = FALSE): No caption
## supplied.

## Warning in definition_nums(name = "addfan", display = FALSE): No caption
## supplied.

## Warning in example_nums(name = "cubic", display = FALSE): No caption supplied.

## Warning in example_nums(name = "sine", display = FALSE): No caption supplied.

## Warning in example_nums(name = "lip", display = FALSE): No caption supplied.
```

```
## Warning in example_nums(name = "fan", display = FALSE): No caption supplied.

## Warning in example_nums(name = "log", display = FALSE): No caption supplied.

## Warning in example_nums(name = "nofan", display = FALSE): No caption supplied.

## Warning in example_nums(name = "norm", display = FALSE): No caption supplied.

## Warning in example_nums(name = "qchain", display = FALSE): No caption supplied.

## Warning in example_nums(name = "quartic", display = FALSE): No caption supplied.

## Warning in note_nums(name = "abbrev", display = FALSE): No caption supplied.

## Warning in note_nums(name = "set", display = FALSE): No caption supplied.

## Warning in note_nums(name = "convex", display = FALSE): No caption supplied.

## Warning in note_nums(name = "bound", display = FALSE): No caption supplied.

## Warning in note_nums(name = "differentiate", display = FALSE): No caption
## supplied.

## Warning in note_nums(name = "scheme", display = FALSE): No caption supplied.

## Warning in note_nums(name = "inf", display = FALSE): No caption supplied.

## Warning in note_nums(name = "lip", display = FALSE): No caption supplied.

## Warning in note_nums(name = "dfan", display = FALSE): No caption supplied.

## Warning in note_nums(name = "additive", display = FALSE): No caption supplied.

## Warning in note_nums(name = "dadd", display = FALSE): No caption supplied.

## Warning in note_nums(name = "ruitenburg", display = FALSE): No caption supplied.

## Warning in note_nums(name = "twopoint", display = FALSE): No caption supplied.

## Warning in note_nums(name = "finite", display = FALSE): No caption supplied.
```

```
## Warning in note_nums(name = "decreasing", display = FALSE): No caption supplied.

