Fractional Programming: a survey, recent developments, and applications – to be presented in Summer School of GCM9

R. L. Sheu

Institute of Applied Mathematics
National Cheng Kung University, Tainan,
Taiwan

July 15 and July 17, 2008

Outline

- Introduction to fractional programming and applications
- Generalized concavity and fractional duality
- Methods and algorithms

Problem Definition

- Optimization involving ratios of functions. That is,

$$(P) \quad \sup_{x \in X} \frac{f(x)}{g(x)}$$

where $X = \{x \in C | h_j(x) \leq 0,\ j = 1, 2, \ldots, n\}$ is called a *fractional program*.

- When f, g, h_j are affine, (P) is a linear fractional program.
- When f, g are quadratic and h_j affine, (P) is called a quadratic fractional program.
- When $f \geq 0$ is concave and $g > 0$ and h_j are convex, (P) is called a concave fractional program.

• In some applications, more than one ratio can be considered.
Particularly,
The *generalized fractional program*:

$$(P) \quad \lambda^* = \max_{x \in X} \min_{1 \le i \le n} \{ \frac{f_i(x)}{g_i(x)} \}$$

The *sum-of-ratios program*:

$$(P) \quad \lambda^* = \max_{x \in X} \sum_{1 \le i \le n} \{ \frac{f_i(x)}{g_i(x)} \}$$

The *multi-objective fractional program*:

$$(P) \quad \lambda^* = \min_{x \in X} \{ (\frac{f_1(x)}{g_1(x)}, \ldots, \frac{f_n(x)}{g_n(x)}) \}$$

All functions $g_i > 0$.

Traditional LP models try to decide on the best combination of different activities over a fixed time horizon by maximizing the profit subject to resources constraints:

$$(LP) \quad \max \; Z = \sum_{j=1}^{n} (p_j - v_j) x_j$$

$$s.t. \; \sum_{j=1}^{m} a_{ij} x_j \leq b_i, \; i = 1, 2, \ldots, m.$$

where p_j is the *given* market prices per unit (under conditions of perfect competition); v_j the variable costs per unit of activity x_j; b_i are the capacities, usually representing the floor spaces of a warehouse; the availability of transportation equipments; sizes of machines; or management staffs, etc.

Problems in the profit maximizing LP model

- Many, perhaps most, real world systems are dynamic, so naturally the time factor should be an important decision variable. The LP model does not take into consideration various decisions over different time horizons.
- In reality, there are no such perfect market prices. Usually the entrepreneur has to decide not only the activity levels x_j, but also the prices p'_j of their products. The profit maximizing model above is of little help to the price fixer.

- The economic fixed costs such as the lease payments of warehouses, the cost of the machines, annual salaries of management staffs are often lurking behind the set of capacity constraints. Let FC be the totality of the fixed cost. Then, the LP model above either ignores the fixed cost, or allocates the fixed cost equally to each activity:

$$\max \sum_{j=1}^{n}(p_j - v_j)x_j = \max \sum_{j=1}^{n}(p_j - v_j)x_j - FC$$

$$= \sum_{j=1}^{n}(p_j - v_j - \frac{FC}{\sum_{j=1}^{n} x_j})x_j$$

In either case, the LP model above provides almost no guideline to the accountant for valuing the finished goods inventory of the firm.

Full-cost Objective

$$Z = \sum_{j=1}^{n} k_j \left(v_j + \frac{v_j x_j}{\sum v_j x_j} \cdot \frac{FC}{x_j} \right) x_j - \sum v_j x_j - FC$$

proposed by Colantoni, Manes, Whinston, *The Accounting Review*, 1969.

• The total fixed costs are allocated based on the proportion which total variable costs of that activity bear to total variable costs of all activities.

• The perfect market price is replaced by k_j, the mark-up of activity j, multiplying the full cost of that activity. (For government projects, k_j often has an upper bound like 110% or 120% subject to audit.)

• Each entrepreneur has his own considerations for the $k_j's$. If k_j's are given parameters in the model, it is a sum-of-linear-ratios problem. If k_j's are also decision variables, a quadratic fractional programming occurs.

Container ship stowage planning (Picture: Port of KaoHsiung)

• A linear programming model may be set up to find the best combination of cargoes to be loaded in a ship, in terms of maximum profit.
• However, depending on weights, lengths, or other characters of goods, some kinds of cargo containers are loaded, or unloaded, more slowly than others.

- For example, to maintain the stability and safety of the ship, heavier containers should be placed lower than the lighter ones, causing shifts of containers in doing so.


7
91
SC-48


YANG MING
YANG MING


CAI
YANG MING
YANG MING
YANG MING
YANG MING
SWL 40 TONNES
CABIN SIDE
SPREADERS


YANG MING LINE
YANG MING
YANG MING LINE
YANG MING
YANG MING
IMPSA
SWL
40t
60t


華航 3
M9612

Cargo Summary

Bay	Weight	LCG	VCG	TCG	Qty
01	95.8	94.87	19.52	2.16	16
02	0.0	91.80	0.00	0.00	0
03	299.9	88.73	19.65	0.86	17
05	59.1	80.64	15.04	-6.03	4
06	270.7	77.57	11.88	0.41	21
07	73.6	74.50	16.94	-7.64	4
09	612.4	66.99	11.19	0.63	33
10	198.6	63.92	17.21	0.20	8
11	660.1	60.85	11.58	0.15	34
13	504.0	52.94	9.35	-4.31	28
14	289.3	49.87	14.57	7.54	13
15	507.1	46.80	9.09	-3.69	28
17	814.9	39.29	10.18	-1.51	47
18	240.8	36.22	14.23	-0.87	15
19	835.1	33.15	9.96	-1.58	47
21	836.2	24.88	11.03	2.36	43
22	99.0	21.81	14.37	12.43	5
23	829.7	18.74	10.82	2.31	43
20'	13062.7				703
40'	5132.5				233
Ttl	18195.2	Hold	Deck		936

Figure: Loading information: Weight, LCG, VCG, TCG

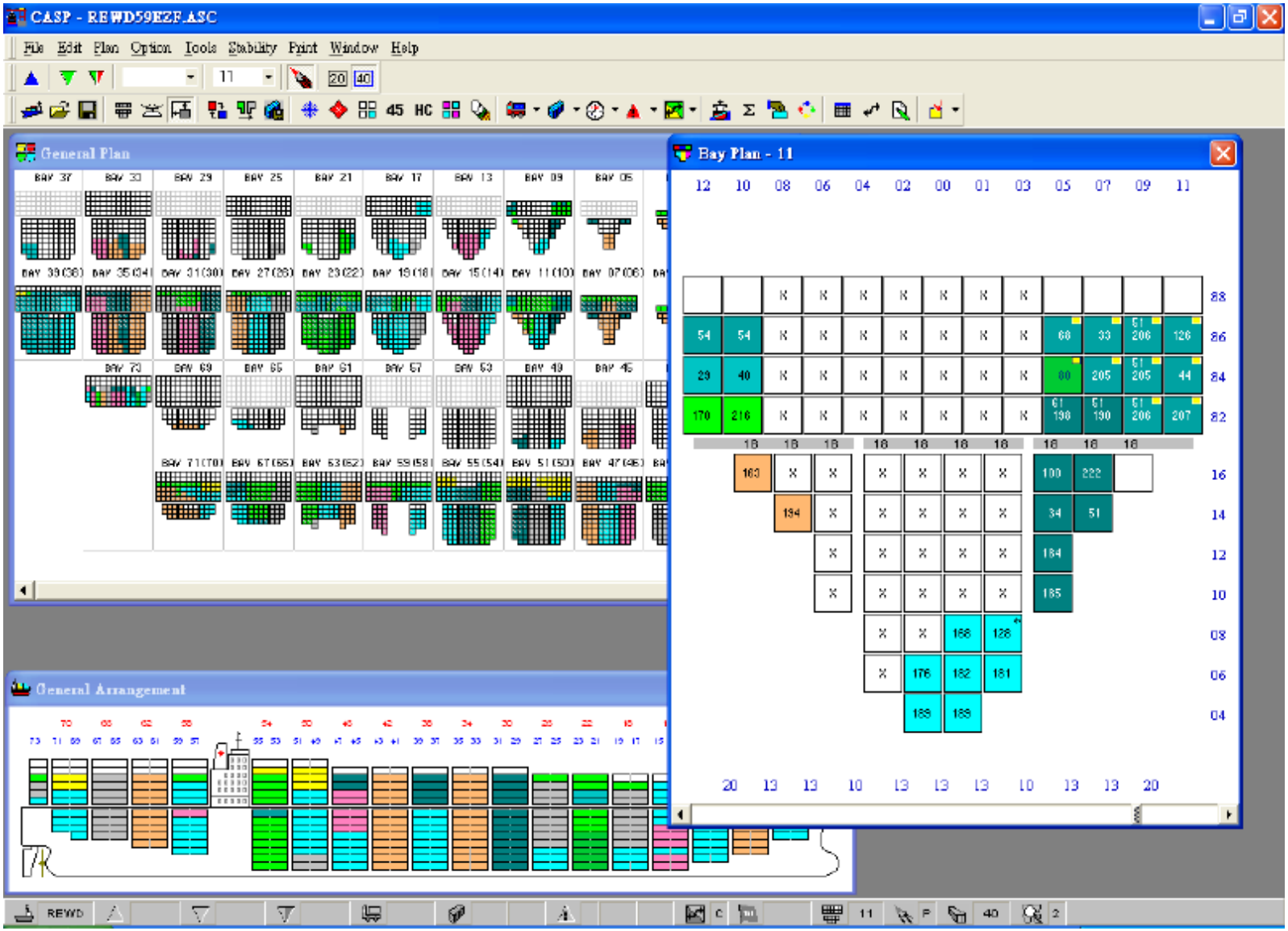


Figure: Bay planning on a vessel

Figure: One of the most important cost for a containership company is the yearly rental fee for a berth.

- A stowage plan is necessary to improve the efficiency in loading/unloading the vessel subject to safety regulations. The purpose is to reduce the number of shifting at ports and on ships.
- To this end, traditional profit-maximization LP model may not be useful as it does not consider the loading time of different cargoes, and thus fails to pass a proper share of the yearly rental fee of the berth to different types of containers.

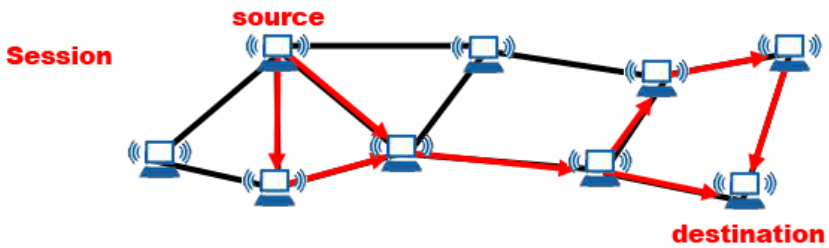
• An objective function taking into account of the loading time is suggested as follows: (Fractional Programming, B.D. Craven, 1988)

$$\max \frac{\sum_i (p_i - C_1 k_i) x_i - C_2 T}{\sum_i k_i x_i + T}$$

where p_i is the unit profit for cargo of type i; C_1 is the cost per unit time at ports; k_i the loading time for cargo i; C_2 is the cost per unit time at sea, and T the journey time.
• Given k_i, this is a single ratio linear fractional programming. If k_i's are also decision variables representing different stowage plans, it becomes a quadratic fractional programming problem.
• The objective function can be viewed as to allocate the fixed costs at ports (berth/equipments rental fees) by prorating to various types of cargoes based on loading/unloading times.

Congestion Control

- In a wireless telecommunications network, two wireless devices (e.g. cell phone base stations, computer accessing points, etc.) are connected through a set of links (I), called *sessions*.
- The capacity of each link (data rate, bits per second, for example, 10Mbit/s) is adjustable, depending upon the transmitted power of the device and the interferences from other nodes. The stronger the power and the less the interferences, the better the throughput.



The average congestion level on a particular link l over a period of T time slots is defined by

$$CL(l, P_{l,t}, f_{l,t}^s) = \frac{\text{Total loaded flows}}{\text{Total Capacity}} = \frac{\sum_{s\in S, t\in\{1,2,\ldots,T\}} f_{l,t}^s}{\sum_{t\in\{1,2,\ldots,T\}} C_{l,t}}$$

The congestion control problem in a wireless telecommunication network is to determine the best power level $P_{l,t}$ for each link l at different time slot t and to decide the data routings $(f_{l,t}^s)$ subject to the capacity defined by $P_{l,t}$, so that all required transmission rates for each session are met in such a way that the highest congestion level in this network is minimized.

$$\min_{P_{l,t}^m, f_{l,t}^s} \max_{l\in L} CL(l, P_{l,t}, f_{l,t}^s)$$

$$\min_{P_{l,t}^m, f_{l,t}^s} \max_{l \in L} \frac{\sum_{s \in S, t \in \{1,2,\ldots,T\}} f_{l,t}^s}{\sum_{t \in \{1,2,\ldots,T\}} C_{l,t}}$$

$$s.t.\ C_{l,t} = \sum_{m \in M} W^m \log_2(1 + \frac{h_l^m P_{l,t}^m}{W^m \sigma + \sum_{l' \in L, l' \neq l} h_{l'}^m P_{l',t}^m}), \forall l, t;$$

$$P_{l,t}^m \leq P^{max}, \forall l \in L, t \in \{1, 2, \ldots, T\}$$

$$\sum_{s \in S} f_{l,t}^s \leq C_{l,t}, \forall l \in L, t \in \{1, 2, \ldots, T\}$$

$$\sum_{l \in IL(i) \cup OL(i), m \in M} x(P_{l,t}^m) = 1, \forall i \in N, t \in \{1, 2, \ldots, T\}$$

$$\sum_{l \in IL(i), t \in \{1,2,\ldots,T\}} f_{l,t}^s - \sum_{l \in OL(i), t \in \{1,2,\ldots,T\}} f_{l,t}^s = r_s T \mathbb{I}_{IS(i)}(s), \forall i \in N, s \in S$$

$$P_{l,t} \geq 0, f_{l,t}^s \geq 0, \forall m \in M, s \in S, l \in L, t \in \{1, 2, \ldots, T\}$$

Decision variables (in red):

$P_{l,t}^m$: Transmission power on channel m, link l and time t.

$f_{l,t}^s$: Flow rate of session s on link l at time t.

$x(P_{l,t})$: $x(P_{l,t}) = 1$ if $P_{l,t} > 0$; 0 otherwise.

Parameters:

S : Set of sessions(demands).

N : Set of nodes(devices).

M : Set of available channels.

W^m : Bandwidth of channel m, $m \in M$.

h_l^m : Channel response of channel m on link l,$m \in M, l \in L$.

r_s: Required transmission rate of session s.

$IL(i) = \{l = (k, i) \in L\}$

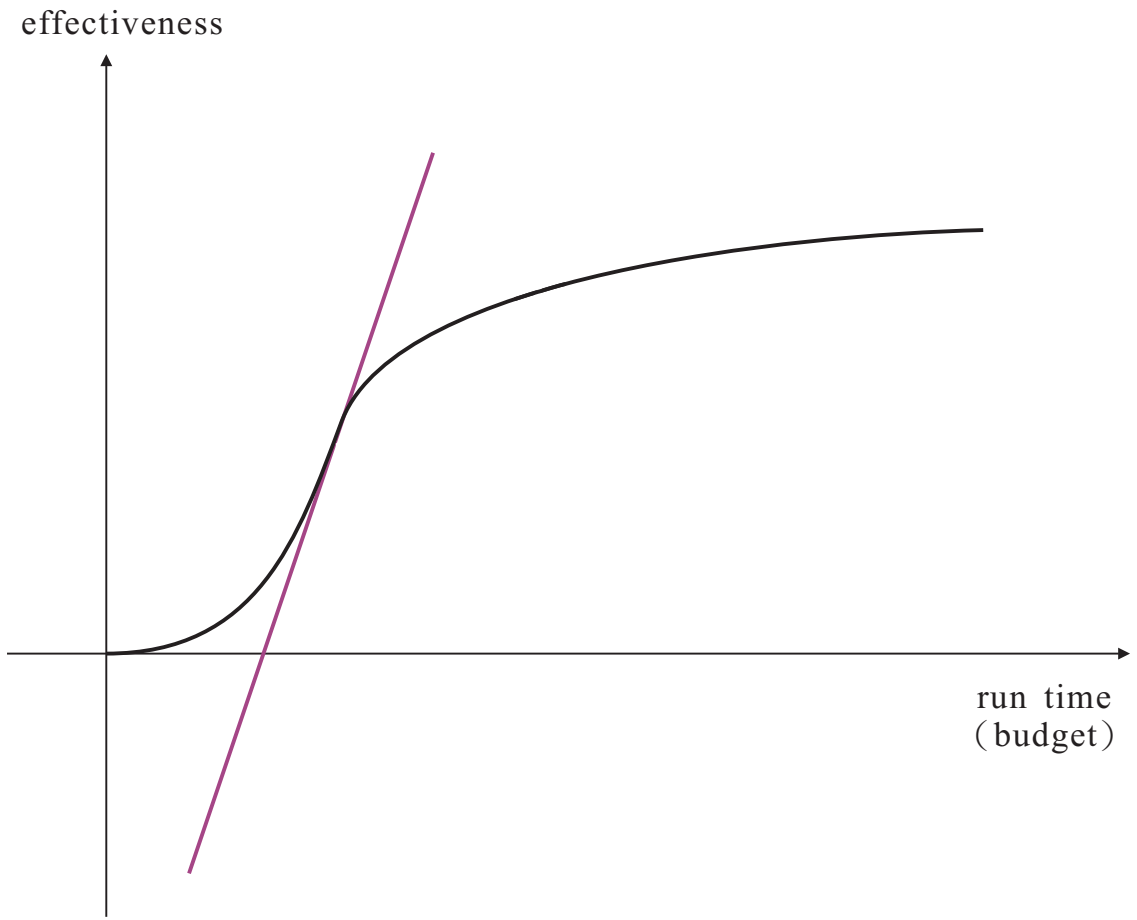
$OL(i) = \{l = (i, j) \in L\}$

$IS(i) = \{s = (n_s, n_d, r_s) \in S : n_d = i\}$

$\mathbb{I}_A(s)$: $\mathbb{I}_A(s) = 1$, if $s \in A$, otherwise $\mathbb{I}_A(s) = 0$.

