

Minimum complexity density estimation

Volkmar Liebscher, Evelyn Rost



EMAU Greifswald
Institut für Mathematik und Informatik

Workshop on Statistical Modelling of Complex Systems
Munich, 10. November 2005

The density estimation problem

Given data

$$X_1, \dots, X_n,$$

find

a **good** density f which **fits** their distribution.

Ansatz

e.c.d.f.

$$\hat{F}_n(x) = \sum_{i=1}^n 1_{[x_i, \infty)}(x)$$

Kolmogoroff: for X_i i.i.d. with c.d.f. F , asymptotically, with high probability

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| < \frac{1}{\sqrt{n}} T_{0.95}$$

Tubes of distribution functions

given: $a < b$ and piecewise constant càdlàg functions $L, U : [a, b] \mapsto \mathbb{R}$

form: a tube

$$T_{L,U}^{a,b} = \{F : [a, b] \mapsto \mathbb{R} : F \text{ left continuous } \nearrow, L \leq F \leq U\}$$

smooth

nice

wanted:

parsimonious

function from the tube $T_{L,U}^{a,b}$

easy to interpret

Smoothness-Scores

$$\Phi(f) = \int_a^b \varphi(\dot{f}(s)) ds, \quad \varphi \text{ strictly convex} \quad (1)$$

$$\Phi(f) = \int_a^b \sqrt{1 + \dot{f}(s)^2} ds \quad (2)$$

$$\Phi(f) = \text{No. of modes of } \dot{f} \quad (3)$$

$$\Phi(f) = \text{No. of plateaus of } \dot{f} \quad (4)$$

(2) is minimised by the *taut string* τ (Hartigan& Hartigan, 1985), (Davies&Kovac,2001)

The Taut String minimises also (1)

Theorem For all strictly convex $\varphi : \mathbb{R} \mapsto [0, \infty)$ there exists a unique location of minimum of Φ on $T_{L,U}^{a,b}$.

It is independent from φ and therefore coincides with τ .

Sketch of Proof (I)

Observation 1 Let $f \in T_{L,U}^{a,b}$, $\Phi(f) < \infty$ and $a \leq a' < b' \leq b$ so chosen that the function $\tilde{f} : [a, b] \mapsto \mathbb{R}$,

$$\begin{aligned}\tilde{f}(s) &= f(s) & s \notin [a', b'] \\ \tilde{f}(\lambda b' + (1 - \lambda)a') &= \lambda f(b') + (1 - \lambda)f(a') & \lambda \in [0, 1]\end{aligned}$$

lies also in $T_{L,U}^{a,b}$. Then $\Phi(\tilde{f}) \leq \Phi(f)$ and $=$ iff $\tilde{f} = f$.

Sketch of Proof (II)

Observation 2 There is a unique location of minimum f_0 which is continuous and piecewise linear.

Further notations:

knots of f : points of discontinuity \dot{f}

L, U -knots: $\dot{f}(s - 0) \geq \dot{f}(s + 0)$

Sketch of Proof (III)

Observation 3 The minimiser is uniquely determined by the following conditions:

1. f_1 is continuous and piecewise linear.
2. Is $s \in (a, b)$ a knot of f_1 then s is a point of discontinuity of L or U .
3. Is $s \in (a, b)$ a $\frac{L}{U}$ -knot of f_1 then

$$f_1(s) = \frac{L(s+0)}{U(s-0)}.$$

The taut string minimises also (3)

Theorem $\hat{\tau}$ has the minimal number of modes in $T_{L,U}^{a,b}$

Idea of Proof

1. indirect: choose a tube the *taut string* τ of which has minimal number of knots among all taut strings, which do not minimise Φ
2. Reduction and Modification until the largest knot
3. The following $L - U - L$ -Lemma allows to throw away this knot contradiction