## Warning in note_nums(name = "interval", display = FALSE): No caption supplied.

## Warning in lemma_nums(name = "unique", display = FALSE): No caption supplied.

## Warning in lemma_nums(name = "minimum", display = FALSE): No caption supplied.

## Warning in theorem_nums(name = "algorithm", display = FALSE): No caption
## supplied.

## Warning in theorem_nums(name = "chain", display = FALSE): No caption supplied.

## Warning in theorem_nums(name = "delta", display = FALSE): No caption supplied.

## Warning in theorem_nums(name = "alpha", display = FALSE): No caption supplied.
```

Note: This is a working paper which will be expanded/updated frequently. All suggestions for improvement are welcome. The directory deleeuwpdx.net/pubfolders/fans has a pdf version, the bib file, the complete Rmd file, and the R and C code (if applicable).

1 Introduction

Majorization algorithms are popular these days. The basic idea is simple. To minimize a real valued target function $f : \mathcal{S} \Rightarrow \mathbb{R}$ on a set $\mathcal{S} \subseteq \mathbb{R}^n$ we use an iterative algorithm in which iteration $k+1$ updates $x^{(k)} \in \mathcal{S}$ to $x^{(k+1)} \in \mathcal{S}$ in two substeps. In the first substep we find a function g that lies above the target function in $\mathcal{S}$ and touches it in the current $x^{(k)}$. In the second substep we find $x^{(k+1)}$ by minimizing the majorizing function g over $\mathcal{S}$. This produces a strictly decreasing sequence of target function values $f^{(k)} = f(x^{(k)})$, which forces convergence under some natural additional conditions (D'Esopo (1959), Zangwill (1969)).

Early majorization algorithms for specific classes of problems were described by Dempster, Laird, and Rubin (1977) and De Leeuw (1977). Both papers suggest that a general class of algorithms lies behind their proposals. As a natural next step, some more general families of majorization methods were discussed in Vosz and Eckhardt (1980) and Böhning and Lindsay (1988). A general theory, inspired by both Dempster, Laird, and Rubin (1977) and De Leeuw (1977), was introduced in De Leeuw (1994) and Heiser (1995), and a much improved and expanded version is now available in book form in Lange (2016 (in press)) and (**deleeuw_B_16b?**).

Minimizing f is done by constructing majorizations g. But of course we can also maximize f by using minorizers g. Thus Lange and co-workers (for example Hunter and Lange (2004)) defined the class of **MM algorithms**, which cleverly covers both majorization-minimization and minorization-maximization. In this paper we only talk about majorization-minimization, because it is trivial to switch from one to the other anyway (by using $-f$ and $-g$).

Now for notation.

- The real numbers are $\mathbb{R}$, the positive reals are $\mathbb{R}_+$, and the vector space of n-tuples of real numbers is $\mathbb{R}^n$. The extended reals (with $\pm\infty$) are $\overline{\mathbb{R}}$.
- I have already used the notation $f : \mathcal{X} \Rightarrow \mathcal{Y}$ for a function from $\mathcal{X}$ to $\mathcal{Y}$. If $f : \mathcal{X} \otimes \mathcal{Y} \Rightarrow \mathcal{Z}$ then $f(\bullet, y) : \mathcal{X} \Rightarrow \mathcal{Z}$ for each $y \in \mathcal{Y}$. Thus $f(\bullet, y)(x) = f(x, y)$. If $f(x) = \|x\|$, for example, I also use the notation $\|\bullet\|$ for the function f. Throughout I try to distinguish between the function and the values it takes, so I avoid saying "the function f(x)=ax+b".
- Successive derivatives of $f : \mathcal{X} \Rightarrow \mathcal{Y}$ are $\mathcal{D}f, \mathcal{D}^2 f$, and so on. If the domain $\mathcal{X}$ is a subset of the real line $\mathbb{R}$ we also use f', f'', f''', f^{iv}, and so on. If $g : \mathcal{X} \otimes \mathcal{Y} \Rightarrow \mathcal{Z}$ we use $\mathcal{D}_1 g, \mathcal{D}_2 g, \mathcal{D}_{11} g = \mathcal{D}_1 \mathcal{D}_1 g$ and so on. See Spivak (1965), p. 44-45.
- The symbol $=^\Delta$ is used for definitions.
- End of proof is ■.

2 Majorization Basics

Definition 1: Suppose $f : \mathcal{S} \Rightarrow \mathbb{R}$ and $g : \mathcal{S} \Rightarrow \mathbb{R}$. Then we say g **majorizes** f on $\mathcal{S}$ if $g(x) \geq f(x)$ for all $x \in \mathcal{S}$. If $\mathcal{S}$ is all of $\mathbb{R}^n$ we usually leave out the "on $\mathcal{S}$".