Advertising strategy

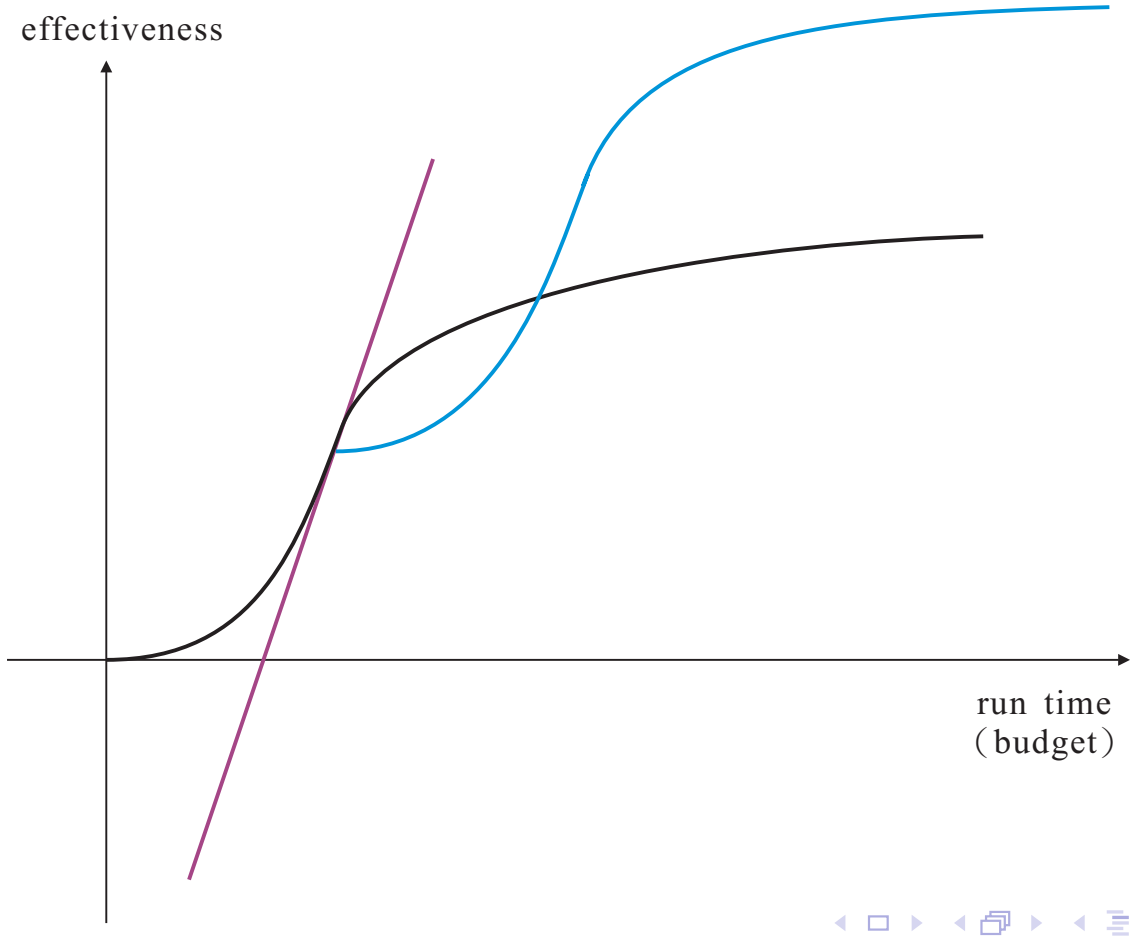
- Companies advertise to promote a product or an idea. For example, in March 2000, Coca-Cola unveiled a major shift for its Diet Coke brand. It aimed to exude energy and create a more upbeat image. The target audiences were among the age group ranging from 25 to 49, and the main theme of the message is "be-true-to-yourself".
- When an advertising campaign is launched, the message is first seen and read. The advertising effectiveness grows slowly until it is believed and memorized at which time the effectiveness bursts out. Finally, the marginal effect of the advertisement fades even with a heavy broadcasting frequency. This can be described by a logistic curve.


effectiveness
run time
(budget)

- It is then important to evaluate the advertising effectiveness subject to a given budget. Does it communicate well with the target audiences? Does it stimulate or change the consumers' buying behavior?
- Maximizing the advertising effectiveness subject to the broadcast runtime will lead to consume all the broadcasting budget. This does not make too much sense as the final stage of the campaign is, though, still "effective", yet "inefficient" with a very low "marginal effect (effectiveness per unit budget)".

- What Coca-Cola did was to campaign "be-true-to-yourself" by a long-run series. The next advertising (sequel) follows immediately when the marginal effect of the previous one starts falling.
- This amounts to determining the largest marginal effect of each advertisement, namely, to maximize a slope function – a ratio of two functions.

The sequel follows immediately when the marginal effect of the previous one starts falling.



Other Applications

- An equilibrium model for an expanding economy introduced by J. von Neumann. (A model of general economic equilibrium, *Review of Economic Studies* 13, 1945)
- The growth rate is determined by

$$\text{growth rate} = \max_{x}\Big(\min_{1\le i\le p}\frac{output_i(x)}{input_i(x)}\Big),$$

where x denotes a feasible production plan of the economy and the efficiency measures are expressed as output-input ratios that are to be optimised under a max-min criterion.

Optimizing efficiency

- In some resource allocation problems, the efficiency measures can be the ratio of "profit to revenue".
- In portfolio selection model, the efficiency measures could be "return to risk"; "earning per share"; "dividend per share"; or "liquidity".
- A routing problem in a transportation network used to minimize the "cost-to-time" ratio.
- In a wireless network, the interference among different local transmissions is one of the most important issues. The signal-to-noise ratios are thus to be max-minimized.

$$SNR_l = \frac{P_l G_{ll}}{n + \sum_{j \neq l} P_j G_{jl}}$$

Optimizing quadratic ratios

- In numerical analysis, the eigenvalues of a matrix can be calculated as constrained maxima of the Rayleigh quotient, a ratio of two quadratic forms.
- In study for predictability in asset returns, the measure for predictability is to be optimized:

$$\lambda(m) = \frac{m'\Gamma_0(\hat{z})m}{m'\Gamma_0(z)m},$$

where $\Gamma_0(z) = \Gamma_0(\hat{z}) + \Sigma$, $\Gamma_0(\hat{z})$ is a covariance matrix. It is assumed that Σ, and therefore $\Gamma_0(z)$ is positive definite. The vector m is a particular linear combination of the primary assets.

Survey papers and books

- S. Schaibled published a bibliography collecting 1198 articles of fractional programming. (in *R. Horst and P. M. Pardalos (Eds.), Handbook of Global Optimization*, 1995). A more recent version by Frenk and Schaible appeared in Encyclopedia of Optimization, C. A. Floudas and P. M. Pardalos, Eds., 2001.
- I. M. Stancu-Minasian has a series of 6 bibliographies. The first three appeared in *Pure and Applied Mathematika Sciences*, 13 (1981), 17 (1983) and 22 (1985). The forth to sixth ones in *Optimization* 23 (1992), 45 (1999), 55 (2006))
- B. D. Craven published a book (1988) discussing important concepts of multi-ratios problems. *Fractional programming, Sigma Series in Applied Mathematics.*


合歡山主峯
3416
公尺
長途電信

Generalized concavity for ratios

- The ratio of two functions is, in general, not concave.
- For a concave fractional function(a concave function divided by a convex function), it is only *quasi-concave*.
- Generalized concavity properties of special family of ratios can be summarized into the following three tables. (in "Generalized concavity," by Avriel, Diewert, Schaible, and Zhang, Plenum Press, 1988.)

The Ratio Φ_1/Φ_2		
	Φ_2	
Φ_1	> 0, concave	> 0, convex
≥ 0 concave	–	s. qcv
≥ 0 convex	s. qcx	–

- s. qcv: semistrictly quasiconcave
- s. qcx: semistrictly quasiconvex

The Ratio Φ/l		
	l	
Φ	> 0	< 0
concave	s. qcv	s. qcx
convex	s. qcx	s. qcv

- l is affine
- s. qcv: semistrictly quasiconcave
- s. qcx: semistrictly quasiconvex

The Ratio l/Φ

	Φ			
	concave		convex	
l	> 0	< 0	> 0	< 0
≥ 0	s. qcx	s. qcx	s. qcv	s. qcv
≤ 0	s. qcv	s. qcv	s. qcx	s. qcx

- l is affine
- s. qcv: semistrictly quasiconcave
- s. qcx: semistrictly quasiconvex

Important notes for GC functions

- Quasi-concavity is a slight generalization of concavity. Especially, any local maximum of a quasi-concave function is global. In other words, quasi-concave functions are unimodal.
- However, for a quasi-concave function, there might be many critical points that are non-local maximum (like points of inflection). This fails most first-order based algorithms.
- Moreover, the ordinary Lagrangian duality no longer possesses the strong duality (but still has the weak duality) for a fractional program.

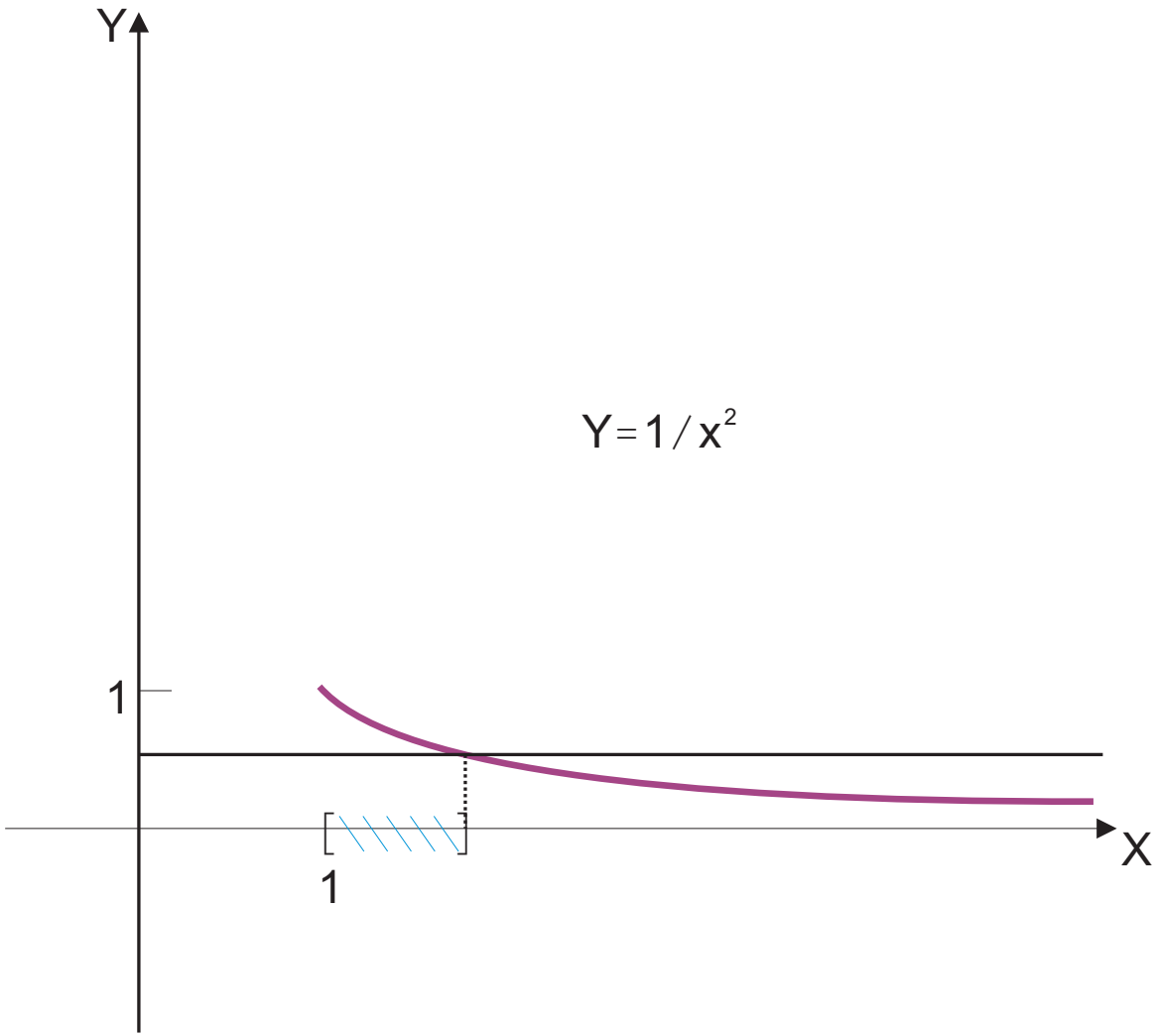
Example

$$\max_{x>0}\{\frac{1}{x^2}|x \geq 1\} = 1,$$

provided by S. Schaible, "Duality in fractional programming: a unified approach", Operations Research 24, 1976.

- This is a concave fractional program.
- The function $\frac{1}{x^2}$ is quasi-concave. It is convex indeed.
- Maximizing the quasi-concave function amounts to maximizing a convex function (generally a tough problem!!)
- The basic Lagrangian duality is not valid since

$$\inf_{u\geq 0}\{\sup_{x>0}[(1/x^2) - u(1-x)]\} = \infty.$$


Y
$Y = 1/x^2$
1
1
X

Duality for generalized concave fractional program

$$(GP) \quad \sup_{x \in X} \{ \min_{1 \le i \le p} \frac{f_i(x)}{g_i(x)} \}$$

where $X = \{x \in C | h_j(x) \le 0,\ j = 1, 2, \ldots, m\}$, C is convex compact.
For notational convenience, let $F(x) = (f_1(x), \ldots, f_p(x))^t$; $G(x) = (g_1(x), \ldots, g_p(x))^t$; $H(x) = (h_1(x), \ldots, h_m(x))^t$.
Problem (GP) has the following equivalent form:

$$\sup\{\tau | F(x) - \tau G(x) \ge 0, H(x) \le 0, x \in C\}.$$

Assume that $F \ge 0, G > 0$. Then $\tau \ge 0$ and the constraint set is convex.

Given a continuous vector-valued convex function k, the *theorem of alternative* asserts that exactly one of the following alternatives is true:

$$I.\ k(x) \leq 0, x \in X \text{ is consistent;}$$

$$II.\ \text{there exists } y \geq 0, s.t.\ y^t k(x) > 0, \forall x \in X.$$

We arrive the following dual program:

$$(GD) \qquad \inf_{u>0, v\geq 0} \{\sup_{x\in C} \frac{u^t F(x) - v^t H(x)}{u^t G(x)}\}.$$

The weak duality between (GP) and (GD) holds, that is,

$$value(GD) \geq value(GP);$$

and the strong duality can be verified too.

Remarks for the duality

- There are other dual forms in literature, and they may not be equivalent.
- Unlike the Lagrange duality, (GD) is a type of "non-linear" dual, which retains the fractional structure.
- The dual problem (GD) is again a generalized fractional programming, involving only linear ratios. However, when C is not finite, there are infinitely many of affine ratios whose maximum is to be minimized.
- There is an Dinkelbach-type algorithm for solving generalized fractional programming of infinitely many ratios. (Lin and Sheu, 2005, "Modified Dinkelbach-type algorithm for generalized fractional programming with infinitely many ratios", JOTA 126.)

Duality for the sum-of-ratios program

- The sum-of-ratios function $\sum_{i=1}^{p} \frac{f_i(x)}{g_i(x)}$ is, in general, not of any type of generalized concavity.
- Schaible (NRLQ, 1977) showed that, a sum of a linear function and a ratio of affine functions need not be quasi-concave.
- Very few theoretic result for the duality of the sum-of-ratios problem has been reported so far.
- Scott and Jefferson (JOTA, 1998) obtained a kind of duality, in a sense similar to what Craven called in 1977 (Bulletin of the Australian Mathematical Society) "the *quasi-duality*".

$$(SP) \quad \min_{x \in X} \sum_{j=1}^{M} (\sum_{j=1}^{N} a_{ij} x_j + g_j) / (\sum_{j=1}^{N} b_{ij} x_j + h_j)$$

$$s.t.\ X = \{x | \sum_{j=1}^{N} c_{kj} x_j \leq 1,\ k = 1, 2, \ldots, K\}$$

All vectors a, b, c, g, h and x are positive.

Scott and Jefferson transform the sum-of-linear-ratios problem into the format of the signomial programming:

$$
\begin{aligned}
(SP1) \quad \min & \sum_{i=1}^{M} s_i t_i^{-1} \\
s.t. & \sum_{j=1}^{N} a_{ij} x_j s_i^{-1} + g_j s_i^{-1} \leq 1, \quad i = 1, \ldots, M, \\
& \sum_{j=1}^{N} b_{ij} x_j t_i^{-1} + h_j t_i^{-1} \geq 1, \quad i = 1, \ldots, M, \\
& \sum_{j=1}^{N} c_{kj} x_j \leq 1, \ k = 1, 2, \ldots, K, \\
& x, s, t > 0.
\end{aligned}
$$

- The signomial programming is also known as an (ordinary) geometric program with reverse constraints.
- The ordinary geometric program (Duffin, Peterson, Zener 1967) can be convexified via an exponential transformation, and a complete duality can be derived.
- As a consequence of the reverse constraints, the signomial program cannot be similarly convexified.
- Duffin and Peterson (JOTA 1973) derived the duality theorem of the geometric program with reversed constraints in the sense of *equilibrium solutions*.