$L - U - L$ -Lemma

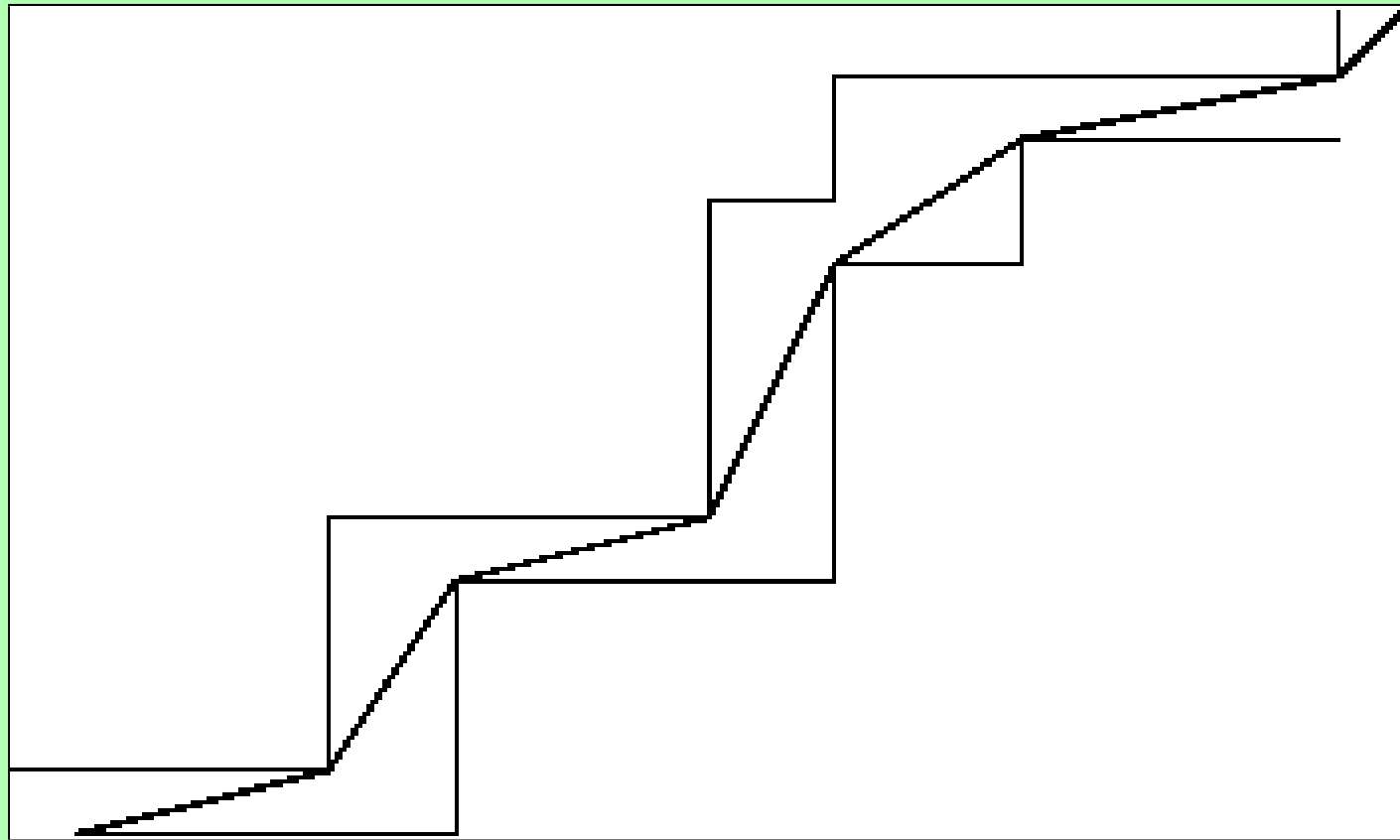
Lemma Are $s_1 < t < s_2$ L, U, L -knots of the *taut string* $\tau \in T_{L,U}^{a,b}$ such that t is the only knot in (s_1, s_2) .