Example 1: If $f : \mathbb{R} \Rightarrow \mathbb{R}$ is a non-trivial cubic and $g : \mathbb{R} \Rightarrow \mathbb{R}$ is a quadratic, then g does not majorize f on $\mathbb{R}$ and f does not majorize g on $\mathbb{R}$. Majorization of f by g would imply $g(x) - f(x) \geq 0$ for all $x \in \mathbb{R}$, and $g - f$ is a non-trivial cubic, which cannot be non-negative on the whole line. Similar for majorization of g by f.

Definition 2: If g majorizes f on $\mathcal{S}$ then the **support set** of the majorization is the set of all $x \in \mathcal{S}$ with $g(x) = f(x)$. Thus for $x \in \mathcal{S}$ not in the support set we have $g(x) > f(x)$. Elements of the support set are **support points**.

Note 1: We usually abbreviate "g majorizes f on $\mathcal{S}$ with $y \in \mathcal{S}$ a support point of the majorization" to "g majorizes f on $\mathcal{S}$ at y".

Note 2: The set $\mathcal{G}$ of all functions majorizing f on $\mathcal{S}$ is convex. If we order $\mathcal{G}$ using majorization and define functions $g \wedge h$ and $g \vee h$ by $(g \wedge h)(x) = \min(g(x), h(x))$ and $(g \vee h)(x) = \max(g(x), h(x))$ then $\mathcal{G}$ becomes an inf-complete lattice. Both the convexity and the lattice property remain true for the set $\mathcal{G}_Y$ of all functions majorizing f on $\mathcal{S}$ with a given support set Y. Both G and G_Y have as their unique minimum element the function f.

Example 2: There can be zero, one, a finite number, or an infinite number of support points of a majorization. The example in figure 1, from De Leeuw and Lange (2009), has $g(x) = x^2$ and $f(x) = x^2 - 10\sin^2(x)$. The support set are all integer multiples of π.

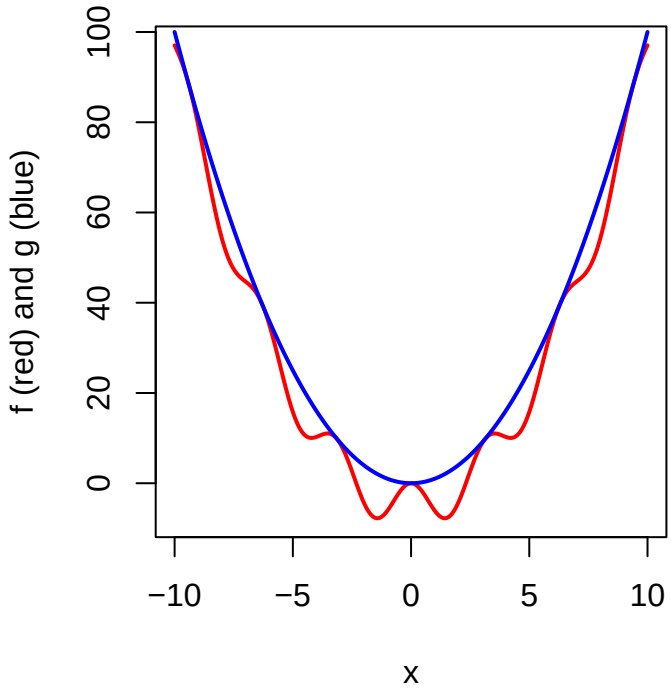


Figure 1: Support Set

Definition 3: A **strict majorization** on $\mathcal{S}$ at y is a majorization with a unique support point.

Note 3: If f is convex, then g with $g(x) = f(y) + z'(x-y)$ majorizes f at y, for any z in the subgradient $\partial f(y)$. If f is strictly convex the majorization is strict.

Note 4: If f is two times differentiable, and there is a B such that $B - \mathcal{D}^2 f(x)$ is positive semi-definite for all x then g with $g(x) = f(y) + (x-y)'\mathcal{D}f(y) + \frac{1}{2}(x-y)'B(x-y)$ majorizes f at y.

Note 5: If a differentiable g majorizes a differentiable f on $\mathbb{R}$ then $\mathcal{D}f(y) = \mathcal{D}g(y)$ at any support point. If a twice differentiable g majorizes a twice differentiable f on $\mathbb{R}$ then $\mathcal{D}^2 g(y) \succeq \mathcal{D}^2 f(y)$ at any support point, i.e. $\mathcal{D}^2 g(y) - \mathcal{D}^2 f(y)$ is positive semi-definite. This is because the differentiable function $h = g - f$ has a minimum, equal to zero, at any support point. If majorization is strict we have $\mathcal{D}^2 g(y) \succ \mathcal{D}^2 f(y)$.

3 Majorization Algorithm

Definition 4: A **majorization algorithm** is an iterative algorithm intended to minimize f over $x \in \mathcal{S}$. Iteration k starts with a current $x^{(k)} \in \mathcal{S}$. Select a g that majorizes f on $\mathcal{S}$ at $x^{(k)}$ and find $x^{(k+1)} \in \mathcal{S}$ such that $g(x^{(k+1)}) < g(x^{(k)})$. If there is no $x \in \mathcal{S}$ with $g(x) < g(x^{(k)})$ the algorithm stops.

The key to why majorization algorithms work (i.e. converge) is the following result.

Theorem 1: Suppose g majorizes f on $\mathcal{S}$ at y. Then, for all $x \in \mathcal{S}$, $g(x) < g(y)$ implies $f(x) < f(y)$.

Proof: $f(x) \leq g(x)$ by majorization, $g(x) < g(y)$ by assumption, and $g(y) = f(y)$ by support. Thus we have the **sandwich inequality** $f(x) \leq g(x) < g(y) = f(y)$. If majorization is strict this becomes $f(x) < g(x) < g(y) = f(y)$. But even if x is a second support point of the majorization we still have $f(x) = g(x) < g(y) = f(y)$. ■

Note 6: Definition 4 does not tell us how to select majorizations. In that sense it is an incomplete definition, which makes it impossible to study the properties of the algorithm. To actually get an implementation going, we need a more complete definition.

Definition 5: A **majorization scheme** for $f : \mathcal{S} \Rightarrow \mathbb{R}$ on $\mathcal{S}$ is a function $g : \mathcal{S} \times \mathcal{S} \Rightarrow \mathbb{R}$ such that

- $g(x, y) \geq f(x)$ for all $x, y \in \mathcal{S}$,
- $g(x, x) = f(x)$ for all $x \in \mathcal{S}$.

In other words, for each $y \in \mathcal{S}$ the function $g(\bullet, y)$ majorizes f on $\mathcal{S}$ at y.