Signomial dual for sum-of-linear-ratios program

$$(SD1) \quad \max_{\delta,\gamma,\beta\geq 0} \prod_{i=1}^{M}(1/\delta_{oi})^{\delta_{oi}} \prod_{i=1}^{M}\prod_{j=1}^{N}(a_{ij}/\delta_{ij})^{\delta_{ij}}(b_{ij}/\gamma_{ij})^{-\gamma_{ij}}$$

$$\times \prod_{i=1}^{M}(g_i/\delta_i)^{\delta_i}(h_i/\gamma_i)^{-\gamma_i} \prod_{k=1}^{K}\prod_{j=1}^{N}(c_{kj}/\beta_{kj})^{\beta_{kj}} \prod_{k=1}^{K}{\alpha_k}^{\alpha_k}$$

$$s.t. \ \sum_{i=1}^{M}\delta_{oi} = 1; \alpha_k = \sum_{j}\beta_{kj}; \delta_{oi} = \sum_{j}\delta_{ij} + \delta_i;$$

$$\delta_{oi} = \sum_{j}\gamma_{ij} + \gamma_i; \sum_{i=1}^{M}(\delta_{ij} - \gamma_{ij}) + \sum_{k=1}^{K}\beta_{kj} = 0.$$

Nonconvex dual

- Taking the logarithm of the dual objective function, we find (SD1) is indeed a DC program.
- The primal equilibrium solution (x, s, t) can be related with the dual equilibrium solution (δ, γ, β) by a system of equations (primal-dual conversion), so there is no duality gap at each pair of primal-dual equilibrium solutions. (Strong duality)
- The dual equilibrium solution may locate at a non-global local maximum of (SD1) whereas the corresponding primal solution could also be a maximizing point (a *max-max* situation) or just a stationary point. In other words, no weak duality theorem for the signomial dual of the sum-of-linear-ratios model.

- Having the strong duality (locally) without the weak duality is typical of a non-convex dual program like the signomial dual.
- Passy and Wilde (SIAM Journal on Applied Mathematics, Vol. 25, 1967) proposed a *quasi-maximization* procedure by finding all maximizing dual stationary points and then choosing the minimum among them.
- This has much in common with the concept of Craven's quasi-duality. (Bulletin of the Australian Math. Society, Vol. 16, 1977) and the recently developed "canonical duality" by David Gao. (*Duality principles in nonconvex systems: theories, methods, and applications*, Kluwer Academic Publishers, 2000.)
- Nowadays, non-convex duality has become an important research area in global optimization.

Craven's quasidual pair

Consider the following pair of problems:

$$(QP) \quad quasimin_x \ f(x), \text{ subject to } k(x) \in K$$

and

$$(QD) \quad quasimax_{u,v} \ g(u,v) = f(u) - vk(u)$$
$$s.t. \quad f'(u) - vk'(u) = 0, \ v \in K^*.$$

where K is a cone and K^* its conjugate cone. Problem (QD) (indeed the Wolfe-type dual without convexity assumption), is called a *quasidual* of (QP) if, (QP) has a quasimin at $x = \xi$, (QD) has a quasimax at $(u, v) = (\mu, \upsilon)$ and $f(\xi) = g(\mu, \upsilon)$.

The Craven concept of quasi-min

- In the simplest situation, the point a is called a *quasimin* of $f(x)$, if
$$f'_+(a) \geq 0; \ \ f'_-(a) \leq 0$$
where $f'_+(a)$ is the right derivative of f at a should it exist. Similarly, $f'_-(a)$ is the left derivative. The definition can be generalized to a much more general space without one-side(directional) derivatives.
- The point a is a *quasimax* of f, if and only if, $-f$ has a quasimin at a.
- The quasimin and the quasimax are defined locally.

Illustrative example for quasi-duality

Consider the concave minimization problem:

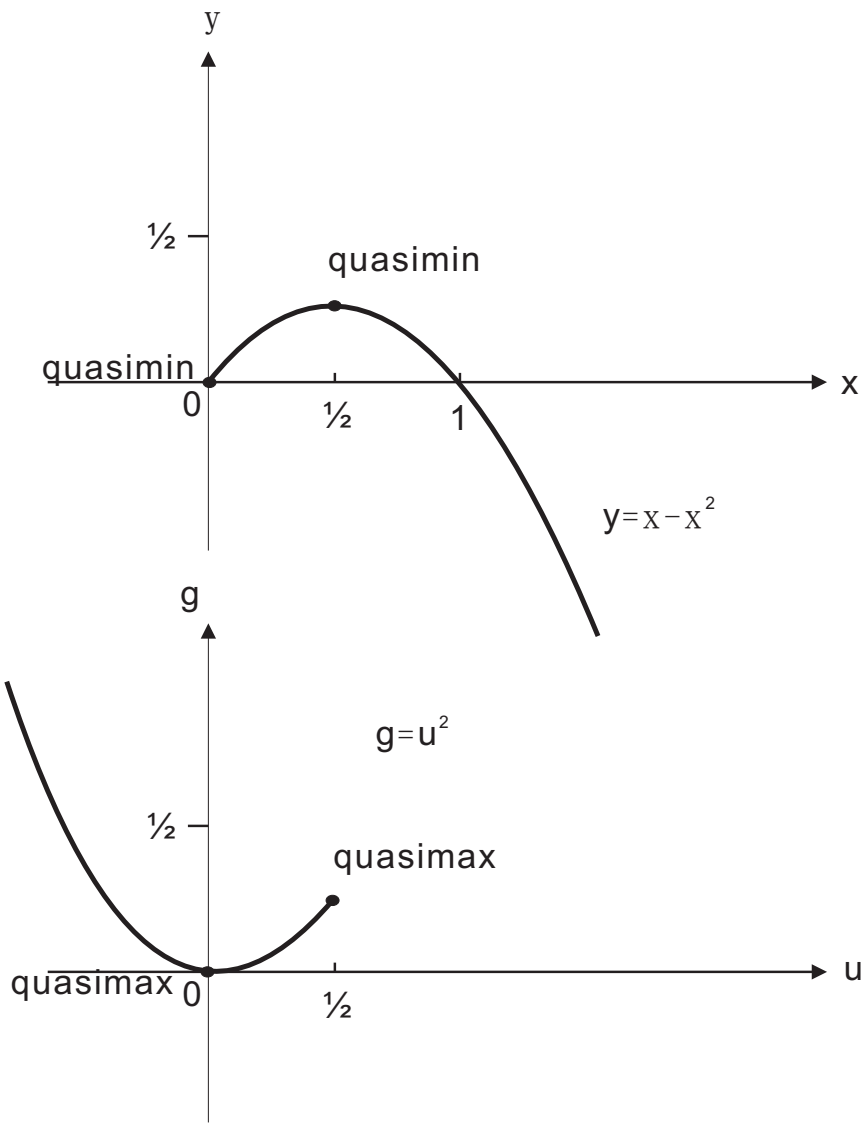
$$(QP) \quad \min_{x \geq 0} \; f(x) = x - x^2.$$

The "quasi-dual" is as follows:

$$(QD) \quad \max_{\lambda \geq 0} \; u - u^2 - \lambda u, \quad \text{subject to } 1 - 2u - \lambda = 0$$

which is equivalent to

$$\max \; g(u) = u^2, \quad \text{subject to } u \leq \frac{1}{2}.$$


y
½
quasimin
quasimin
0
½
1
x
y=x−x²
g
g=u²
½
quasimax
quasimax
0
½
u

- Therefore, (QP) and (QD) have two pairs of quasi-critical points that do not have a duality gap.
- But there is no weak duality:

$$\text{On } x \geq 0,\ u \leq 1/2, \quad f(x) \not\geq g(u).$$

David Gao's canonical duality

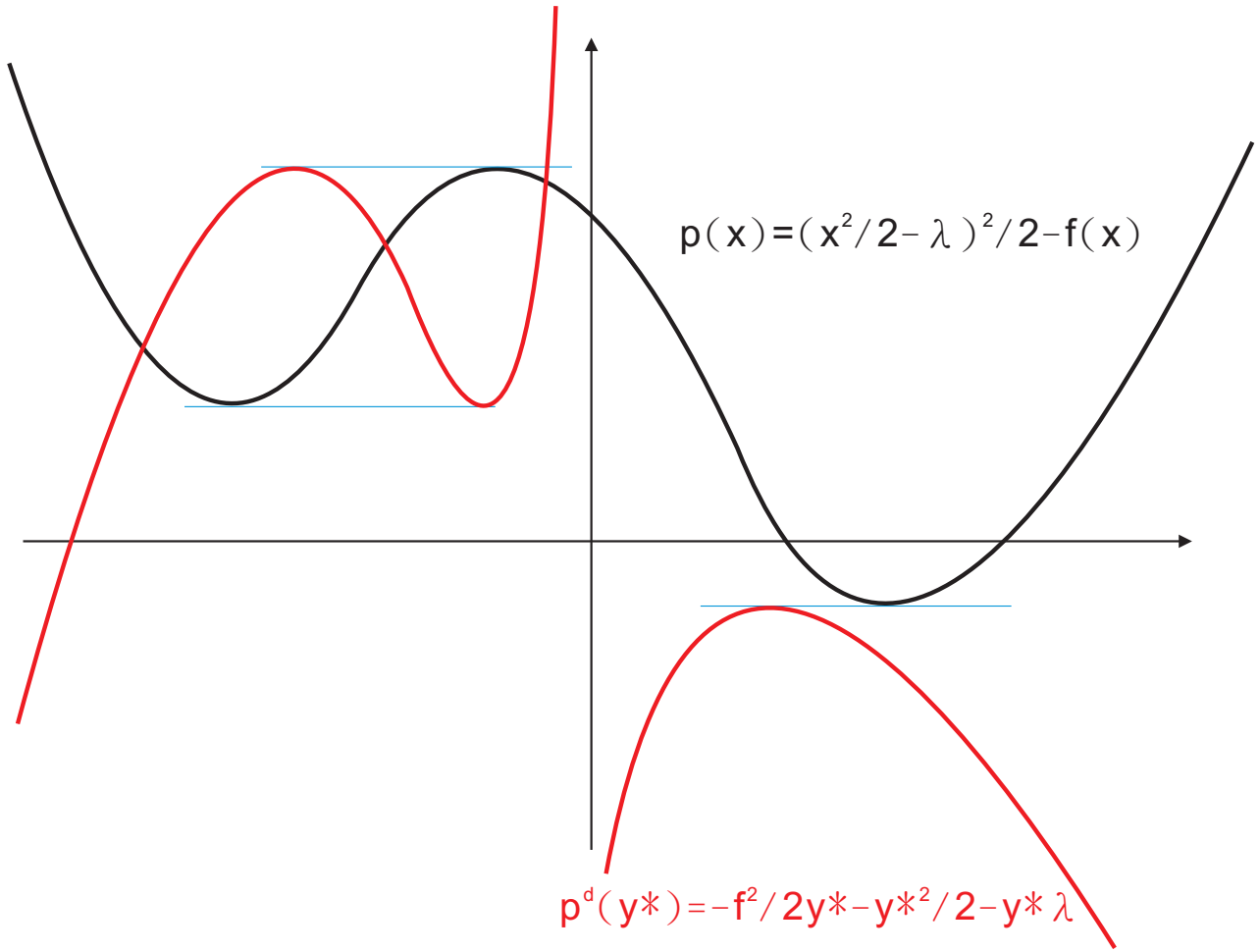
Consider

$$p(x) = \frac{1}{2}(\frac{1}{2}x^2 - \lambda)^2 - fx,$$

where f is a given real value and $\lambda > 0$. Notice that $p(x)$ is a non-convex polynomial of degree 4. His canonical dual functional is derived and defined as follows:

$$\begin{aligned} p^d(y^*) &= L(y^*, y^*) \\ &= (\frac{f^2}{2y^*} - \lambda)y^* - \frac{1}{2}(y^*)^2 - \frac{f^2}{y^*} \\ &= -\frac{f^2}{2y^*} - \frac{1}{2}(y^*)^2 - \lambda y^*. \end{aligned}$$

Graphic interpretation for canonical dual



It can be proved (and also can be seen from the graph) that, both $p(x)$ and $p^d(y^*)$ have three critical points:

1. On $x > 0, y^* > 0$, the strong duality and the weak duality hold (and thus called the perfect duality) so that $\min_{x \in R} p(x) = \min_{x>0} p(x) = \max_{y^*>0} p^d(y^*)$. The global minimum of $p(x)$ can be easily obtained by maximizing the concave function $p^d(y^*)$ over the domain $y^* > 0$.
2. On $x < 0, y^* < 0$, the other two pairs of primal-dual local solutions have only strong duality, but no weak duality.

Algorithms for generalized fractional programming

• J.P. Crouzeix, J.A. Ferland and S. Schaible, "An algorithm for generalized fractional programs," JOTA 47 (1985)
• J.P. Crouzeix and J.A. Ferland, "Algorithms for generalized fractional programming," Math. Prog. 52 (1991)
• J.C. Bernard and J.A. Ferland, "Convergence of interval-type algorithms for generalized fractional programming," Math. Prog. 43 (1989)
• Barros, Frenk, Schaible, and Zhang, "A new algorithm for generalized fractional programs," Math. Prog. 72 (1996)
• Barros, Frenk, Schaible, and Zhang, "Using duality to solve generalized fractional programming problems," JOGO 8 (1996)
• Chen, Schaible, Sheu, "Generic algorithm for generalized fractional programming," JOTA, to appear. (2009)

Dinkelbach-type algorithm

- Problem format:

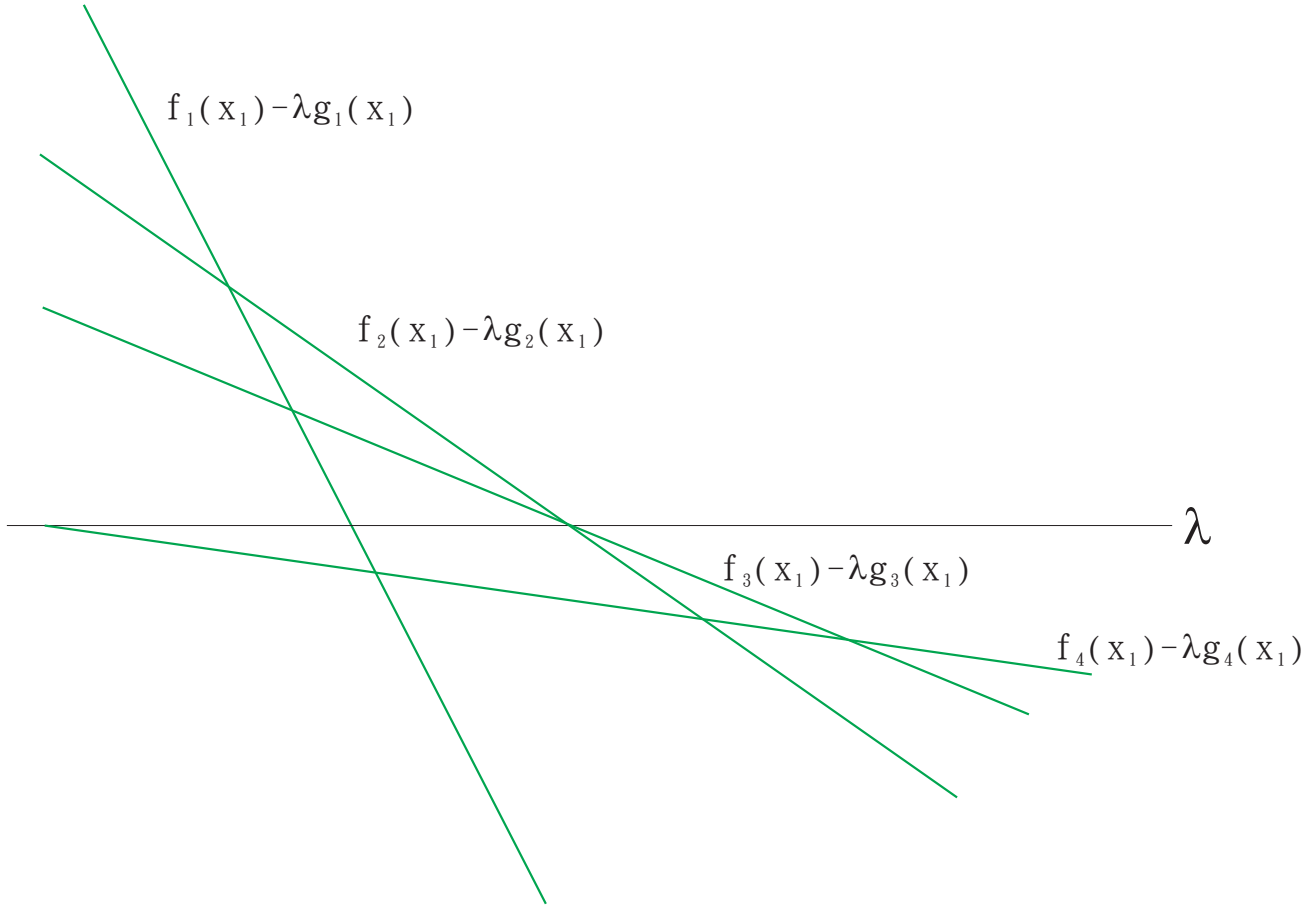
$$(P) \quad \lambda^* = \min_{x \in X} \max_{1 \le i \le n} \{ \frac{f_i(x)}{g_i(x)} \}$$

where X is a nonempty compact subset of $\mathbb{R}^p$, the functions f_i and g_i are continuous on X, and g_i are positive on X.