Let $F \in T_{L,U}^{a,b}$ be concave in $[s'_1, s_2]$. Then $s'_1 > s_1$ and, if $s'_1 \leq t$, also $F(s'_1) \leq \tau(s'_1)$.

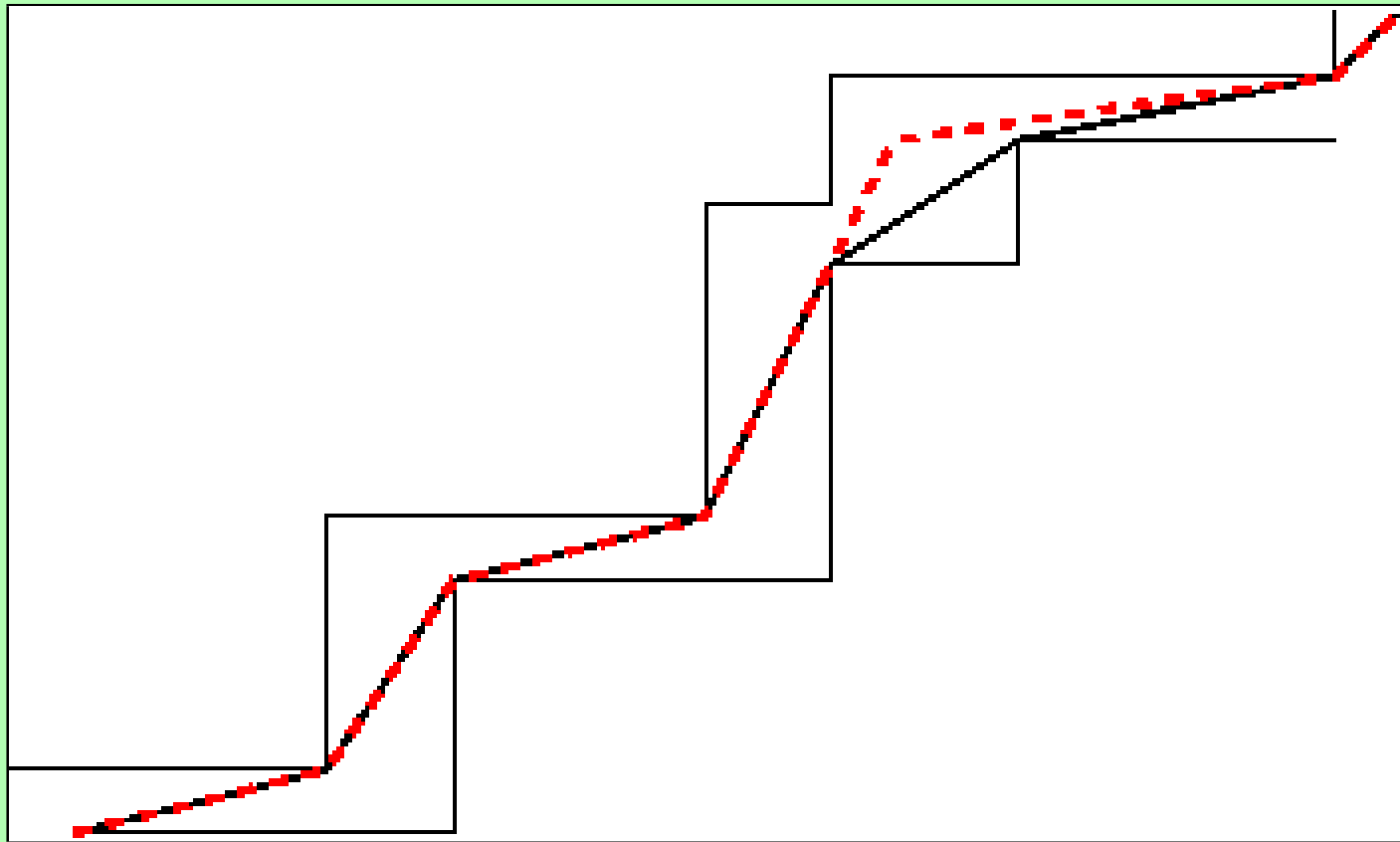
The taut string does not minimise (4)

Observation 1: There are L, U for which $\dot{\tau}$ has not the least number of plateaus in $T_{L,U}^{a,b}$.