Definition 6: [**Redone**] Suppose g is a majorization scheme for f on $\mathcal{S}$. In a **majorization algorithm** iteration k starts with a current $x^{(k)} \in \mathcal{S}$. We then choose $x^{(k+1)} \in \mathcal{S}$ such that $g(x^{(k+1)}, x^{(k)}) < g(x^{(k)}, x^{(k)})$. The **sandwich inequality** becomes

$$f^{(k+1)} \leq g(x^{(k+1)}, x^{(k)}) < g(x^{(k)}, x^{(k)}) < f(x^{(k)}).$$

If there is no $x \in \mathcal{S}$ with $g(x, x^{(k)}) < g(x^{(k)}, x^{(k)})$ the algorithm stops.

Note 8: Not every majorization scheme leads to a useful majorization algorithm. The function g with $g(x, y) = f(y) + \alpha|x - y|$ is a majorization scheme for f for any $\alpha > 0$. But it is impossible to choose $x^{(k+1)}$ such that $g(x^{(k+1)}, x^{(k)}) < g(x^{(k)}, x^{(k)})$, so the algorithm stops immediately at any initial solution $x^{(0)}$.

4 Fans of Functions

Definition 7: A **fan** on $\mathcal{S}$ at $y \in \mathcal{S}$ is a function $g : \mathcal{S} \otimes \mathcal{A} \Rightarrow \mathbb{R}$, with $\mathcal{A}$ a real interval, such that

- $g(y, \bullet)$ is a constant function,
- If $\alpha < \beta$ then $g(x, \alpha) < g(x, \beta)$ for all $x \neq y$.

Thus for $\alpha < \beta$ we have $g(\bullet, \beta)$ strictly majorizing $g(\bullet, \alpha)$ at y.

Note 7: Suppose

$$g(x, 0) \triangleq \inf_{\alpha \in \mathcal{A}} g(x, \alpha) > -\infty.$$

Then $g(\bullet, \alpha)$ strictly majorizes $g(\bullet, 0)$ for every $\alpha \in \mathcal{A}$.

Note 9: Suppose $g(\bullet, \alpha)$ is differentiable at y for all α. Then $g(\bullet, \beta) - g(\bullet, \alpha)$ has either a minimum or a maximum at y, and thus $\mathcal{D}_1 g(y, \alpha) = \mathcal{D}_1 g(y, \beta)$. Thus all $g(\bullet, \alpha)$ have the same tangent at y. If the $g(\bullet, \alpha)$ are twice differentiable and $\alpha < \beta$ then $\mathcal{D}_{11} g(y, \alpha) \preceq \mathcal{D}_{11} g(y, \beta)$ in the Loewner order.

Example 4: Figure 2 is an example of a fan that is quadratic in x at $y = 3$ and linear in $\alpha \in \mathcal{A}$, with

$$g(x, \alpha) = 1 + 2(x - 3) + \frac{1}{2}\alpha(x - 3)^2.$$

Figure 2 plots the quadratic functions $g(\bullet, \alpha)$ for $\alpha = 1, \cdots, 10$. They have their minimum at $3 - 2/\alpha$, with minimum value $1 - 2/\alpha$. The common tangent is the blue line $g(\bullet, 0) = 1 + 2(x - 3)$.

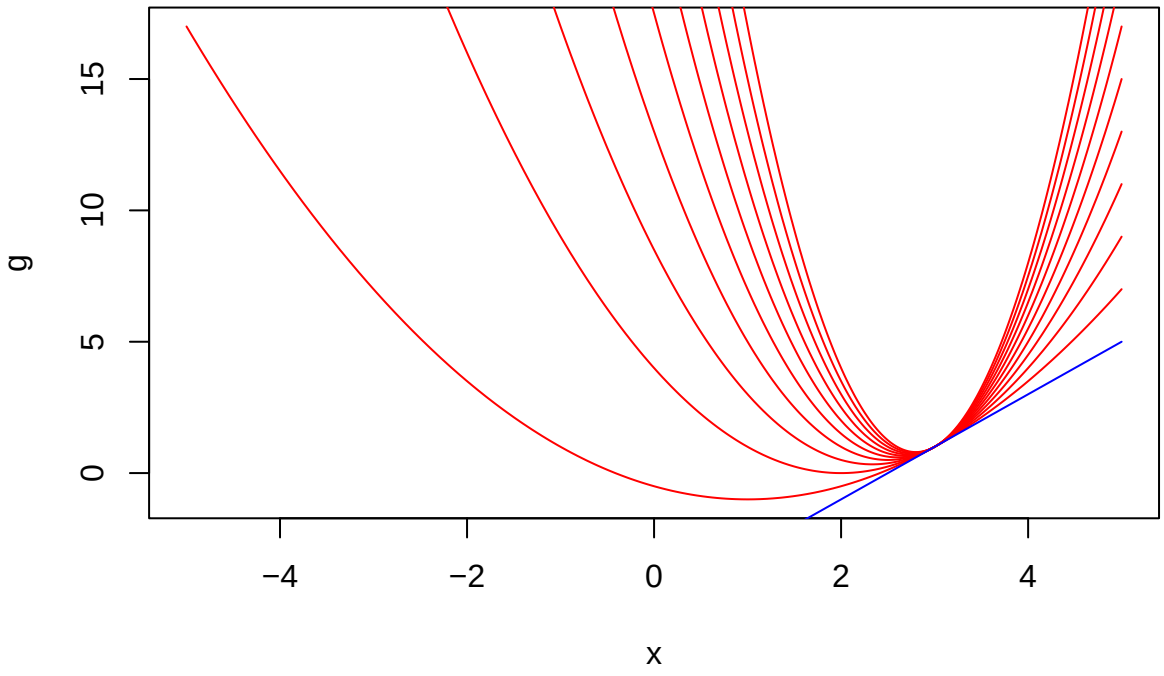


Figure 2: Fan, Quadratic in x

The linear functions $g(x, \bullet)$ are plotted in figure 3, for x taking 50 values between -10 and $+10$. The blue line is the function $\min_x g(x, \alpha) = 1 - 2/\alpha$.

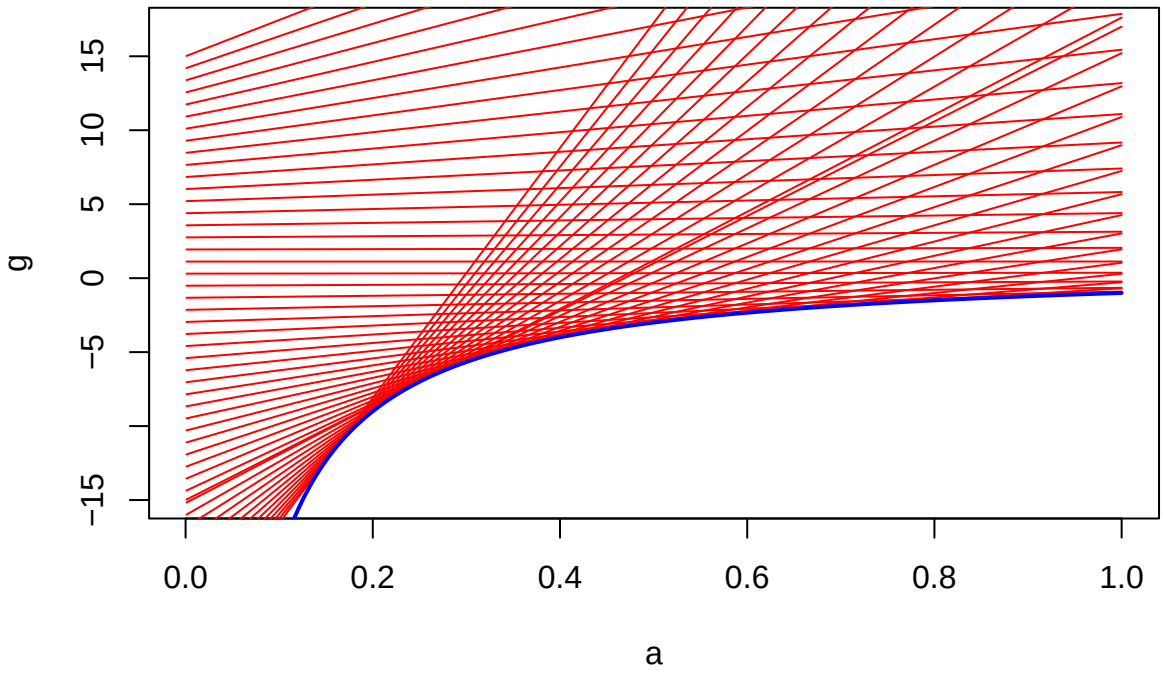


Figure 3: Fan, Linear in alpha

Example 5: This is a minor modification of example 4. We use

$$g(x, \alpha) = 1 + 2(x - 3) + \frac{1}{2}\log(\alpha)(x - 3)^2.$$

This fan is still quadratic in x, but for $0 < \alpha < 1$ the quadratics are concave. Also for this example there is no $g(x, 0)$, because $\inf_{\alpha>0} g(x, \alpha) = -\infty$ for all $x \neq 3$.

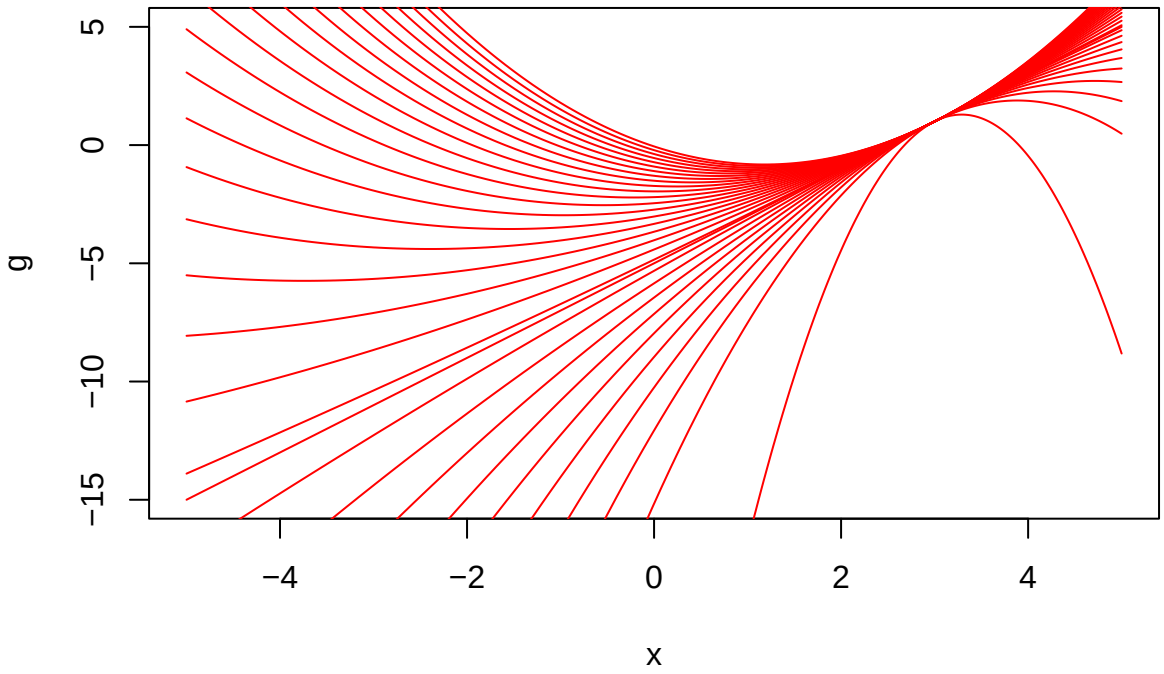


Figure 4: Fan, Logarithmic in alpha

Definition 8: If $g : \mathcal{S} \times \mathcal{A} \Rightarrow \mathbb{R}$ then $y \in \mathcal{S}$ is a **common point** of g if $g(y, \alpha) = g(y, \beta)$ for all α and β in $\mathcal{A}$. i.e. if $g(y, \bullet)$ is a constant function on $\mathcal{A}$. Thus if g is a fan on $\mathcal{S}$ at y, then y is a common point of the fan.

Theorem 5: A fan cannot have more than one common point.

Proof: Suppose g is a fan on $\mathcal{S}$ at y and $z \in \mathcal{S}$ is a second common point. Then $g(z, \alpha) = g(z, \beta)$, contradicting that either $g(x, \alpha) < g(x, \beta)$ for all $x \neq y$ or $g(x, \beta) < g(x, \alpha)$ for all $z \neq y$. ■

Example 6: Suppose the function $g : \mathbb{R} \otimes \mathcal{A} \Rightarrow \mathbb{R}$ consists of the quartics $g(\bullet, \alpha)$ with $g(x, \alpha) = \alpha(x-1)^2(x+1)^2$. Then g has the two common points ± 1 and thus g is not a fan.

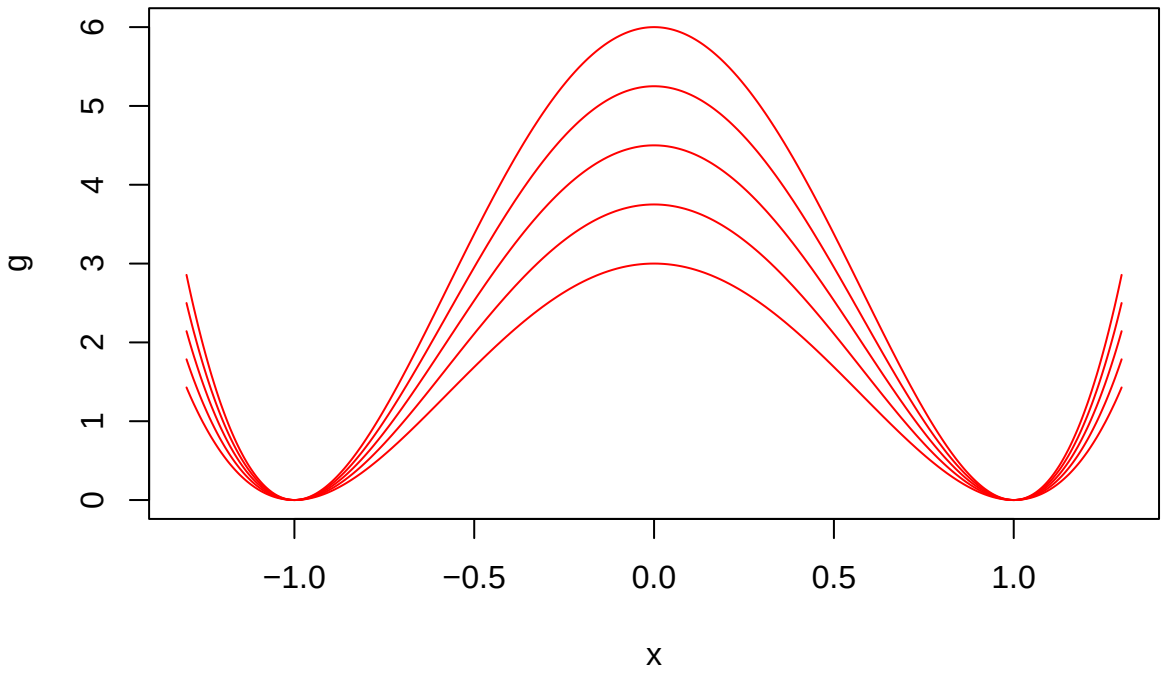


Figure 5: This is not a Fan

Note 10: The fan in figures 2 and 3, like most of the fans considered in this paper, actually has additional structure, captured in the following definition.

Definition 10: An **additive fan** on $\mathcal{S}$ at y is a fan of the form $g(x, \alpha) = p(x) + \alpha q(x)$ with $q(x) > 0$ for all $x \neq y$.