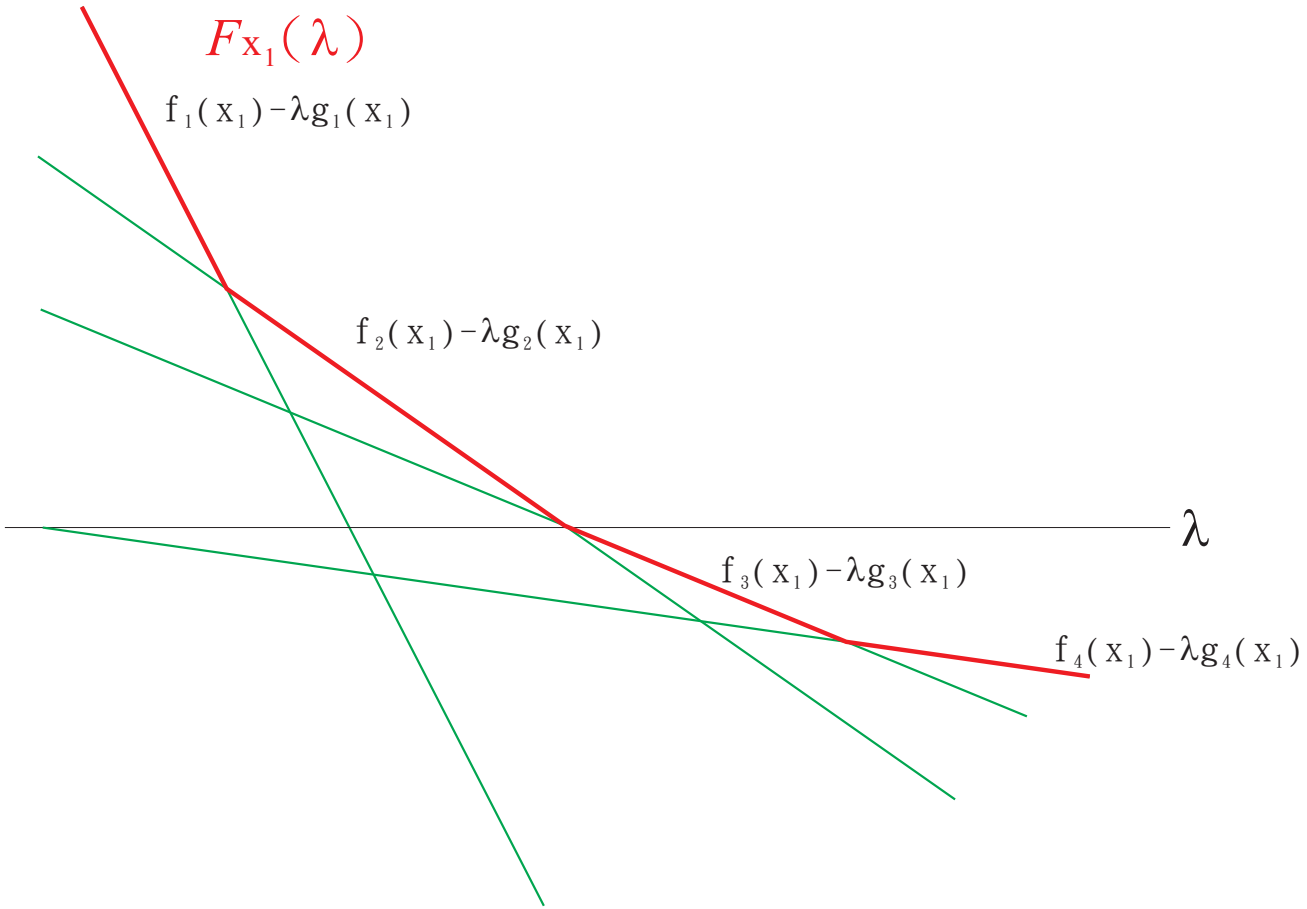
- The "Dinkelbach-type" algorithm considers the following parametric subproblems:

$$(P_\lambda) \quad F(\lambda) = \min_{x \in X} \max_{1 \le i \le n} \{ f_i(x) - \lambda g_i(x) \}.$$

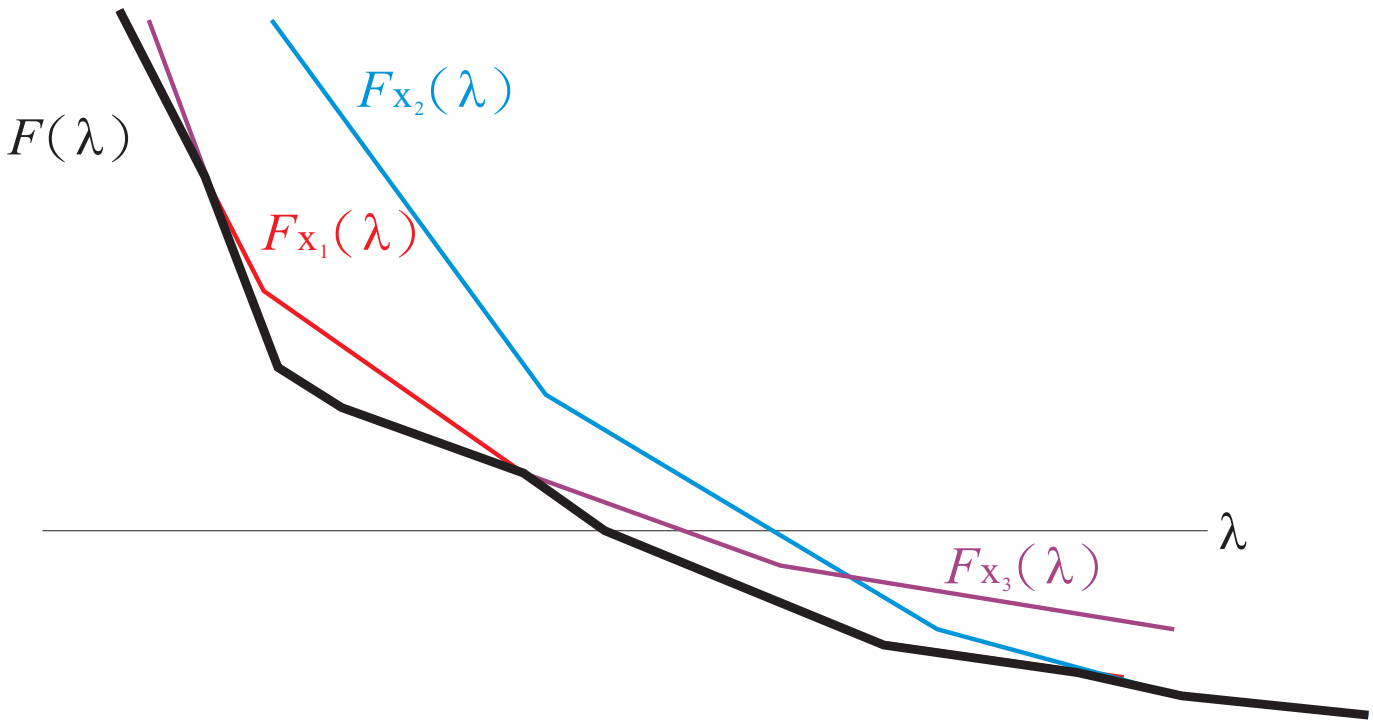
Solving (P) amounts to finding the root of $F(\lambda) = 0$.



- $F(\lambda) = \min\limits_{x \in X} \max\limits_{1 \le i \le n} \{f_i(x) - \lambda g_i(x)\}$.
- Given $x_1 \in X$, we obtain n linear functions in λ.
- The slopes are $\{-g_i(x_1)\}, i = 1, 2, ..., n$.
- $-\infty < -M \le -g_i(x) \le -m < 0$.



- Given $x_1 \in X$, the maximum function $F_{x_1}(\lambda) = \max_i \{f_i(x_1) - \lambda g_i(x_1)\}$ is piecewise linear, convex and decreasing.

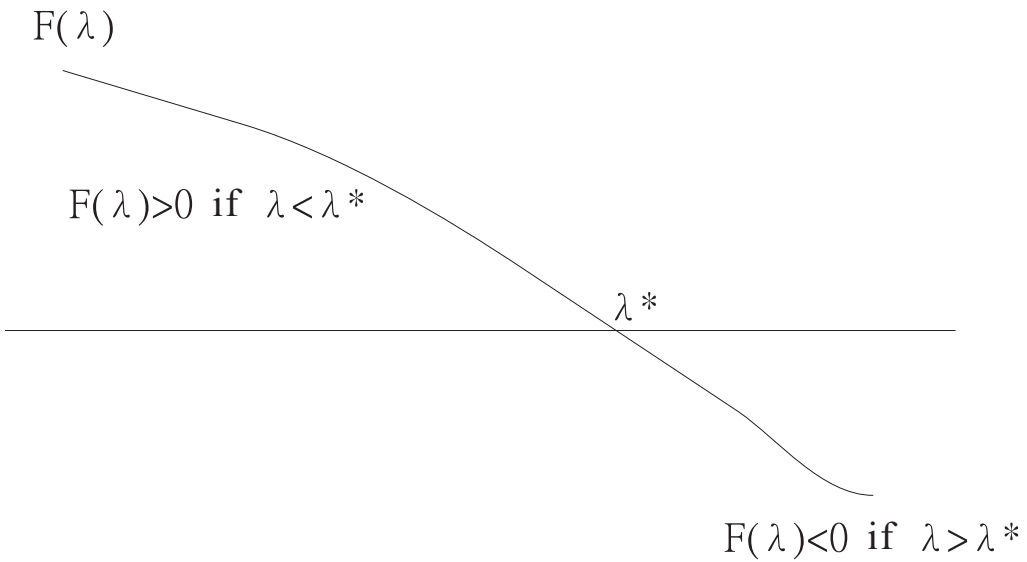


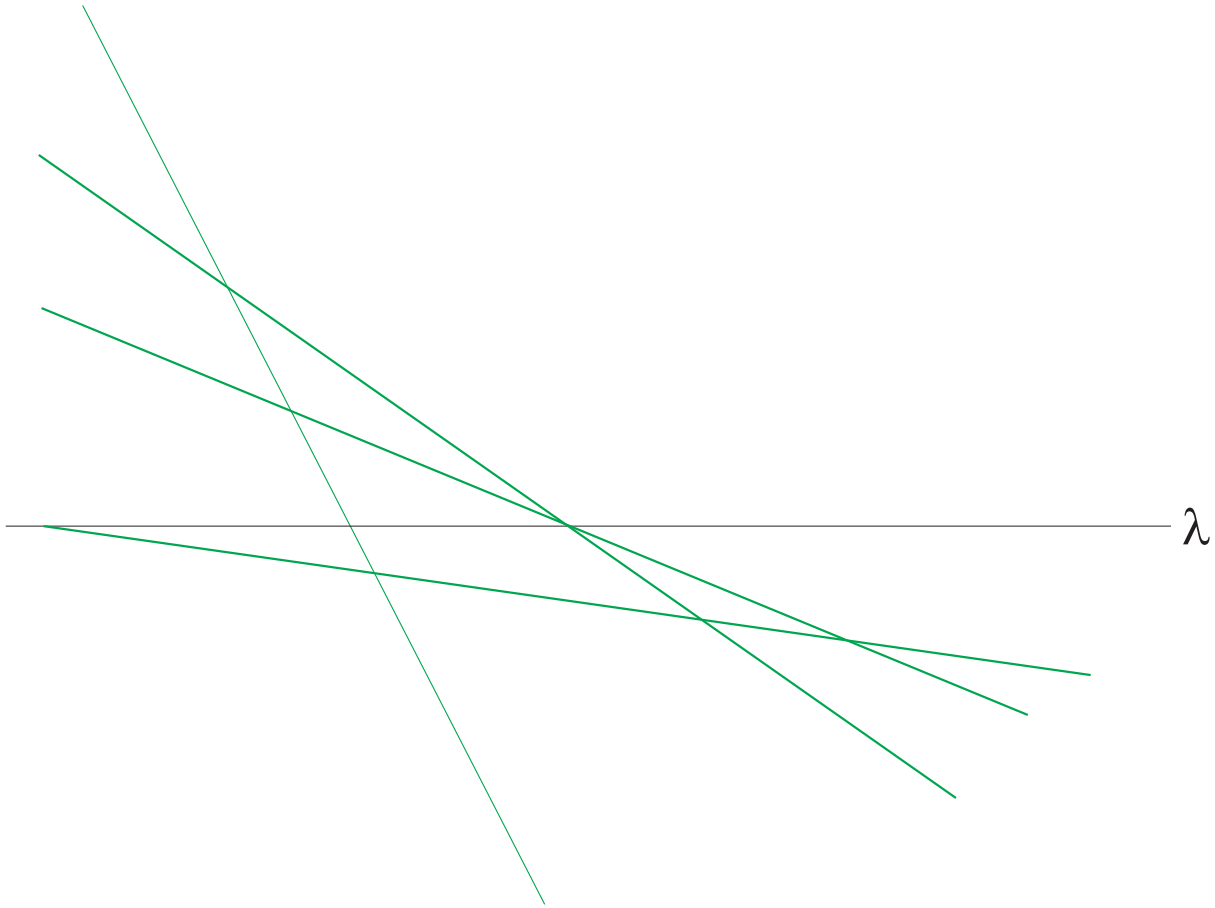
- $F(\lambda) = \min\limits_{x \in X} \max\limits_{1 \le i \le n} \{f_i(x) - \lambda g_i(x)\} = \min\limits_{x \in X} F_x(\lambda)$
- $F(\lambda)$ is the greatest lower bound function of $F_x(\lambda),\ \forall x \in X.$

General properties for $F(\lambda)$

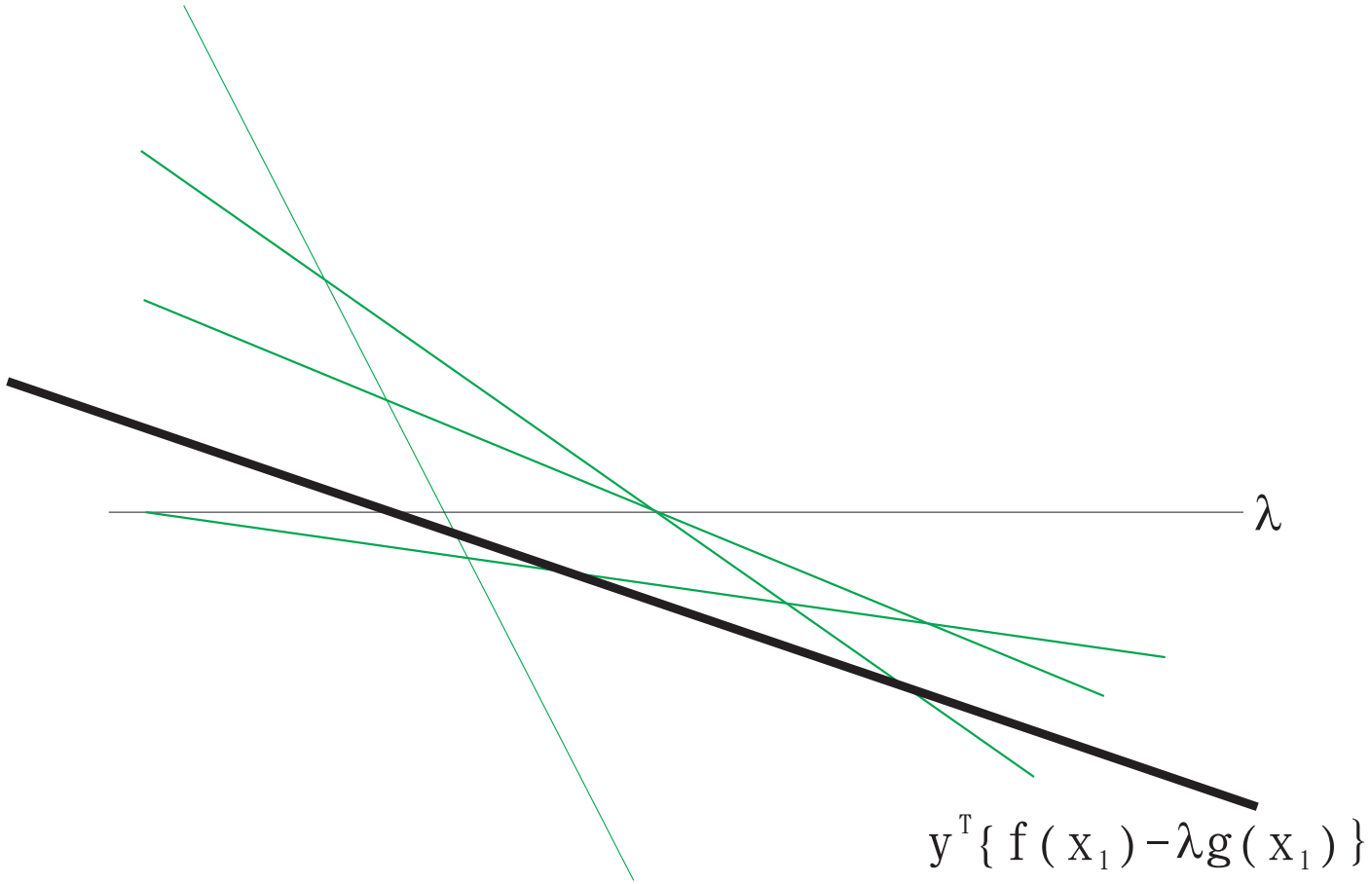
- F is decreasing and continuous.
- Denote the optimal value of (P) to be λ^*. Then,

$F(\lambda) < 0$, for $\lambda > \lambda^*$; $F(\lambda) > 0$, for $\lambda < \lambda^*$; and $F(\lambda^*) = 0$.

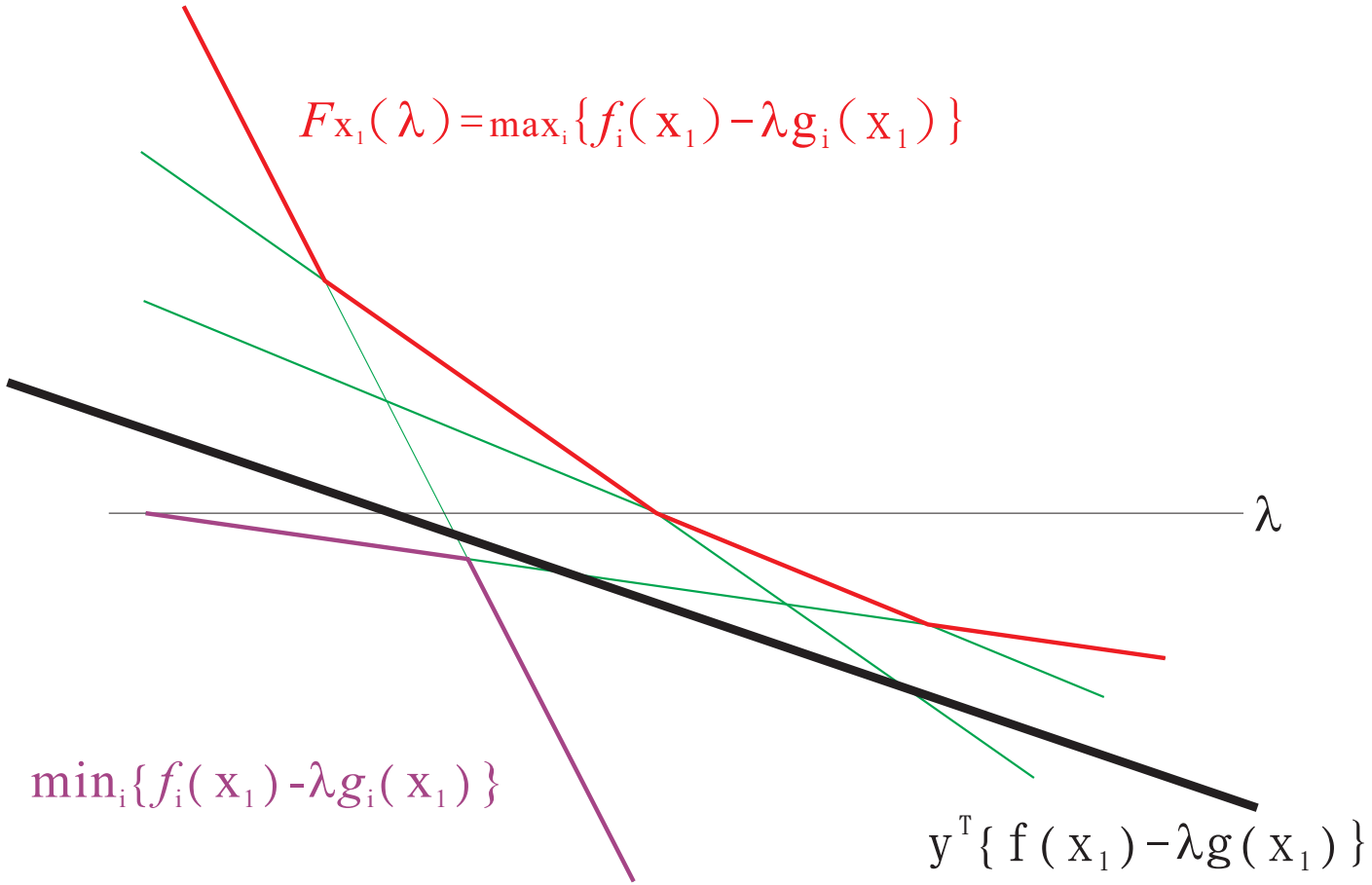




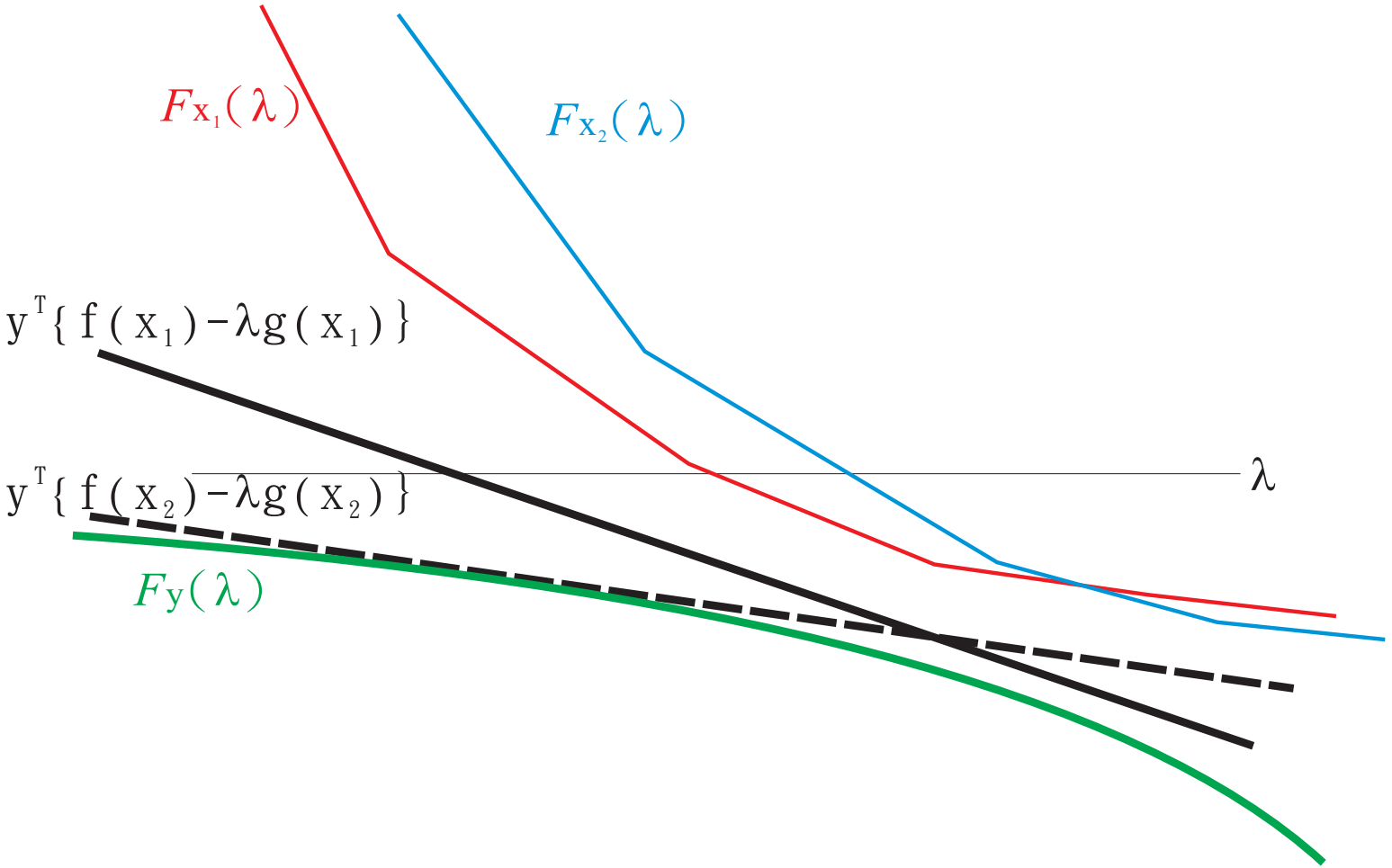
- $y \in Y = \{y_i \in \mathbb{R}^n | \sum y_i = 1, y_i \geq 0\}$.
- $y^T\{f(x_1) - \lambda g(x_1)\}$ is the surrogate of $\{f_i(x_1) - \lambda g_i(x_1)\}_{i=1}^n$.



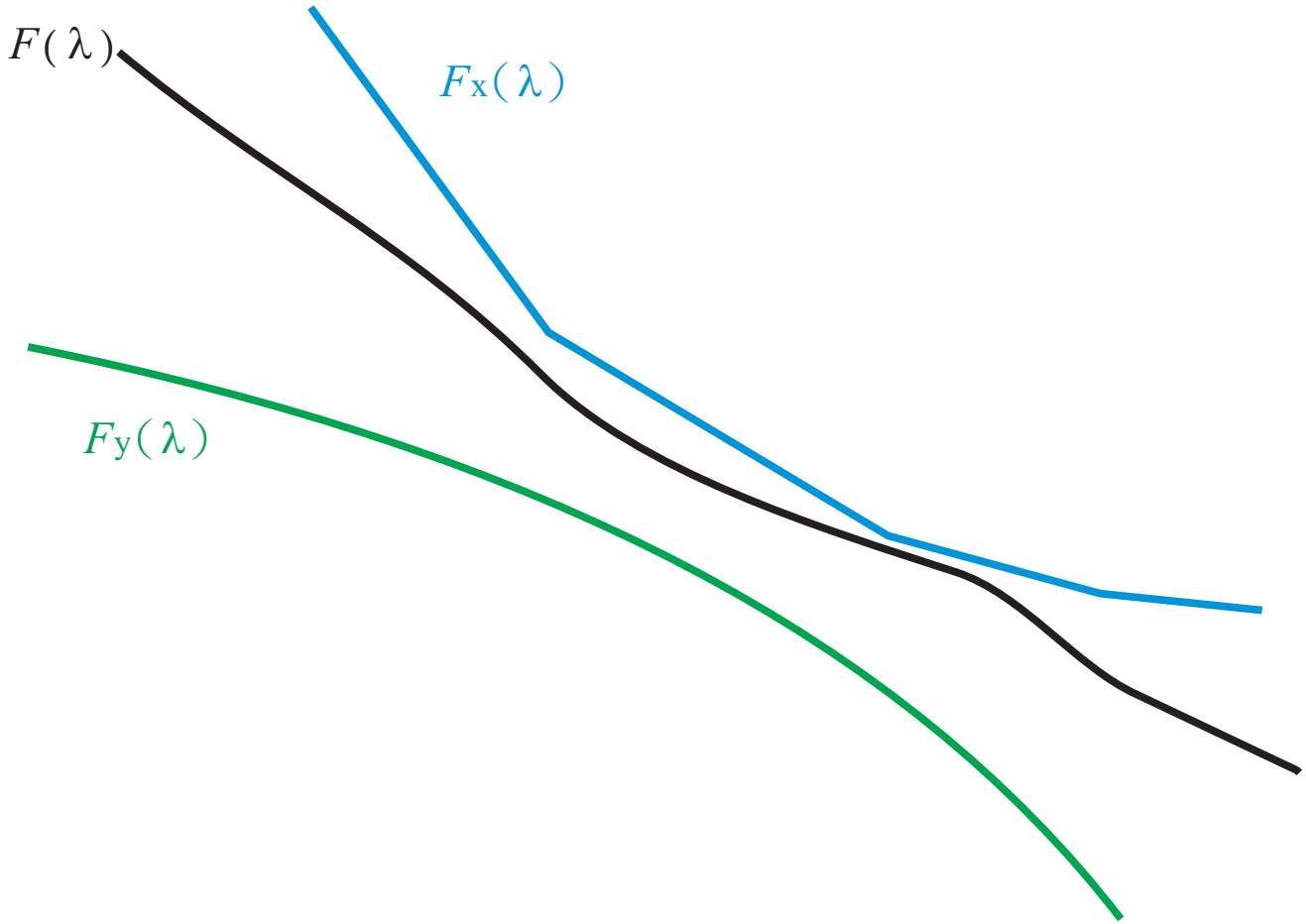
- $y \in Y = \{y_i \in \mathbb{R}^n | \sum y_i = 1, y_i \geq 0\}$.
- $y^T\{f(x_1) - \lambda g(x_1)\}$ is the surrogate of $\{f_i(x_1) - \lambda g_i(x_1)\}_{i=1}^n$.



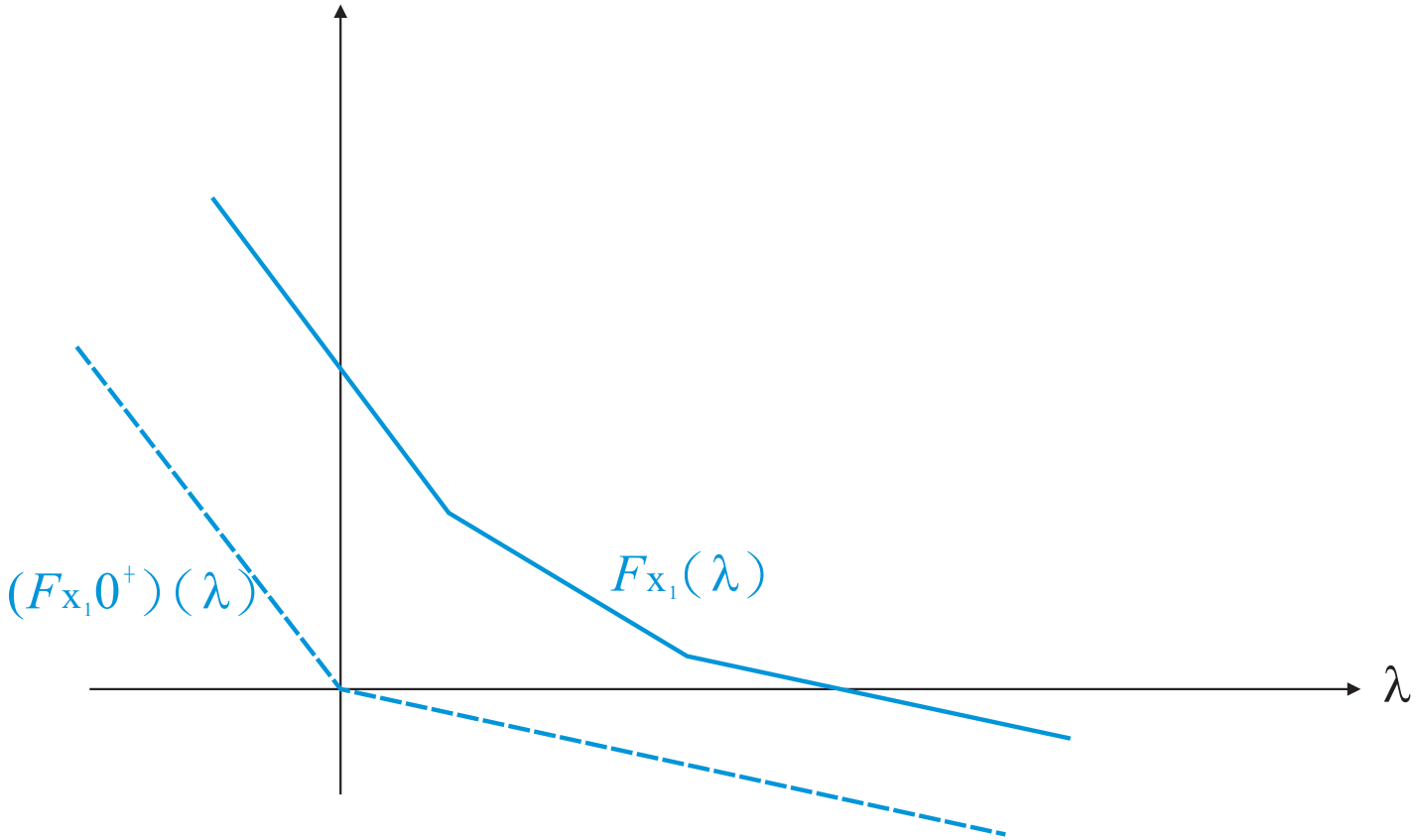
- $y \in Y = \{y_i \in \mathbb{R}^n | \sum y_i = 1, y_i \geq 0\}$.
- $y^T\{f(x_1) - \lambda g(x_1)\}$ is the surrogate of $\{f_i(x_1) - \lambda g_i(x_1)\}_{i=1}^n$.