Note 11: In an additive fan at y, since $g(y, \alpha)$ is the same for all $\alpha \in \mathcal{A}$, it follows that $q(y) = 0$ and $g(y, \alpha) = p(y)$. Suppose the additive fan is differentiable. Then, in the same way, $\mathcal{D}_1 g(y, \alpha)$ does not depend on α, and thus $\mathcal{D}q(y) = 0$ and $\mathcal{D}_1 g(y, \alpha) = \mathcal{D}p(y)$. Since q has a strict local minimum at y we also have $\mathcal{D}^2 q(y) \succ 0$.

5 Majorizing Fans

Definition 11: A **majorizing fan** on $\mathcal{S}$ at y is a fan $g : \mathcal{S} \otimes \mathcal{A} \Rightarrow \mathbb{R}$ on $\mathcal{S}$ such that

- For all $\alpha \in \mathcal{A}$ the function $g(\bullet, \alpha)$ majorizes f on $\mathcal{S}$ at y.

Thus the common point of the fan is the support point of all majorizations in the fan.

Note 12: Our key result for majorizing fans is a generalization of a result first proved by Van Ruitenburg (2005). Our version of Van Ruitenberg's result does not suppose any particular functional form for the majorizations in the fan, while Van Ruitenburg (2005) uses a restricted class of polynomials. Our general definition is not even restricted to univariate functions.

Theorem 2: Suppose g is a majorizing fan for f on $\mathcal{S}$ with common support point y. There can be at most one $g(\bullet, \alpha)$ with more than one support point. This g, if it exists, is the minimum element of the fan.

Proof: From the following two lemmas. ■

Lemma 1: Suppose g is a fan majorizing f on $\mathcal{S}$ at y. If $g(\bullet, \alpha)$ has a second support point $u \neq y$ and $g(\bullet, \beta)$ has a second support point $v \neq y$, then $\alpha = \beta$.

Proof: Suppose $g(\bullet, \alpha)$ strictly majorizes $g(\bullet, \beta)$. Then $g(u, \beta) < g(u, \alpha) = f(u)$ and thus $g(\bullet, \beta)$ does not majorize f. If $g(\bullet, \beta)$ strictly majorizes $g(\bullet, \alpha)$ then $g(v, \alpha) < g(v, \beta) = f(v)$ and thus $g(\bullet, \alpha)$ does not majorize f. The contradictions proves $g(\bullet, \alpha) = g(\bullet, \beta)$. ■

Lemma 2: Suppose g is a fan majorizing f on $\mathcal{S}$ at y. If $g(\bullet, \alpha)$ has a second support point $z \neq y$ then $g(\bullet, \alpha)$ is majorized strictly by all $g(\bullet, \beta)$ with $\beta \neq \alpha$, and is consequently the minimum point of the fan.

Proof: Suppose $g(\bullet, \alpha)$ strictly majorizes $g(\bullet, \beta)$. Then $g(z, \beta) < g(z, \alpha) = f(z)$ and thus $g(\bullet, \beta)$ does not majorize f. Consequently $g(\bullet, \beta)$ strictly majorizes $g(\bullet, \alpha)$. ■

Note 13: Note that we have not shown that in a majorizing fan a majorization with more than one support point always exists. And neither have we shown that having two or more support points is necessary for minimality in the fan.

Example 10: The functions $g(\bullet, \alpha)$ with $g(x, \alpha) = \alpha x^2$ are a majorizing fan for f with $f(x) = -x^2$ at $y = 0$. For each α there is only a single support point. $g(\bullet, 0) \equiv 0$ is the minimal element.

Example 7: Suppose g has $g(x, \alpha) = f(x) + \alpha\|x - y\|$ for any norm $\| \bullet \|$. For $\alpha \geq 0$ this is a majorizing fan of f with common support point y and minimum element f. If $g(\bullet, \alpha)$ has a second support point $z \neq y$ then we must have $g(z, \alpha) = f(z) + \alpha\|z - y\| = f(z)$ and thus $\alpha = 0$, and $g(\bullet, \alpha) = f$.

Example 8: Suppose g with

$$g(x) = f(y) + f'(y)(x - y) + \frac{1}{2}\alpha(x - y)^2$$

is a majorizing fan for the differentiable f at y. If for some α the majorization has a second support point then it is the minimum element of the majorizing fan.

Note 13: If g with $g(x,\alpha) = p(x) + \alpha q(x)$ is an additive majorizing fan for f on $\mathcal{S}$ at y then $p(y) = f(y)$ and $\mathcal{D}p(y) = \mathcal{D}f(y)$. Since $\mathcal{D}_{11}g(y,\alpha) \succeq \mathcal{D}^2 f(y)$ we have

$$\alpha \mathcal{D}^2 q(y) \succeq \mathcal{D}^2 f(y) - \mathcal{D}^2 p(y),$$

which is particularly interesting in the one-dimensional case, where it becomes

$$\alpha \geq \frac{f''(y) - p''(y)}{q''(y)}.$$

In the quadratic fan of example 8 this simply becomes $\alpha \geq f''(y)$.

6 Checking Majorization

Theorem 3: Suppose g is a fan on $\mathcal{S}$ at y. Define $\delta : \mathcal{A} \Rightarrow \overline{\mathbb{R}}$ by

$$\delta(\alpha) \triangleq \inf_{x \in \mathcal{S}} \{g(x,\alpha) - f(x)\}.$$

- If g is majorizing fan for f on $\mathcal{S}$ at y then $\inf_{\alpha \in \mathcal{A}} \delta(\alpha) = 0$.