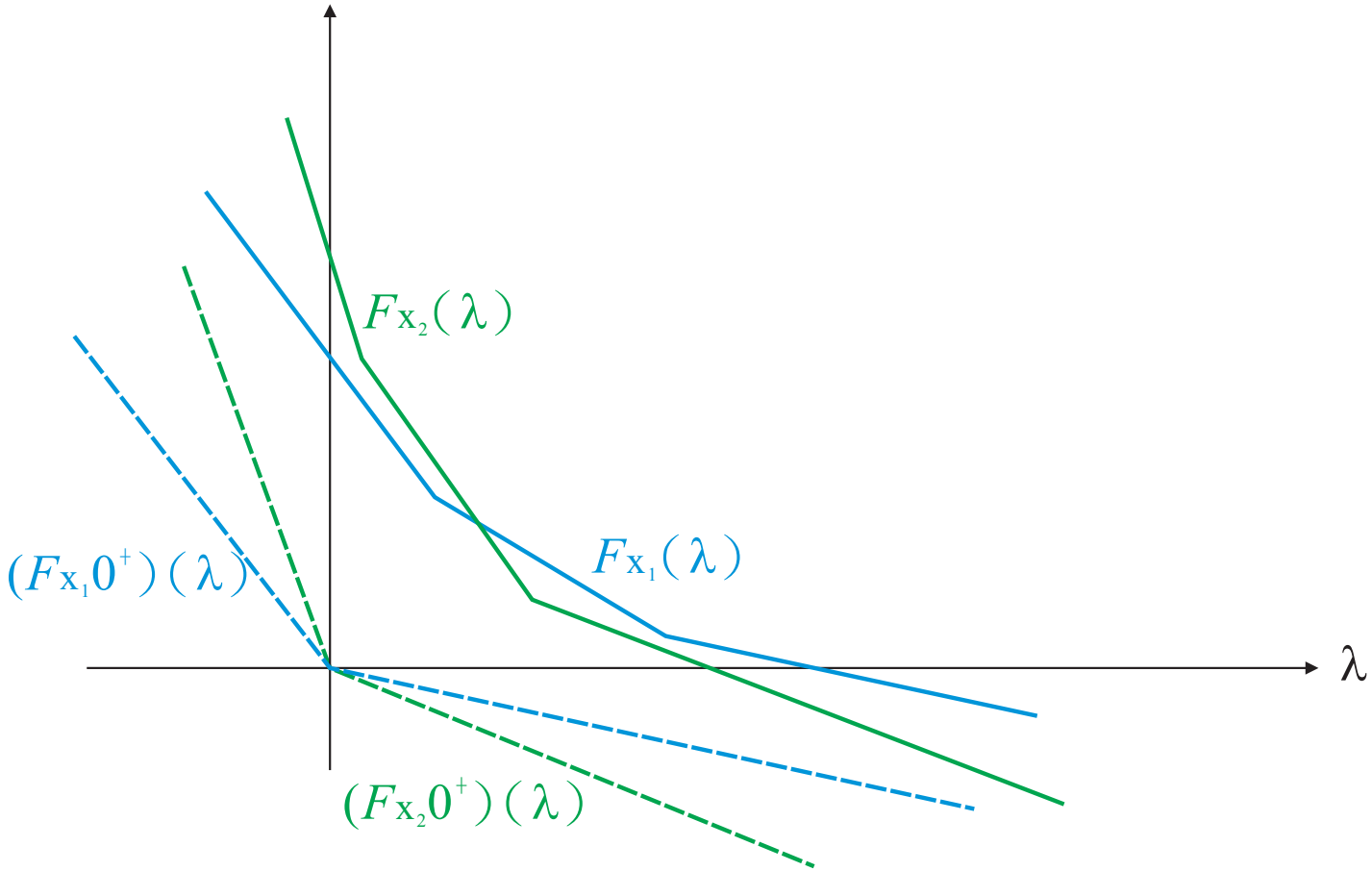
- $F_y(\lambda) = \min_x y^T\{f(x) - \lambda g(x)\} \leq \min_x \max_i\{f_i(x) - \lambda g_i(x)\} = F(\lambda)$



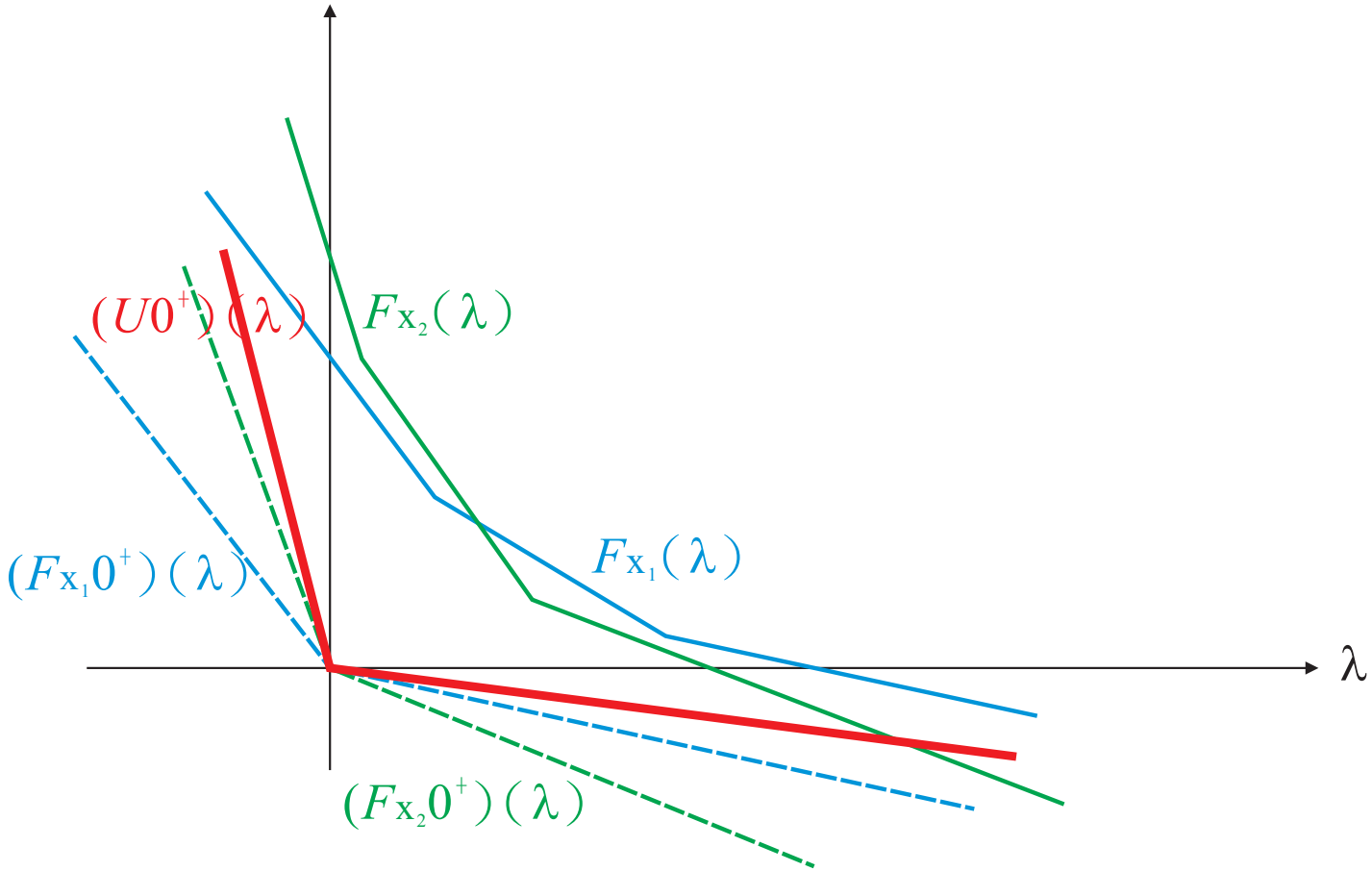
• $F_y(\lambda)$ is concave, decreasing and $F_y(\lambda) \leq F(\lambda) \leq F_x(\lambda)$, $\forall x \in X, y \in Y, \lambda \in \mathbb{R}$.



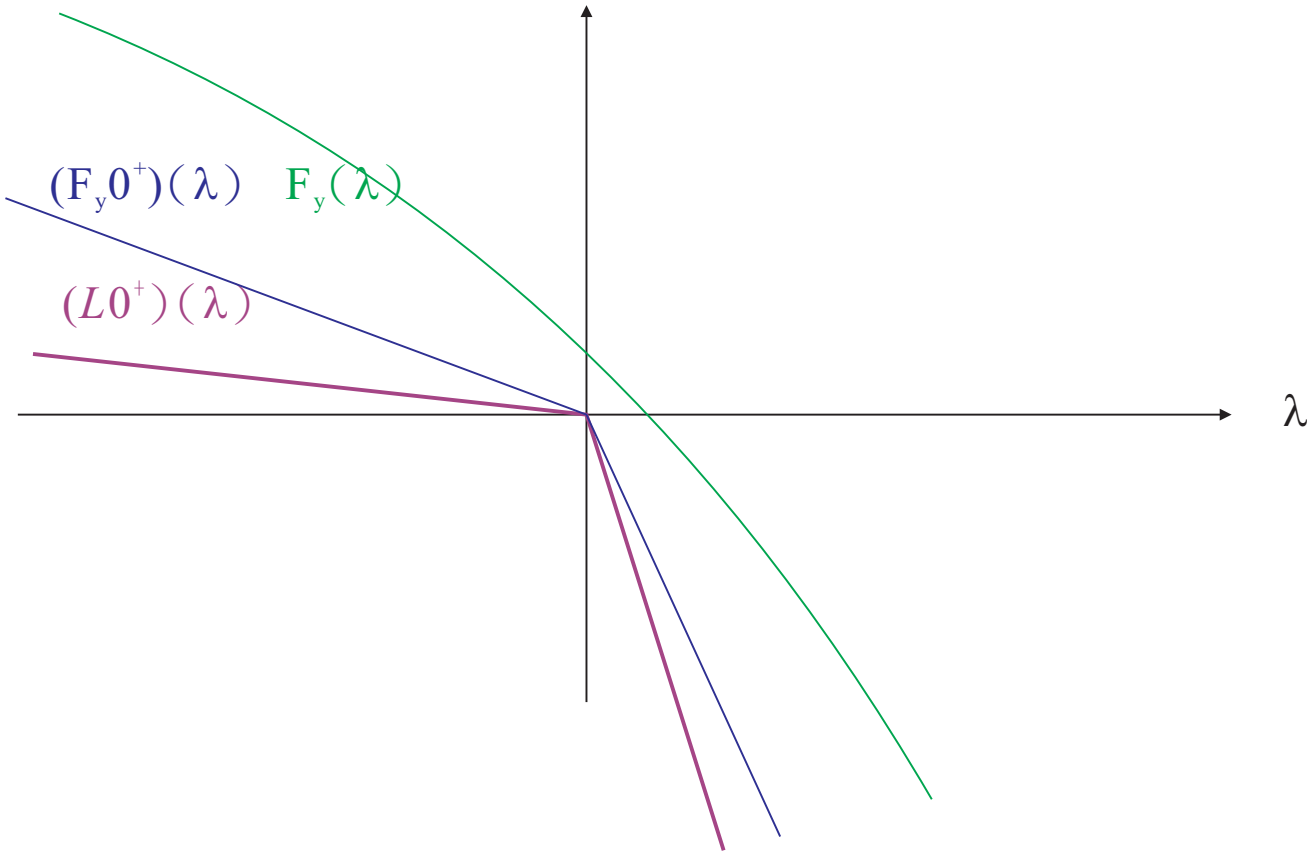
- $(U0^+)(\lambda) = \max\{-m\lambda, -M\lambda\}$.
- $m = \min\limits_{x\in X}\{\min\limits_{1\le i\le n} g_i(x)\},\ M = \max\limits_{x\in X}\{\max\limits_{\le i\le n} g_i(x)\}$



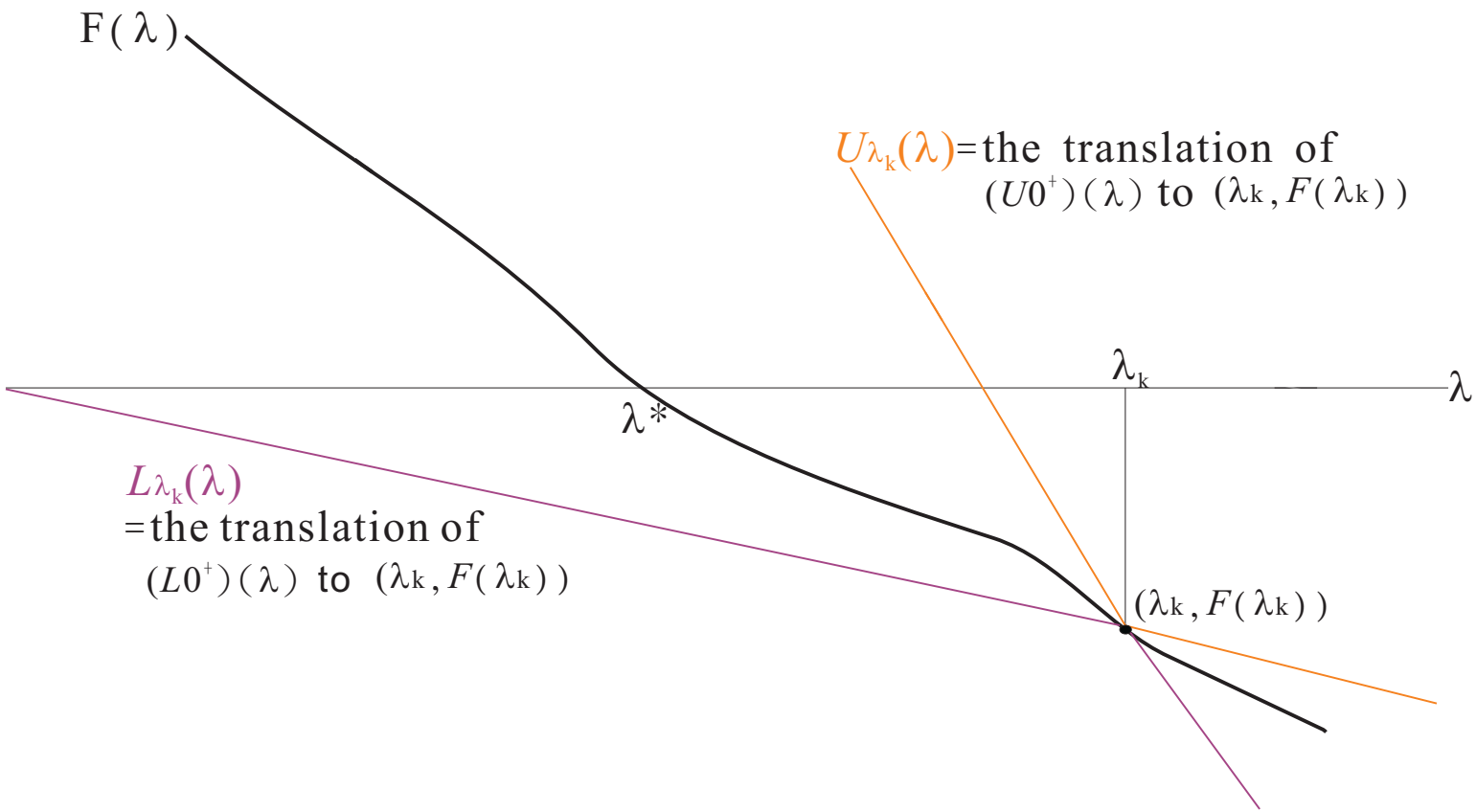
- $(U0^+)(\lambda) = \max\{-m\lambda, -M\lambda\}$.
- $m = \min\limits_{x\in X}\{\min\limits_{1\le i\le n} g_i(x)\},\ M = \max\limits_{x\in X}\{\max\limits_{\le i\le n} g_i(x)\}$