- If f is continuous and g is jointly continuous on $\mathcal{S} \times \mathcal{A}$ then $\inf_{\alpha \in \mathcal{A}} \delta(\alpha) = 0$ is also sufficient for g to be a majorizing fan.

Proof: For a majorizing fan the minimum of $g(x,\alpha) - f(x)$ over $\mathcal{S}$ is attained at y, and $\delta(\alpha) = 0$ for all $\alpha \in \mathcal{A}$. For any fan $g(y,\alpha) - f(y) = 0$ and thus $\delta(\alpha) \leq 0$. If $\inf_{\alpha \in \mathcal{A}} \delta(\alpha) < 0$ there is at least one $\alpha \in \mathcal{A}$ and one $x \in \mathcal{S}$ such that $g(x,\alpha) - f(x) < 0$, and thus g is not a majorizing fan. ■

Note 14: δ may take the value $-\infty$. But δ is finite-valued if $\mathcal{S}$ is compact and the majorizer and majorant are continuous, or if we can assume that $g(x,\alpha) - f(x)$ attains its minimum in $\mathcal{S}$ for all $a \in \mathcal{A}$.

Note 17: Because $g(x,\alpha) < g(x,\beta)$ if $\alpha < \beta$ the function δ is increasing, and because of continuity $\inf_{\alpha \in \mathcal{A}} \ \delta(\alpha) = \lim_{\alpha \downarrow \underline{a}} \ \delta(\alpha) = \delta(\underline{a})$.

Note 16: For any fan

$$\mathcal{A} \triangleq \{\alpha \mid \delta(\alpha) \geq 0\}$$

then $\mathcal{A} = [\underline{\alpha}, +\infty)$.

Theorem 4: Suppose g with $g(x) = p(x) + \alpha q(x)$ is an additive fan on $\mathcal{S}$ at y. Define

$$\underline{\alpha} \triangleq \sup_{x \in \mathcal{S} \setminus y} \frac{p(x) - f(x)}{q(x)} < +\infty$$

Proof: We have majorization if $p(x) + \alpha q(x) - f(x) \geq 0$ or

$$\alpha \geq \frac{f(x) - p(x)}{q(x)}$$

for all $x \in \mathcal{S}$ with $x \neq y$, which is equivalent to the condition in the theorem. ■

Two points

$$g(z, \alpha) - f(z) = 0$$
$$\mathcal{D}_1 g(z, \alpha) = f'(z)$$

7 Additive Taylor Fans

7.1 Even-order Taylor Polynomials

For r odd, i.e. $r + 1$ even, we can use the additive fan

$$g(x, a) = \sum_{s=0}^{r} \frac{1}{s!} f^{(s)}(y)(x - y)^s + \frac{1}{(r+1)!} \alpha (x - y)^{r+1}.$$

Van Ruitenburg (2005) considers the additive fan

$$g(x, a) = f(y) + f'(y)(x - y) + \frac{1}{(r+1)!} a (x - y)^{r+1}.$$

I see no reason to drop the terms of degree 2 to r, although of course for these intermediate terms we are not forced to use the $f^{(s)}(y)$ as coefficients. In fact, including degree 2 and higher allows us to design majorization algorithms with faster convergence.

7.2 Odd-order Taylor Polynomial

$$g(x) = \sum_{s=0}^{r} \frac{1}{s!} f^{(s)}(y)(x - y)^s + \frac{1}{(r+1)!} a |x - y|^{r+1}.$$

8 Univariate Examples

8.1 A Quartic with a Quadratic Fan

Suppose f is the quartic

$$f(x) = f(y) + f'(y)(x - y) + \frac{1}{2} f''(y)(x - y)^2 + \frac{1}{6} f'''(y)(x - y)^3 + \frac{1}{24} f^{iv}(x - y)^4$$

and we majorize f at y with the additive quadratic fan

$$g(x,a) = f(y) + f'(y)(x-y) + \frac{1}{2}a(x-y)^2.$$

$$\underline{\alpha} = \sup_x \frac{f(x) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2} = \sup_x f''(y) + \frac{1}{3}f'''(y)(x-y) + \frac{1}{12}f^{iv}(x-y)^2$$

If $f^{iv} > 0$ we have $\underline{\alpha} = +\infty$, and the fan does not majorize f at y. If $f^{iv} < 0$ the maximum is attained at

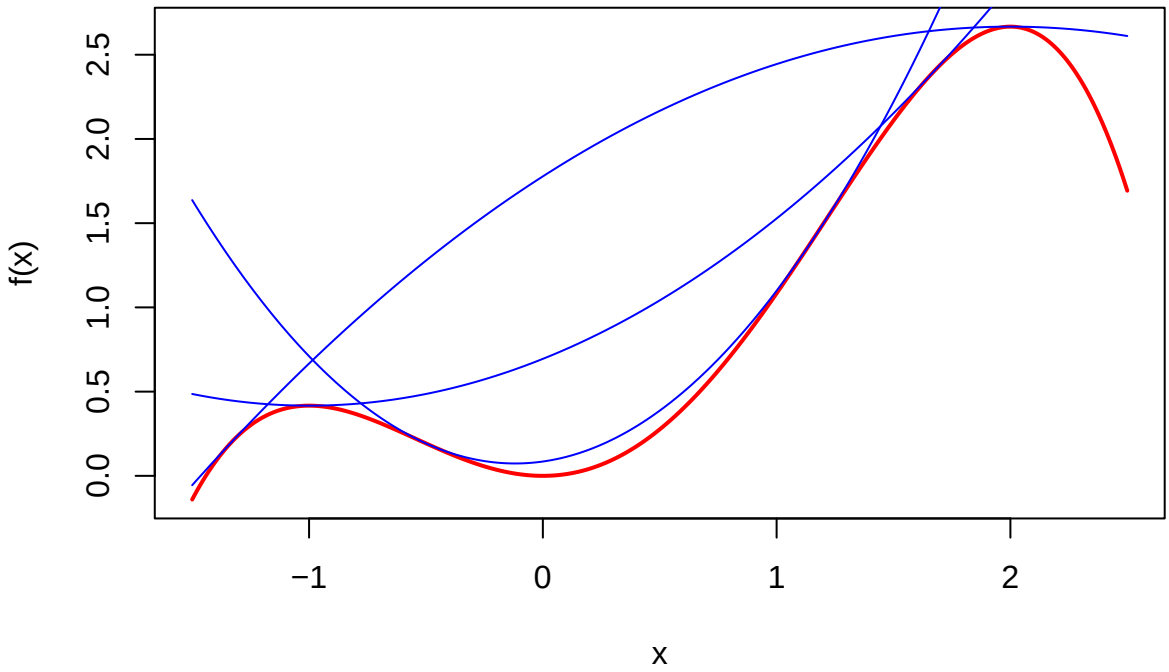
$$\hat{x} - y = -\frac{2f'''(y)}{f^{iv}}$$

and is equal to

$$\underline{a} = f''(y) - \frac{1}{3}\frac{(f'''(y))^2}{f^{iv}}$$

There is no guarantee that $\underline{a} > 0$, so the sharp majorizing quadratic may be concave, in which case it does not have a minimum. The algorithm indicates, appropriately, that $\inf_x f(x) = -\infty$. The convergence rate of the majorization algorithm at a local minimum x is

$$\kappa = 1 - \frac{f''(y)}{\underline{a}} = \frac{f^{iv}(f'''(y))^2}{3f''(y) - f^{iv}(f'''(y))^2}$$



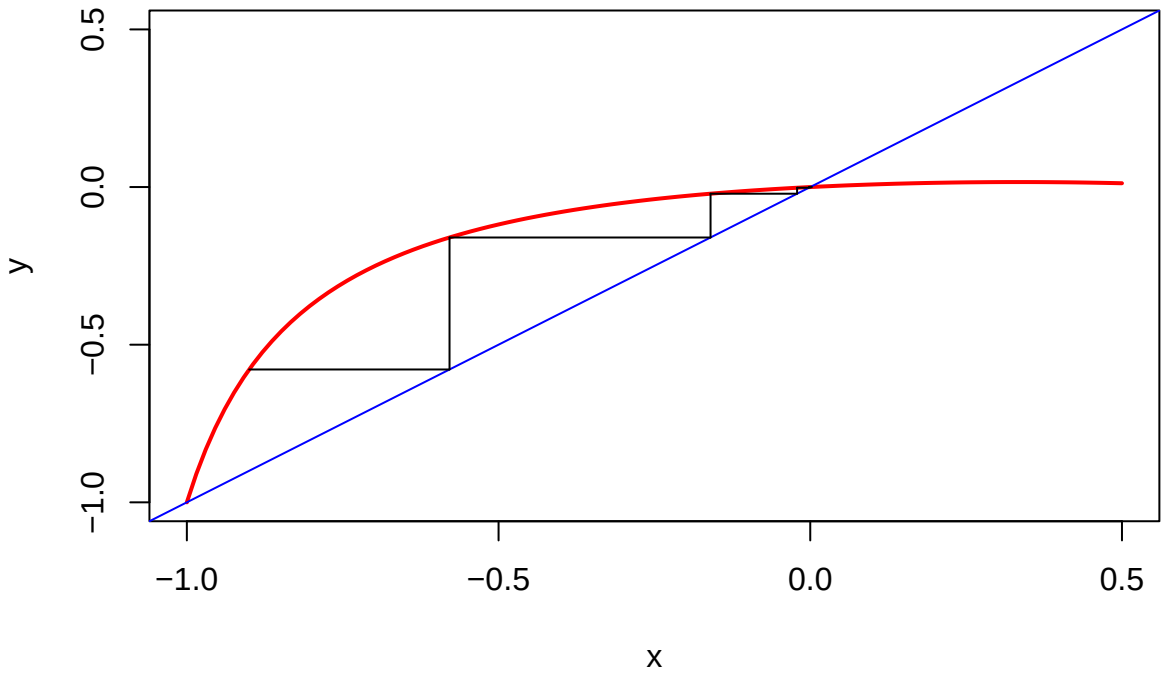
For the quartic with $f(x) = x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4$ we show the sharp quadratic majorizations at $y = -1$, $y = -.5$, and $y = 2$. For $y = -1$ the second support point is at 1.6666666667, for $y = -.5$ it is at 1.1666666667, and for $y = 2$ it is at -1.3333333333. The convergence rate at the local minimum zero is 0.1, which is plenty fast. Figure xxx also illustrates things that can go wrong. If we start the algorithm at $1\frac{2}{3}$, then the successor is -1, and the algorithm

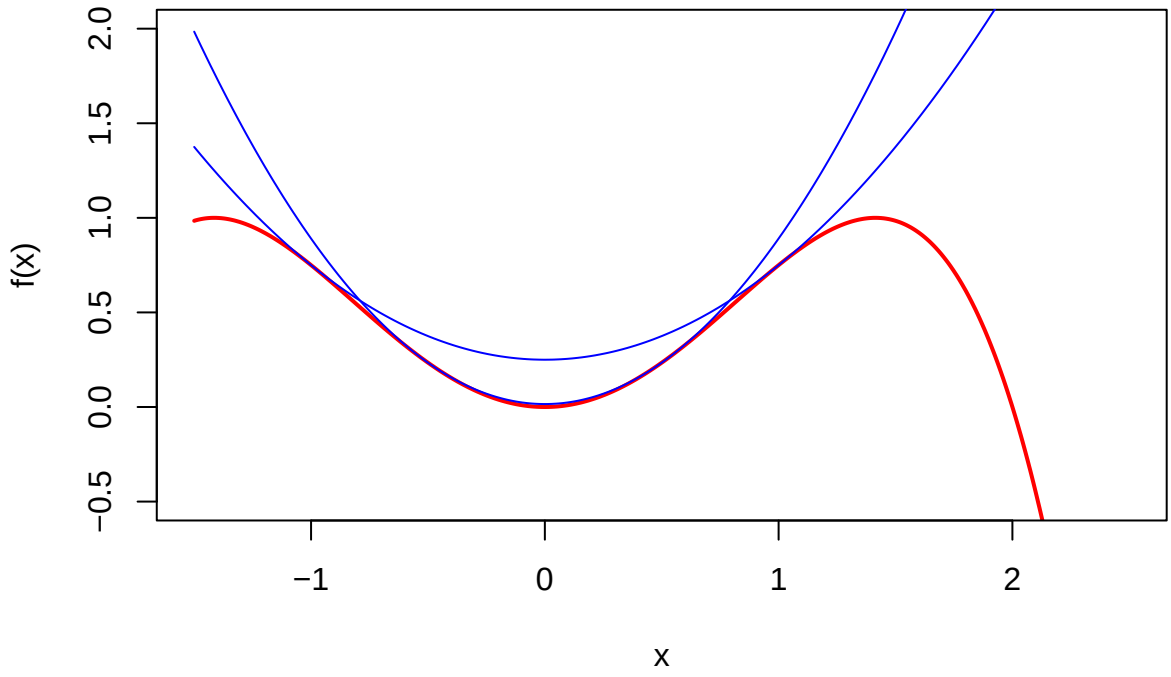
stops at that local maximum. If we start the algorithm at 2 the majorizer is concave and does not have a minimum.