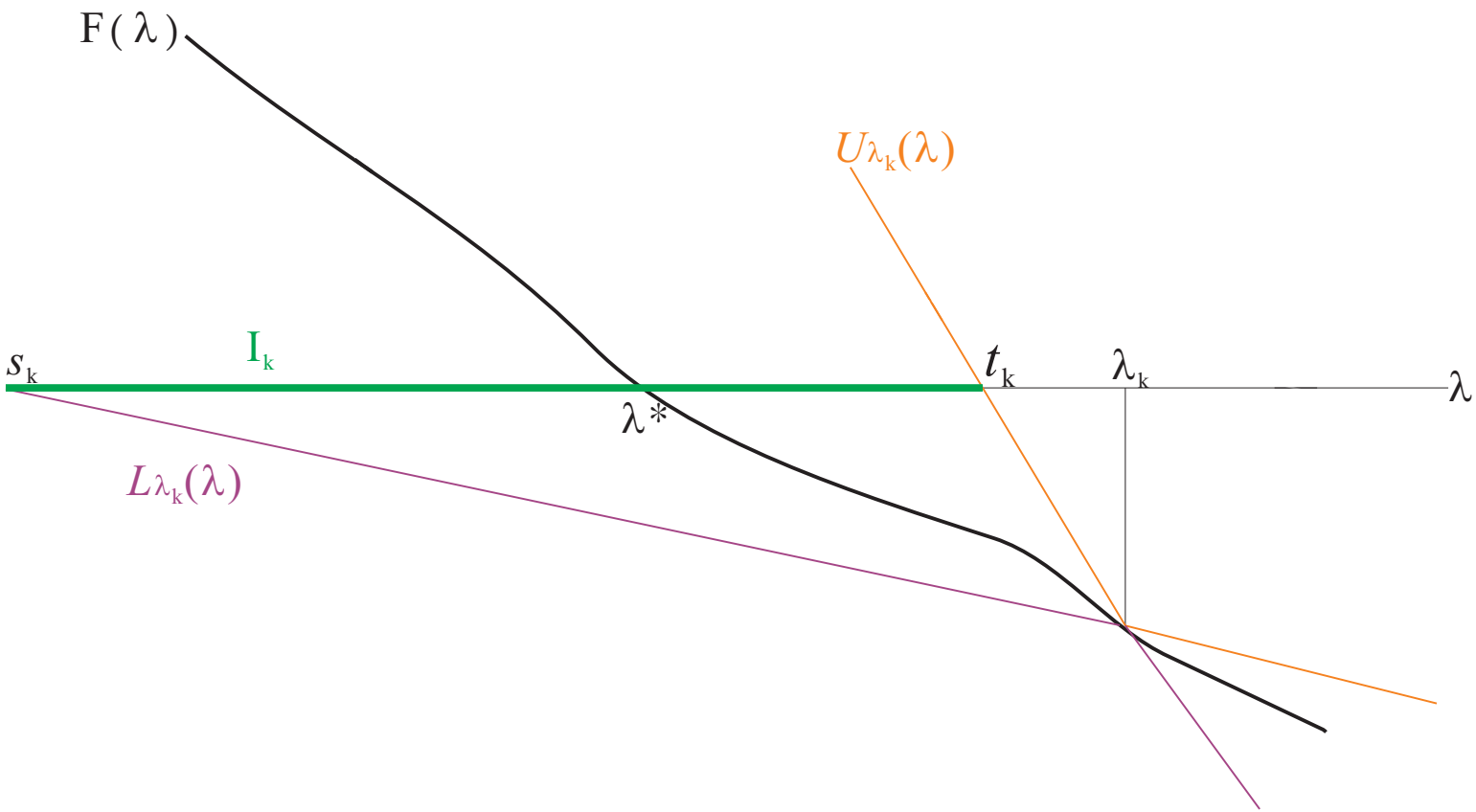


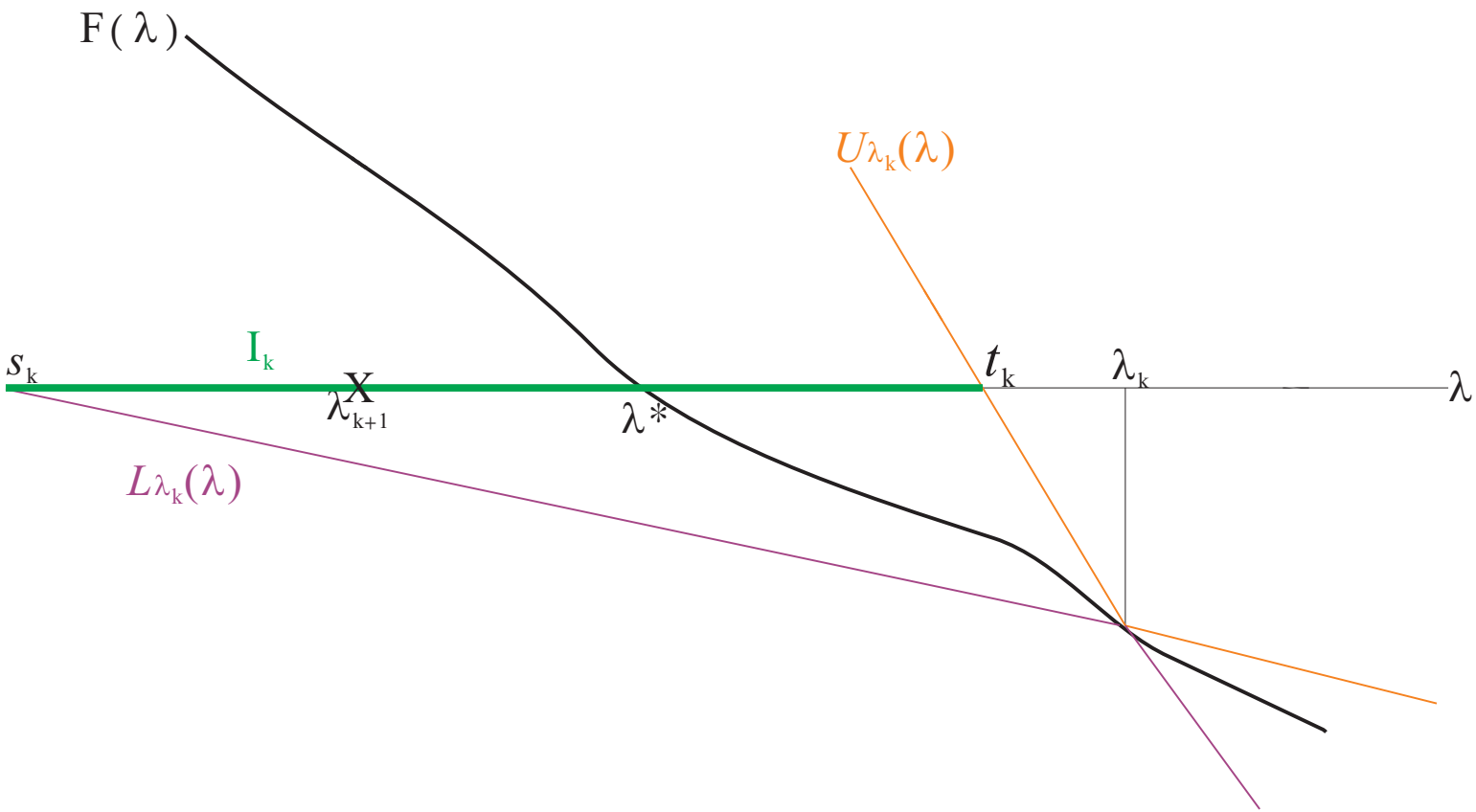
- $(U0^+)(\lambda) = \max\{-m\lambda, -M\lambda\}$.
- $m = \min_{x\in X}\{\min_{1\le i\le n} g_i(x)\}$, $M = \max_{x\in X}\{\max_{\le i\le n} g_i(x)\}$

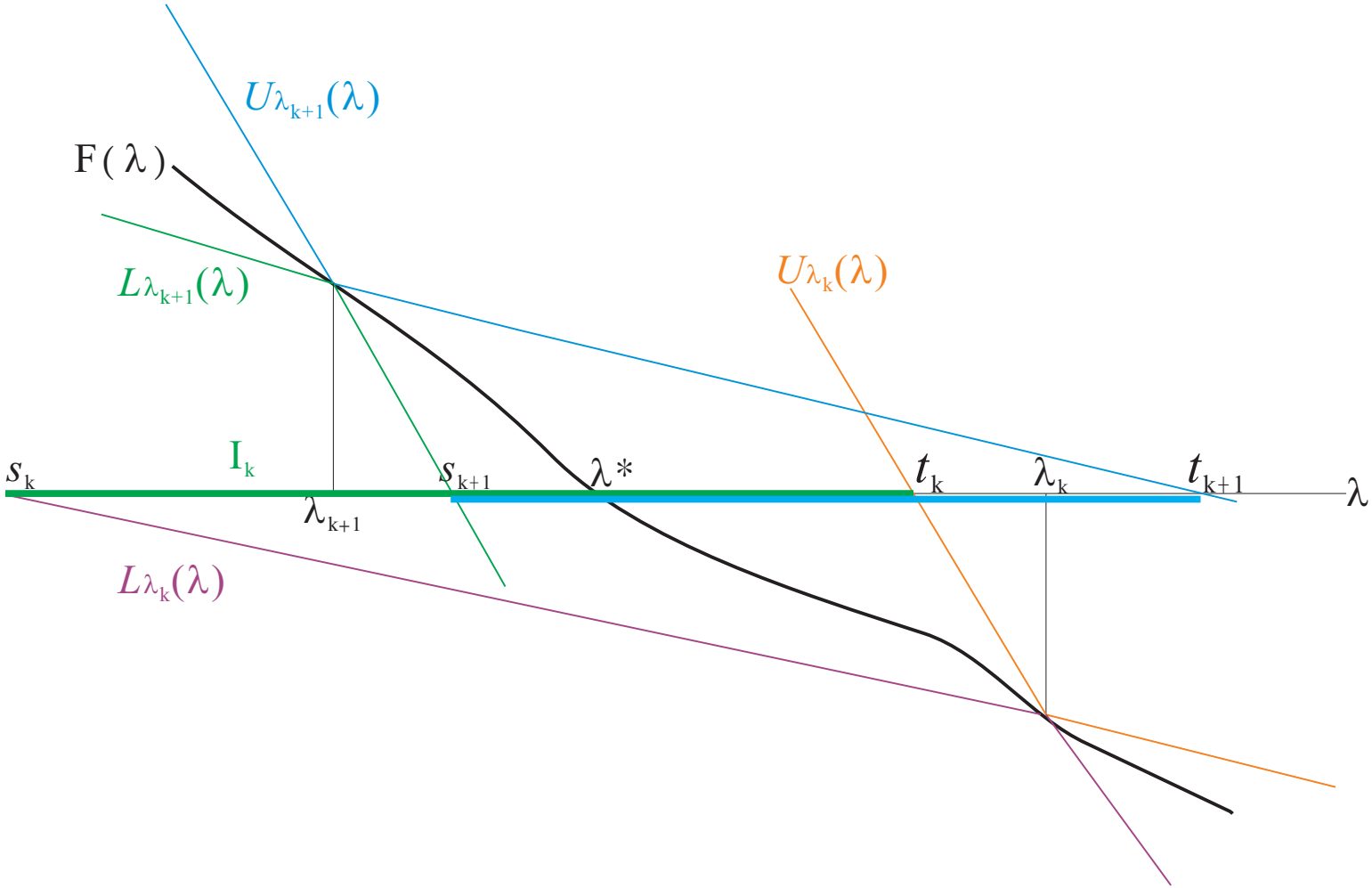


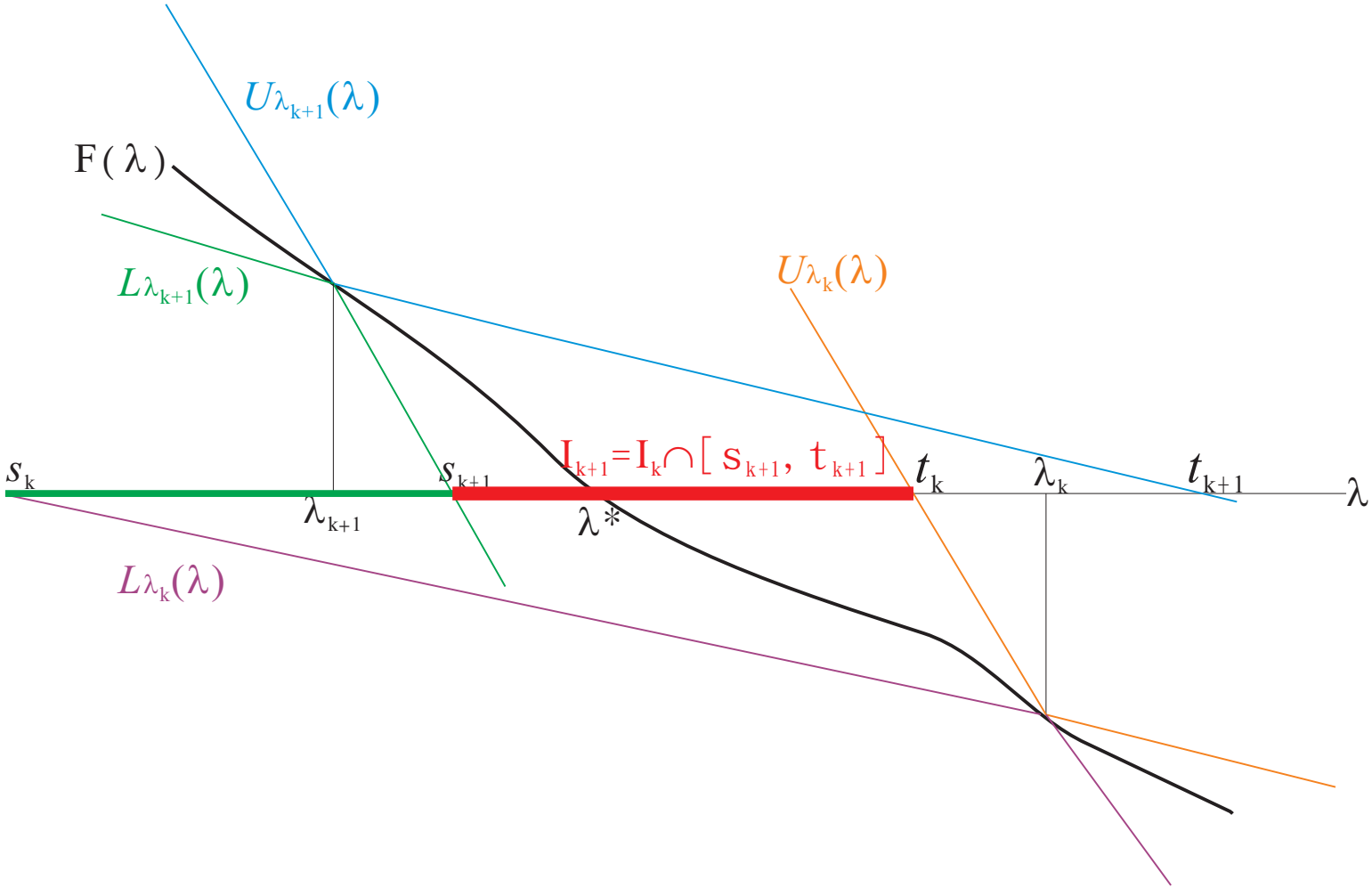
- $(L0^+)(\lambda) = \max\{-m\lambda, -M\lambda\}.$

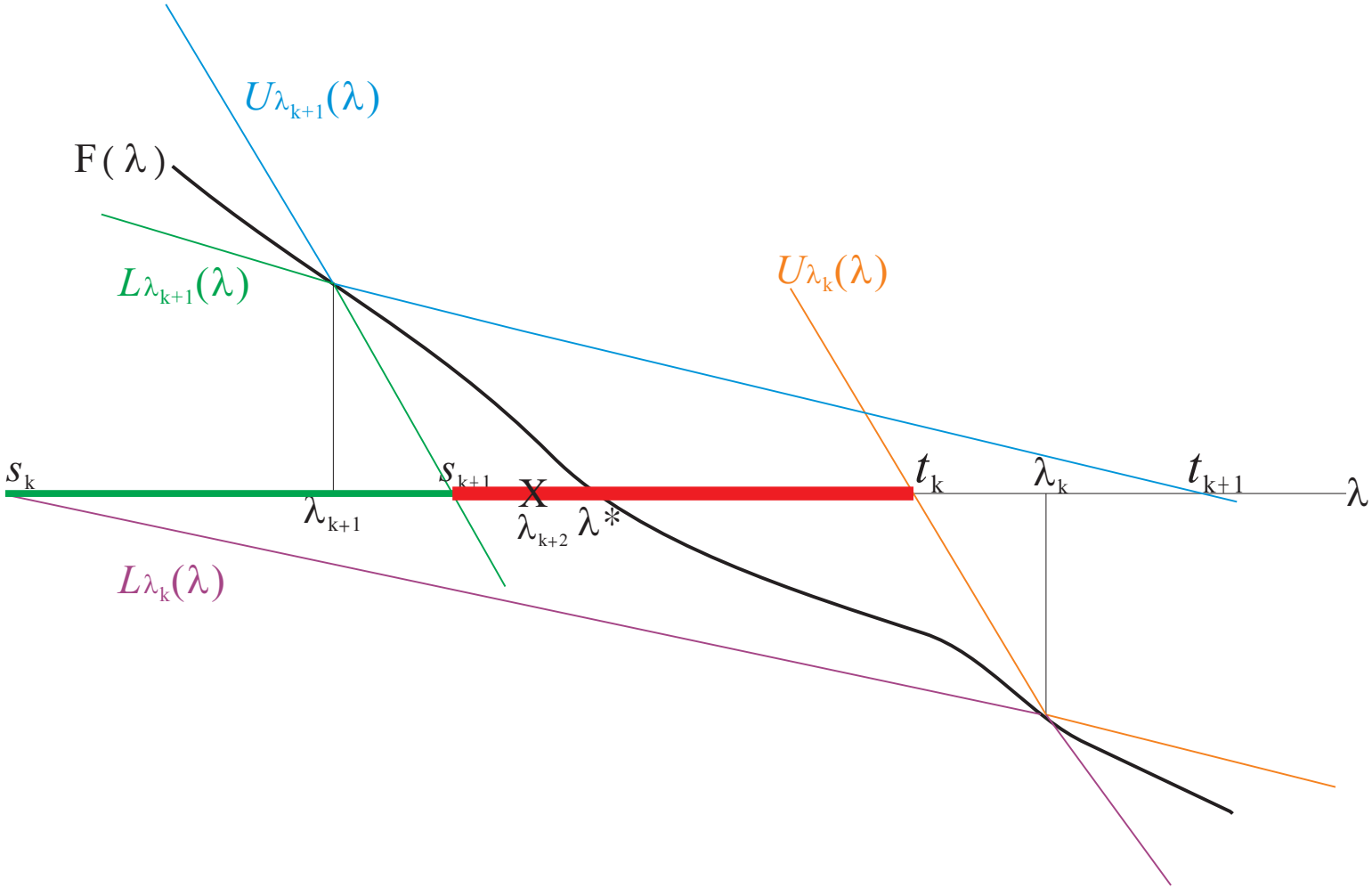

F(λ)
$U\lambda_k(\lambda)$=the translation of
$(U0^+)(\lambda)$ to $(\lambda_k, F(\lambda_k))$
λ_k
λ
λ*
$L\lambda_k(\lambda)$
=the translation of
$(L0^+)(\lambda)$ to $(\lambda_k, F(\lambda_k))$
$(\lambda_k, F(\lambda_k))$


F(λ)
$U\lambda_k(\lambda)$
s_k
I_k
t_k
λ_k
λ
λ*
$L\lambda_k(\lambda)$


F(λ)
$U\lambda_k(\lambda)$
I_k
s_k
λ_{k+1}
λ^*
t_k
λ_k
λ
$L\lambda_k(\lambda)$


$U\lambda_{k+1}(\lambda)$
F (λ)
$L\lambda_{k+1}(\lambda)$
$U\lambda_k(\lambda)$
s_k
I_k
s_{k+1}
λ*
t_k
λ_k
t_{k+1}
λ
λ_{k+1}
$L\lambda_k(\lambda)$


$U\lambda_{k+1}(\lambda)$
F (λ)
$L\lambda_{k+1}(\lambda)$
$U\lambda_k(\lambda)$
$I_{k+1}=I_k\cap[\,s_{k+1},\,t_{k+1}\,]$
s_k
s_{k+1}
t_k
λ_k
t_{k+1}
λ
λ_{k+1}
λ*
$L\lambda_k(\lambda)$


$U\lambda_{k+1}(\lambda)$
F (λ)
$L\lambda_{k+1}(\lambda)$
$U\lambda_k(\lambda)$
s_k
s_{k+1}
X
t_k
λ_k
t_{k+1}
λ
λ_{k+1}
λ_{k+2}
λ*
$L\lambda_k(\lambda)$

Generic Dinkelbach-type algorithm

Step0 (Initialization) Let $I_1 = (-\infty, +\infty)$, $k = 1$ and choose $\lambda_1 \in I_1$.

Step1 Determine an optimal solution x^k for

$$\begin{aligned} F(\lambda_k) &= \min_{x \in X} \{ \max_{1 \le i \le n} \{ f_i(x) - \lambda_k g_i(x) \} \} \\ &= \max_{1 \le i \le n} \{ f_i(x^k) - \lambda_k g_i(x^k) \}. \end{aligned}$$

If $F(\lambda_k) = 0$, then stop.