```
y<--.9
for (i in 1:10) {
  z <- succ(y)
  print(c(y,z))
  y <- z
}
```

```
## [1] -0.9000000000 -0.5786593707
## [1] -0.5786593707 -0.1599667718
## [1] -0.1599667718 -0.0210900511
## [1] -0.021090051099 -0.002190018272
## [1] -0.002190018272 -0.000219866182
## [1] -2.198661820e-04 -2.199532066e-05
## [1] -2.199532066e-05 -2.199619150e-06
## [1] -2.199619150e-06 -2.199627859e-07
## [1] -2.199627859e-07 -2.199628730e-08
## [1] -2.199628730e-08 -2.199628817e-09
```

```
cobwebPlotter(-.9,succ,-1,.5,itmax=100)
```





If we modify f to an even biquadratic by leaving out the third order term, i.e. $f(x) = x^2 - \frac{1}{4}x^4$. The convergence rate at zero is now 0, which indicates superlinear convergence. In fact the algorithm converges to zero in a single step if started anywhere between the two local maxima at $\pm\sqrt{2}$. The second support point for majorization at y is always $-y$. The best quadratic majorizer at $y = \pm\sqrt{2}$ is the horizontal line with function value identically equal to 1.

8.2 A Quartic with a Cubic Fan

Alternatively, again for

$$f(x) = f(y) + f'(y)(x-y) + \frac{1}{2}f''(y)(x-y)^2 + \frac{1}{6}f'''(y)(x-y)^3 + \frac{1}{24}f^{iv}(x-y)^4$$

consider the additive cubic fan

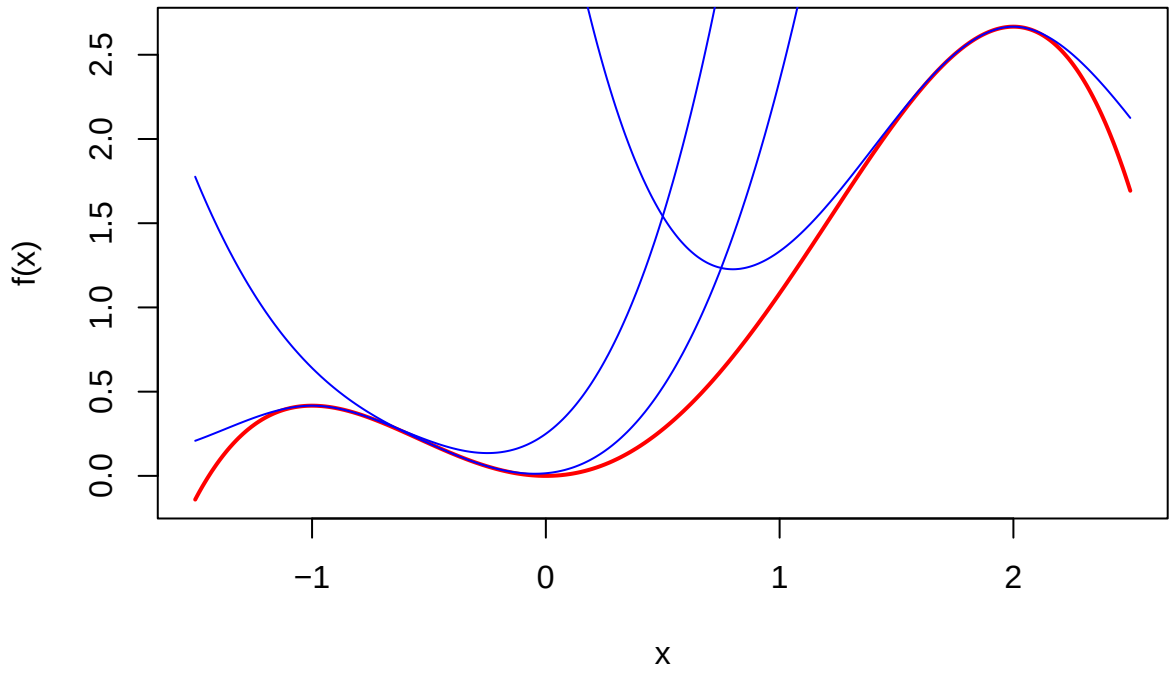
$$g(x,a) = f(y) + f'(y)(x-y) + \frac{1}{2}f''(y)(x-y)^2 + \frac{1}{6}a|x-y|^3.$$

Note that $g(\bullet, a)$ is two times continuously differentiable. Since $g''(x,a) = f''(y) + a|x-y|$ it follows that if $f''(y) \geq 0$ and $a \geq 0$ then $g(\bullet, a)$ is convex.

For this example

$$\underline{a} = \sup_x \frac{\frac{1}{6}f'''(y)(x-y)^3 + \frac{1}{24}f^{iv}(x-y)^4}{\frac{1}{6}|x-y|^3} = \sup_x \mathbf{sign}(x-y)\left(f'''(y) + \frac{1}{4}f^{iv}(x-y)\right).$$

If $f^{iv} > 0$ the fan does not majorize f at y because $\underline{a} = +\infty$. If $f^{iv} < 0$ then $\underline{a} = |f'''(y)|$, and the maximum is attained at $\hat{x} = y$. Note there is only a single support point in this case, or, if you like, the second support point coincides with the first.



The majorization algorithm minimizes

$$g(x,\underline{a}) = f(y) + f'(y)(x-y) + \frac{1}{2}f''(y)(x-y)^2 + \frac{1}{6}|f'''(y)||x-y|^3.$$

Minimization problems of this form are analyzed in an Appendix. If y is close to a local minimum we will have $f''(y) \geq 0$, and consequently $\hat{x} - y = -b + \sqrt{b^2 - 2c}$ if $c < 0$ and $\hat{x} - y = b - \sqrt{b^2 + 2c}$ if $c > 0$.

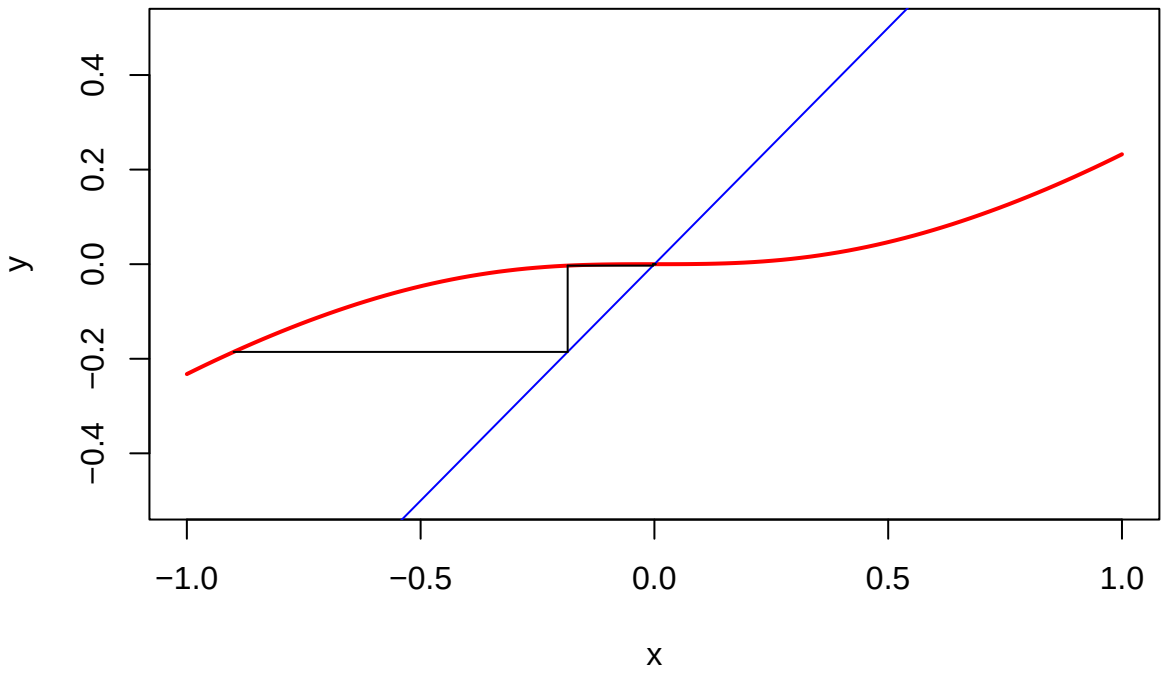
$$b(y) = \frac{f''(y)}{|f'''(y)|}$$

$$c(y) = \frac{f'(y)}{|f'''(y)|}$$

```
y<--.9
for (i in 1:10) {
  if (y == 0) next
  z <- succ(y)
  print(c(y,z))
  y <- z
}
```

```
## [1] -0.9000000000 -0.1855416182
## [1] -0.18554161820 -0.00303932397
## [1] -3.039323970e-03 -1.403766703e-08
## [1] -1.403766703e-08  0.000000000e+00
```

```
cobwebPlotter(-.9,succ,-1,1,-.5,.5,itmax=10)
```



8.3 The Logit

Consider f with $f(x) = -\log \pi(x)$

$$\pi(x) \stackrel{\Delta}{=} \frac{1}{1+\exp(-x)}$$

Because π strictly increases from zero to one, f strictly decreases from $+\infty$ to zero. From $\pi'(x) = \pi(x)(1-\pi(x))$ we see that

$$f'(x) = \pi(x) - 1.$$

Thus, as x goes from $-\infty$ to $+\infty$, f' strictly increases from -1 to 0. It follows that f is strictly convex. Also

$$f''(x) = \pi'(x) = \pi(x)(1-\pi(x)).$$

This implies

$$\lim_{x\to+\infty} f''(x) = \lim_{x\to-\infty} f''(x) = \inf_x f''(x) = 0,$$

as well as

$$\max_x f''(x) = f''(0) = \frac{1}{4}.$$

It also follows that the r-th order derivative is a polynomial of degree r in $\pi(x)$. So if $f^r(x) = \mathcal{P}_r(\pi(x))$ then, because π is strictly monotone,

$$\inf_x \mathcal{P}_r(\pi(x)) = \min_{0\le z\le 1} \mathcal{P}_r(z) \le f^r(x) \le \max_{0\le z\le 1} \mathcal{P}_r(z) = \sup_x \mathcal{P}_r(\pi(x))$$

8.4 The Probit

Consider f with $f(x) = -\log(\Phi(x))$, where

$$\Phi(x) \stackrel{\Delta}{=} \int_{-\infty}^{x} \phi(z)dz,$$

and

$$\phi(x) \stackrel{\Delta}{=} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}.$$

Thus

$$f'(x) = -\frac{\phi(x)}{\Phi(x)} = \mathbf{E}\{z \mid z < x\},$$

where z is a standard normal random variable. The connection between the derivative and the conditional mean is due to Sampford (1953). Thus f' is strictly increasing from $-\infty$ to

$$\lim_{x \to +\infty} f'(x) = \sup_x f'(x) = 0,$$

and f is strictly convex. Also

$$f''(x) = x\frac{\phi(x)}{\Phi(x)} + \left\{\frac{\phi(x)}{\Phi(x)}\right\}^2 = 1 - \mathbf{V}\{z \mid z < x\},$$

which connect the second derivative to the conditional variance of a standard normal. This implies that $f''(x) > 0$, and f'' strictly decreases with

$$\lim_{x \to -\infty} f''(x) = \sup_x f''(x) = 1,$$

and

$$\lim_{x \to +\infty} f''(x) = \inf_x f''(x) = 0,$$

which implies f' is strictly concave.

9 Multivariate Fans

Now, of course, if we allow more general polynomials of the form

$$g(x) = f(y) + f'(y)(x - y) + \sum_{s=2}^{r+1} \frac{1}{s!} a_s (x - y)^s$$

we can presumably do better, in the sense that we can find smaller majorizations. But the vectors $(a_s, \cdots, a_{r+1})$ make the problem multivariate, and do not allow us to define improvement chains in any natural or obvious way.

9.0.1 Scaled Quadratic

$$g(x) = f(x) + (x-y)'\mathcal{D}f(y) + \frac{1}{2}a(x-y)'\Sigma(x-y)$$

$$\underline{a} = \sup_x \frac{f(x) - f(y) - (x-y)'\mathcal{D}f(y)}{\frac{1}{2}(x-y)'\Sigma(x-y)}$$

$$f(x) - f(y) - (x-y)'\mathcal{D}f(y) \leq \sup_{0\leq\lambda\leq 1} \frac{1}{2}(x-y)'\mathcal{D}^2 f(\lambda x + (1-\lambda)y)(x-y),$$

and thus

$$\underline{a} \leq \sup_{0\leq\lambda\leq 1} \sup_x \frac{(x-y)'\mathcal{D}^2 f(\lambda x + (1-\lambda)y)(x-y)}{(x-y)'\Sigma(x-y)} \leq \|\Sigma^{-1}$$

9.0.2 Shifted Newton

$$g(x) = f(y) + (x-y)'\mathcal{D}f(y) + \frac{1}{2}(x-y)'(\mathcal{D}^2 f(y) + a\Sigma)(x-y)$$

$$\sup_x \frac{f(x) - f(y) - (x-y)'\mathcal{D}f(y)}{\frac{1}{2}(x-y)'\Sigma(x-y)} - \frac{(x-y)'\mathcal{D}^2 f(y)(x-y)}{(x-y)'\Sigma(x-y)}$$

9.0.3 Nesterov-Polyak

$$g(x) = f(x) + (x-y)'\mathcal{D}f(y) + \frac{1}{2}(x-y)'\mathcal{D}^2 f(y)(x-y) + \frac{1}{6}a\|x-y\|^3$$

10 Appendix

Consider the problem of finding all critical points (i.e. all points where the derivative vanishes) of the function $f : \mathbb{R} \Rightarrow \mathbb{R}$ given by

$$f(x) = cx + \frac{1}{2}bx^2 + \frac{1}{6}|x|^3$$

with $a \neq 0$. Note that f is coercive, and consequently has a global minimum. Also note that if $b \geq 0$ then f is strictly convex, and the global minimum is the unique critical point.

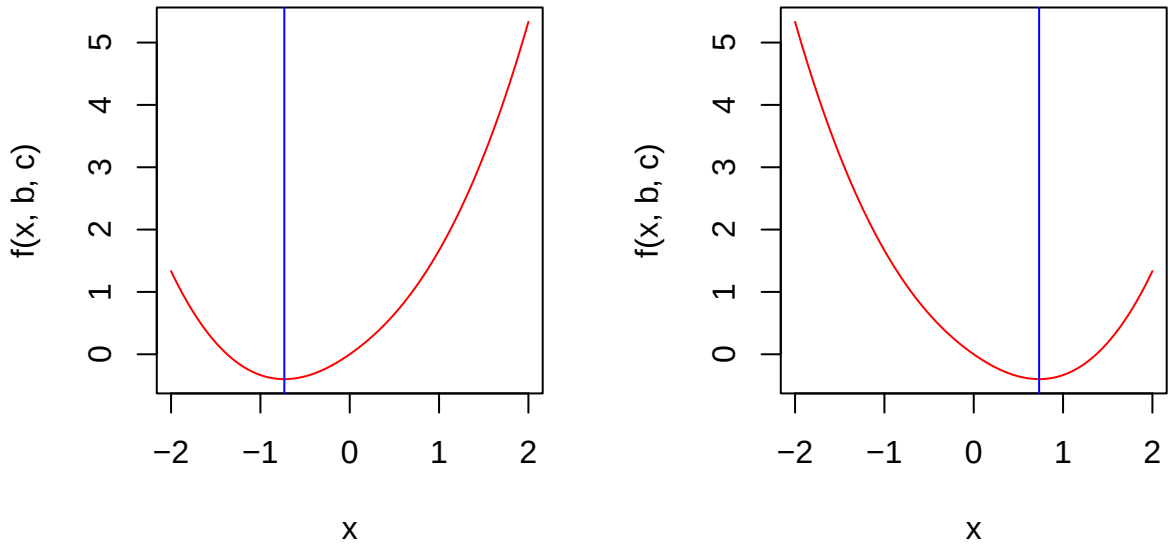
The first and second derivative are

$$f'(x) = c + bx + \frac{1}{2}\mathbf{sign}(x)x^2,$$

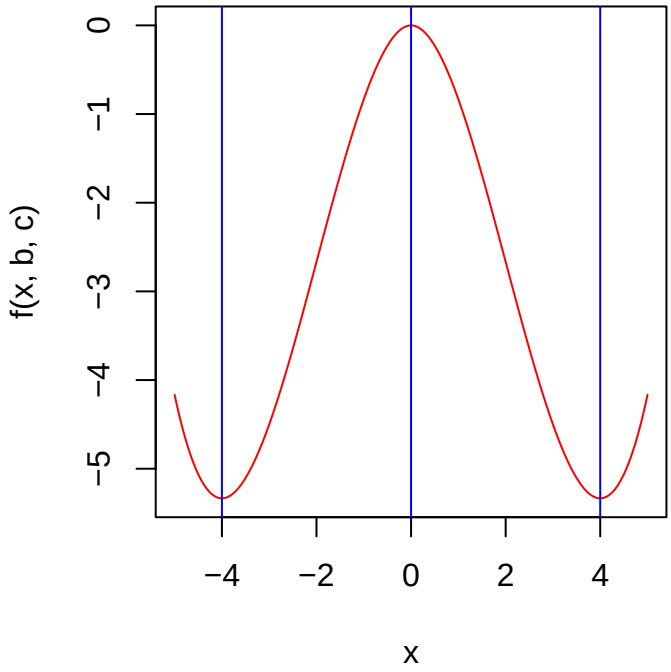
and

$$f''(x) = b + |x|.$$

If b is positive f is strictly convex. Thus f' is increasing. Since $f'(0) = c$ the unique solution of $f'(x) = 0$, corresponding with the global minimum, is positive if $c < 0$ and negative if $c > 0$. If $c < 0$ we have $\hat{x} = -b + \sqrt{b^2 - 2c}$ and if $c > 0$ we have $\hat{x} = b - \sqrt{b^2 + 2c}$. If $b > 0$ and $c = 0$ then $\hat{x} = 0$. If $b = 0$ then $\hat{x} = -\sqrt{2|c|}$ if $c > 0$ and $\hat{x} = \sqrt{2|c|}$ if $c < 0$. If $b = c = 0$ then $\hat{x} = 0$.



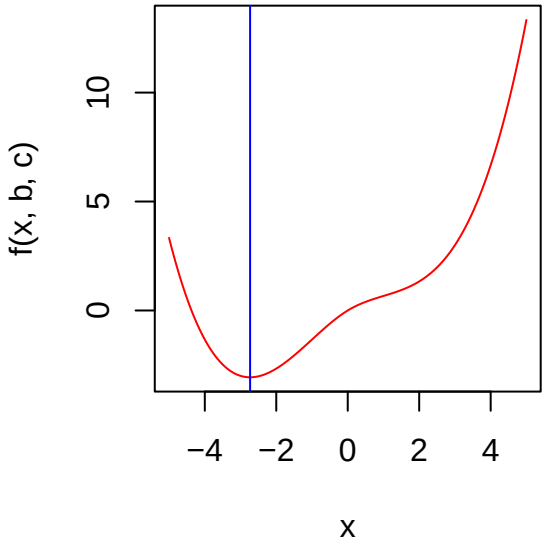
If $b < 0$ the situation is more complicated. If $b < 0$ and $c = 0$ then $f'(x) = x(b + \frac{1}{2}|x|)$ and thus $f'(x) = 0$ for $x = 0$, $x = 2b$, and $x = -2b$. f has a maximum equal to zero for $x = 0$ and two minima equal to $\frac{2}{3}b^3$ for $x = \pm 2b$.

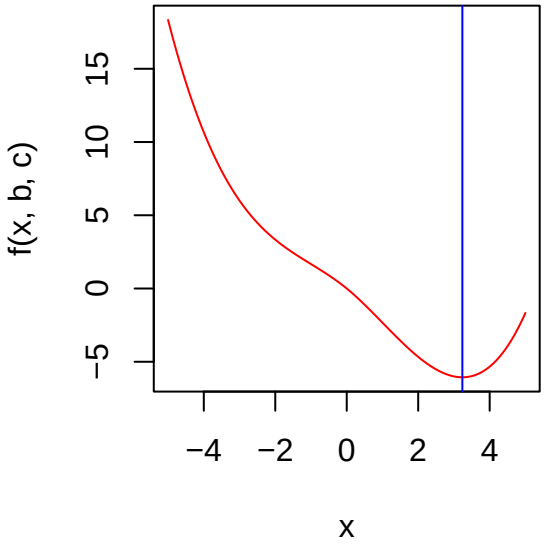


Now consider the case that $b < 0$ and $c \neq 0$. We see that $f''(x) > 0$ for $x < b$ and $x > -b$, while $f''(x) < 0$ for $b < x < -b$. Thus $f'(x)$ first increases from $-\infty$ to its maximum

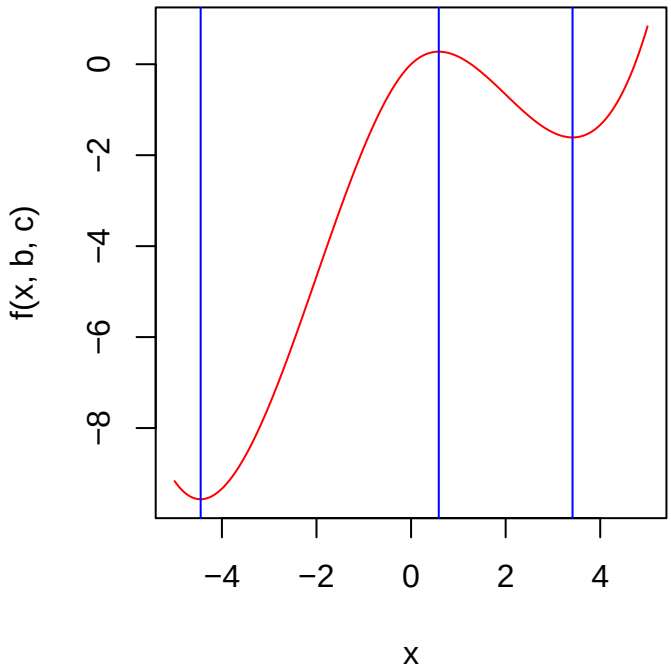
$f'(b) = c + \frac{1}{2}b^2$, then decreases to its minimum $f'(-b) = c - \frac{1}{2}b^2$, and then increases again to $+\infty$.

- If $c + \frac{1}{2}b^2 > c - \frac{1}{2}b^2 > 0$ then $f'(x) = 0$ has a unique solution $\hat{x} < b < 0$. It is the global minimum at $\hat{x} = b - \sqrt{b^2 + 2c}$.
- If $0 > c + \frac{1}{2}b^2 > c - \frac{1}{2}b^2$ then $f'(x) = 0$ has a unique solution $\hat{x} > -b > 0$. It is the global minimum $\hat{x} = -b + \sqrt{b^2 - 2c}$.





- If $c + \frac{1}{2}b^2 > 0 > c - \frac{1}{2}b^2$ then $f'(x) = 0$ has three solutions. There are two local minima at $\hat{x} = b - \sqrt{b^2 + 2c} < b < 0$, and $\hat{x} = -b + \sqrt{b^2 - 2c} > -b > 0$, and one local maximum at $b < \hat{x} = -b - \sqrt{b^2 - 2c} < -b$.



References

Böhning, D., and B. G. Lindsay. 1988. "Monotonicity of Quadratic-approximation Algorithms." *Annals of the Institute of Statistical Mathematics* 40 (4): 641–63.

D'Esopo, D. A. 1959. "A Convex Programming Procedure." *Naval Research Logistic Quarterly* 6: 33–42.

De Leeuw, J. 1977. "Applications of Convex Analysis to Multidimensional Scaling." In *Recent Developments in Statistics*, edited by J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, 133–45. Amsterdam, The Netherlands: North Holland Publishing Company.

———. 1994. "Block Relaxation Algorithms in Statistics." In *Information Systems and Data Analysis*, edited by H. H. Bock, W. Lenski, and M. M. Richter, 308–24. Berlin: Springer Verlag.

De Leeuw, J., and K. Lange. 2009. "Sharp Quadratic Majorization in One Dimension." *Computational Statistics and Data Analysis* 53: 2471–84.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood for Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society* B39: 1–38.

Heiser, W. J. 1995. "Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis." In *Recent Advantages in Descriptive Multivariate Analysis*, edited by W. J. Krzanowski, 157–89. Oxford: Clarendon Press.

Hunter, D. R., and K. Lange. 2004. "A Tutorial on MM Algorithms." *American Statistician* 58 (30–37).

Lange, K. 2016 (in press). *MM Optimization Algorithms*.

Sampford, M. R. 1953. "Some Inequalities on Mill's Ratio and Related Functions." *Annals of Mathematical Statistics* 24: 130–32.

Spivak, M. 1965. *Calculus on Manifolds*. Westview Press.

Van Ruitenburg, J. 2005. "Algorithms for Parameter Estimation in the Rasch Model." Measurement and Research Department Reports 2005-04. Arnhem, Netherlands: CITO.

Vosz, H., and U. Eckhardt. 1980. "Linear Convergence of Generalized Weiszfeld's Method." *Computing* 25: 243–51.

Zangwill, W. I. 1969. *Nonlinear Programming: a Unified Approach*. Englewood-Cliffs, N.J.: Prentice-Hall.