Generic Dinkelbach-type algorithm

Step2 Determine an interval of convergence.
Compute

$$s_k = \min\{\lambda_k + \frac{F(\lambda_k)}{m},\ \lambda_k + \frac{F(\lambda_k)}{M}\}$$

and

$$t_k = \max\{\lambda_k + \frac{F(\lambda_k)}{m},\ \lambda_k + \frac{F(\lambda_k)}{M}\}.$$

Update $I_{k+1} = I_k \cap [s_k, t_k]$.

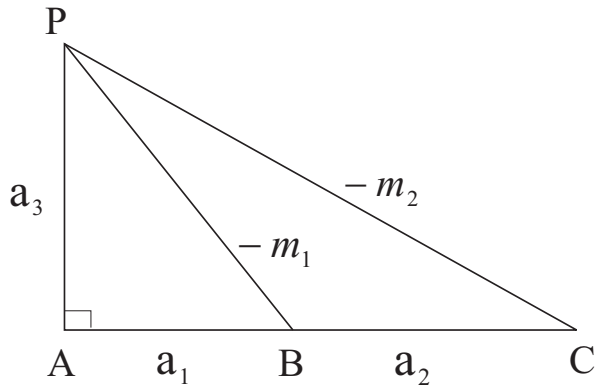
Step3 Choose $\lambda_{k+1} \in I_{k+1}$.
Replace k by $k+1$. Go to Step1.

Theorem

Let $\{\lambda_k\}$ *and* $\{I_k\}_{k=1}^{\infty}$ *be infinite sequences generated by the generic Dinkelbach-type algorithm. Then,* $|I_{k+1}| \leq \left(1 - \frac{m}{M}\right)|I_k|$ *and* $\lim_{k\to\infty} I_k = \{\lambda^*\}$. *Moreover, the two subsequences* $\{\lambda_j\}_{j\in J}$ *and* $\{\lambda_j\}_{j\notin J}$ *converge to* λ^* *linearly, from the right and the left, respectively.*

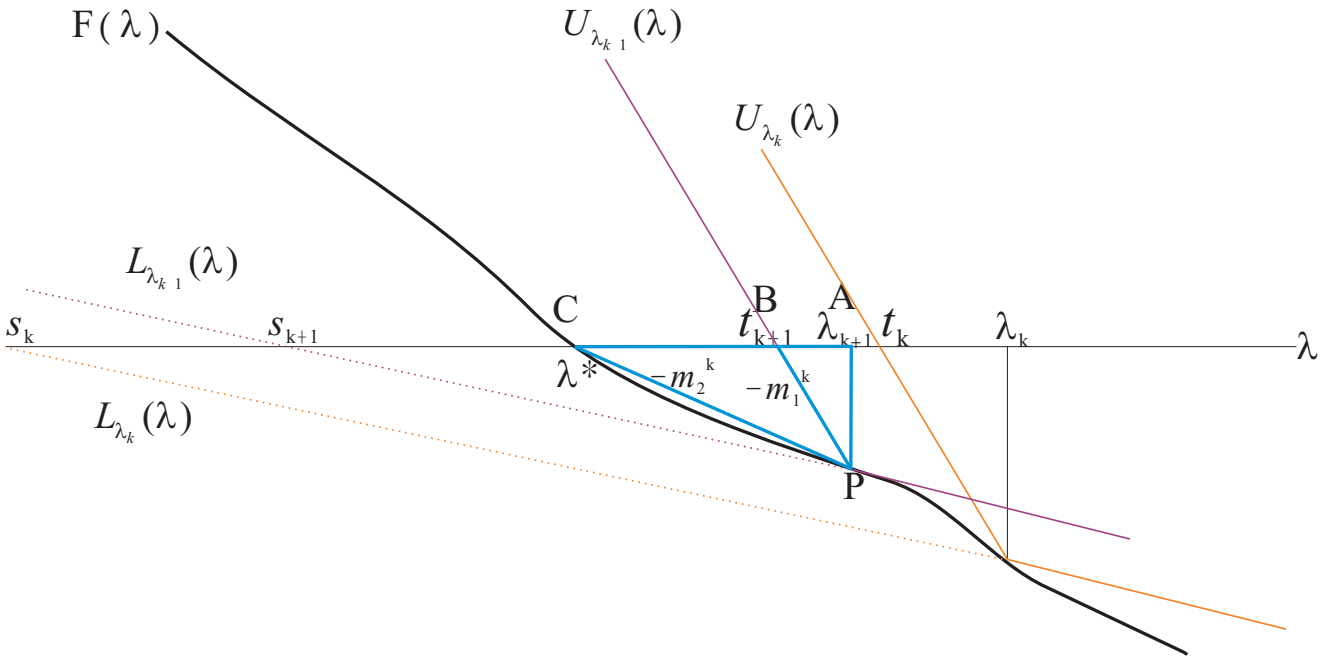
† This theorem, and therefore the generic algorithm unifies all the Dinkelbach-type algorithms before 1996.
‡ The convergence proof is carried out through geometric observations and fundamental properties of convex functions. Consequently, the classical results are either refined or strengthened with new insights.

Lemma 1

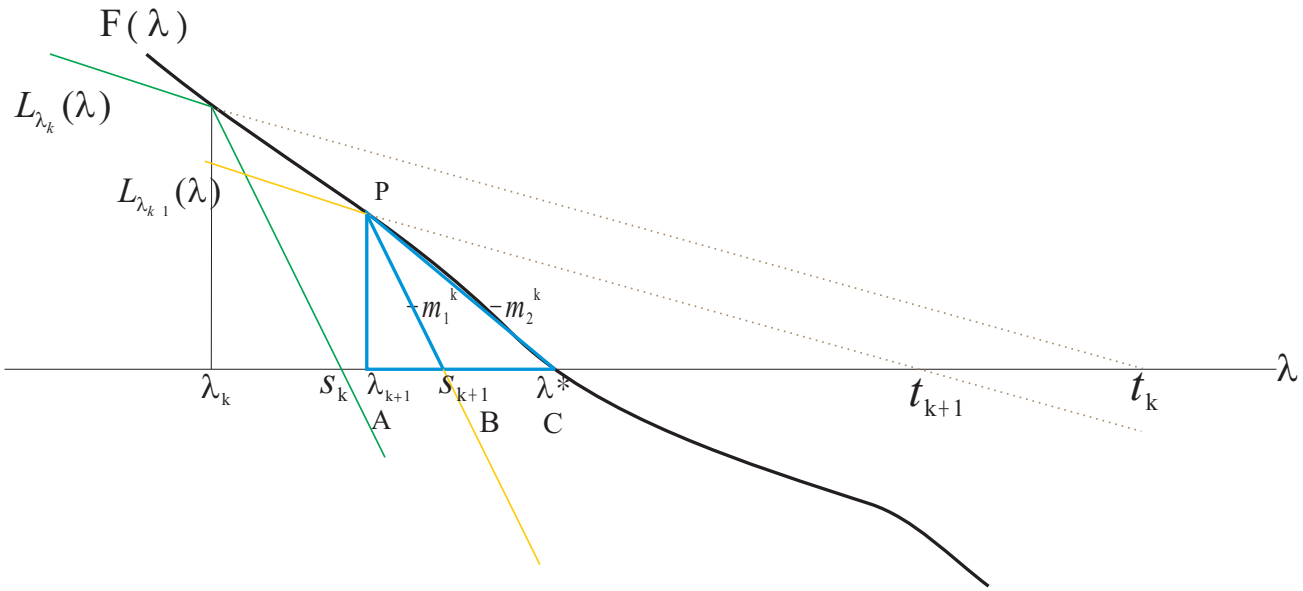


$$\frac{a_2}{a_1+a_2} = 1 - \frac{m_2}{m_1}$$

Consider a triangle PAC with $\overline{PA} \perp \overline{AC}$ and B lies on $\overline{AC}$. Let $\overline{AB} = a_1$, $\overline{BC} = a_2$, $\overline{PA} = a_3$. Suppose that the slopes of the line PB, PC are $-m_1$ and $-m_2$ respectively. Convergence is assured if $\overline{PB}$ is not vertical and $\overline{PC}$ is not horizontal.



- To estimate m_1^k and m_2^k for the convergence of $\{t_k\}$.
- $\lambda_{k+1} - \lambda^* \leq (1 - \frac{m}{M})(\lambda_k - \lambda^*), \ \forall k.$



- To estimate m_1^k and m_2^k for the convergence of $\{s_k\}$.

Methods for the sum-of-ratios problem

Consider the sum-of-ratios problem of the following form:

$$\min_{x \in D} \sum_{s=1}^{q} \frac{f_s(x)}{g_s(x)}$$

where $f_s : R^n \longrightarrow R$ and $g_s : R^n \longrightarrow R$ are continuous on D and $g_s(x) > 0, \ \forall x \in D, \ s = 1, 2, \ldots, q$, where D is a compact connected subset in R^n.

- This is a NP-complete problem due to Freund and Jarre. (JOGO 19, 2001)
- Most studies have focused on the sum-of-linear ratios problem. (E.g., Benson, JOTA 121, 2004; Konno and Abe, JOGO 15, 1999; Konno and Fukaishi, JOGO 18, 2000; Kuno, JOGO 22, 2002.)
- For the sum-of-nonlinear-ratios, see for example, Benson, JMAA 263, 2001; Benson, JOTA 112, 2002; Benson JOGO 22, 2002; Phuong and Tuy, JOGO 26, 2003.
- The branch-and-bound approach is the most popular. However, due to the combinatorial nature, it is quite difficult to handle a high dimensional feasible region or go beyond the sum of ten linear ratios.

Stochastic Search Algorithm

- A sort of method equipped with a random mechanism.
- The term "randomness" does not mean to find the global optimum totally by chance. Indeed, most probabilistic algorithms adopt deterministic policies to direct the local search, whereas repeatedly random sampling is used only to escape non-global local optima.
- Random sampling does not require much from the problem structure. Therefore, it is robust to hard core problems.
- Commonly used stochastic algorithms include simulated annealing, evolutionary methods, multilevel methods, and partitioning methods. Most methods are not just heuristics but have a proof for convergence in probability.

The Electromagnetism-like method

- A kind of stochastic algorithm developed by Birbil, Fang and Sheu (JOGO 30, 2004) for solving global optimization problems with a box constraint.
- The method draws a finite population of sample particles from the domain; associate each of them a "charge" according to their objective values; and move the particles by mimicking the electromagnetism theory of physics.
- The method is applied to the image space of f_s, g_s by decomposing (P) into a series of single ratio problems each of which is solved by the Dinkelbach algorithm.
- Large scale numerical experiments on problems up to sum of eight linear ratios with a thousand variables are reported. (By Wu, Sheu, Birble, to appear in JOGO 2008).

The canonical dual approach†

The problem is to minimize the sum of a quadratic function and the ratio of two quadratic functions:

$$(P) \qquad \min_{x\in\chi} f(x) + \frac{g(x)}{h(x)}$$

where $f(x) = \frac{1}{2}x^tQx - p^tx$, $g(x) = \frac{1}{2}x^tGx$, $h(x) = \frac{1}{2}x^tHx - b^tx$ with $Q \in R^{n\times n}$ being symmetric, $G \in R^{n\times n}$ symmetric positive semi-definite, $H \in R^{n\times n}$ negative definite and $f, b \in R^n$. Assume that $\mu_0^{-1} = h(H^{-1}b) > 0$, $\delta \in (0, \mu_0^{-1}]$, and that the feasible domain $\chi = \{x \in R^n \mid \ h(x) \geq \delta > 0\}$.

† This work is jointly by Fang, Gao, Sheu, and Xin, submitted to JOGO, 2008

- We first parameterize (P) into a family of subproblems:

$$\theta(\mu) = \inf_{x \in \chi_\mu} f(x) + \mu g(x),$$

according to the values of $h(x)$, where $\frac{1}{h(x)} = \mu \in [\mu_0, \delta^{-1}]$ and $\chi_\mu = \{x \in R^n \mid \ h(x) \geq \mu^{-1} \geq \delta > 0\}$, which is a (possibly non-convex) quadratic program subject to one quadratic constraint.

- The original problem is reduced to the one-dimensional problem of minimizing $\theta(\mu)$ over $\mu \in [\mu_0, \delta^{-1}]$.

$$\min_{x \in \chi} f(x) + \frac{g(x)}{h(x)} = \inf_{\mu \in [\mu_0, \delta^{-1}]} \theta(\mu).$$

- The canonical dual functional of $\theta(\mu)$ is derived:

$$P_\mu^d(\sigma) = \frac{\sigma}{\mu} - \frac{1}{2}(p - \sigma b)^T G_\mu^{-1}(\sigma)(p - \sigma b)$$

  over the domain $S_\mu^+ = \{\sigma \geq 0 \mid \ G_\mu(\sigma) \succ 0\}$ where $G_\mu(\sigma) = Q + \mu G - \sigma H$.
- The topological properties of S_μ^+: It is a ray with the boundary point $\max\{0, \sigma_{max}\}$ where σ_{max} represents the maximum root of $\det G_\mu(\sigma) = 0$. If $\sigma_{max} < 0$, the ray is closed and $S_\mu^+ = [0, \infty)$. If $\sigma_{max} \geq 0$, the ray is open and $S_\mu^+ = (\sigma_{max}, \infty)$.

Analytic properties of $P_\mu^d(\sigma) = \frac{\sigma}{\mu} - \frac{1}{2}(p - \sigma b)^T G_\mu^{-1}(\sigma)(p - \sigma b)$

1. It is C^1 concave.
2. $\frac{d}{d\sigma} P_\mu^d(\sigma) = \frac{1}{\mu} - x(\sigma)^T(\frac{1}{2}Hx(\sigma) - b)$.
3. $\frac{d^2}{d\sigma^2} P_\mu^d(\sigma) = -(Hx(\sigma)^T - b)^T G_\mu^{-1}(\sigma)(Hx(\sigma)^T - b)$
4. $P_\mu^d(\sigma)$ can not be unbounded on S_μ^+.

where $x(\sigma) = G_\mu^{-1}(\sigma)(f - \sigma b)$.

Perfect Duality

Theorem: If σ_μ is a global maximizer of $P_\mu^d(\sigma)$ over S_μ^+, then (P_μ^d) is perfectly dual to (P_μ) in the sense that the vector

$$x_\mu = G_\mu^{-1}(\sigma_\mu)(p - \sigma_\mu b)$$

is a global minimizer of (P_μ), and

$$\min_{x \in \chi_\mu} P_\mu(x) = P_\mu(x_\mu) = P_\mu^d(\sigma_\mu) = \max_{\sigma \in S_\mu^+} P_\mu^d(\sigma).$$

Theorem: If

$$\lim_{\sigma \to \infty} \frac{dP_\mu^d(\sigma)}{d\sigma} < 0$$

and

$$\lim_{\sigma(\in S_\mu^+) \to \partial S_\mu^+} \frac{dP_\mu^d(\sigma)}{d\sigma} \geq 0, \quad or\ S_\mu^+ = \{0\}$$

hold for every $\mu \in [\mu_0, \delta^{-1}]$, then

$$\min_{x \in chi} f(x) + \frac{g(x)}{h(x)} = \min_{\mu \in [\mu_0, \delta^{-1}]} P_\mu^d(\sigma_\mu).$$

where σ_μ is a global maximizer of $P_\mu^d(\sigma)$ over S_μ^+.

Numerical example for minimizing f(x)+g(x)/h(x)

$$f(x) = \frac{1}{2}x^T Qx - p^T x, \quad g(x) = \frac{1}{2}x^T Gx, \quad h(x) = \frac{1}{2}x^T Hx - b^T x,$$

with

$$Q = \begin{bmatrix} -1 & 6 \\ 6 & 5 \end{bmatrix}, \quad G = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}, \quad H = \begin{bmatrix} -7 & 3 \\ 3 & -2 \end{bmatrix},$$

$$p = \begin{bmatrix} -8 \\ 2 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ 3 \end{bmatrix}.$$

The constraint set $\chi = \{x \in R^n \mid \ h(x) \geq \delta = 0.01\}$ is an ellipse together with all its interior.

Existence of global maximizer σ_μ of $P_\mu^d(\sigma)$ for $\mu \in [0.04926, 100]$

- First compute the boundary of S_μ^+ as

$$\partial S_\mu^+ = \begin{cases} \{-6.9 - 3\mu + 0.1\sqrt{5581 + 3920\mu + 720\mu^2}\}, \\ \text{if } \mu \in [0.04926, 1.609); \\ \{0\}, \text{ if } \mu \in [1.609, 100]. \end{cases}$$

-

$$-20.29 \leq \lim_{\sigma \to \infty} \frac{dP_\mu^d(\sigma)}{d\sigma} \leq -13.6, \quad \text{for } \mu \in [0.04926, 100]$$

$$\lim_{\sigma \to \partial S_\mu^+} \frac{d}{d\sigma} P_\mu^d(\sigma) = \infty, \text{ when } \mu \in [0.04926, 1.609);$$

- The primal problem can be solved by $\min_{\mu \in [0.04926, 100]} P_\mu^d(\sigma_\mu)$.

Minimizing $P_\mu^d(\sigma_\mu)$

- We use the line search with the Armijo's rule and thus it is not necessary to solve σ_μ for each $\mu \in [0.04926, 100]$.
- It took only 6 times of line search to reach the global minimum of $P_\mu^d(\sigma_\mu)$ at $\mu = 0.69076$ with a value of -7.0766.
- A total of 22 function evaluations in 1.0615 cpu seconds to reach the optimal solution for this example.
- Compared to 82.84 seconds taken by applying the grid method (with a grid size of 0.01), our approach is much faster in speed and more precise in solution quality.

Thank you for your attention!