Illustration



Illustration



... , but sometimes it does

Observation 2: Has τ Zig-Zag Form, then it has the least number of plateaus in $T_{L,U}^{a,b}$.

Some Obstacles

A function with least number of plateaus is not unique.

\implies look for 2nd criterion

Ansatz: minimal length of graph

Notation

A function with minimal length of graph among the functions with least number of plateaus is denoted F^* .

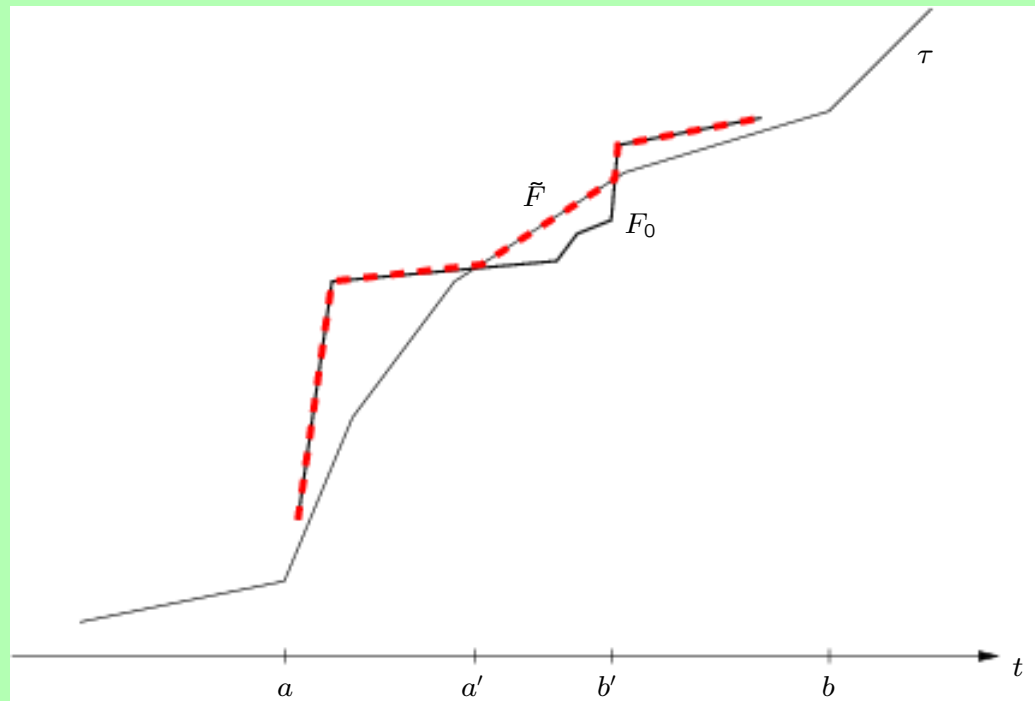
Lemma F^* exists

Lemma 1

Assume the taut string τ is concave in $[a', b']$. Then for any piecewise linear monotonously increasing function $F \in T_{L,U}^{a,b}$ there is a function $\tilde{F} \in T_{L,U}^{a,b}$ with

- $\#L\text{-knots}(\tilde{F}) \leq \#L\text{-knots}(F)$
- $\text{length}(\tilde{F}) \leq \text{length}(F)$
- $\tilde{F}(t) \geq \tau(t)$ for all $t \in [a, b]$

Illustration of Lemma 1

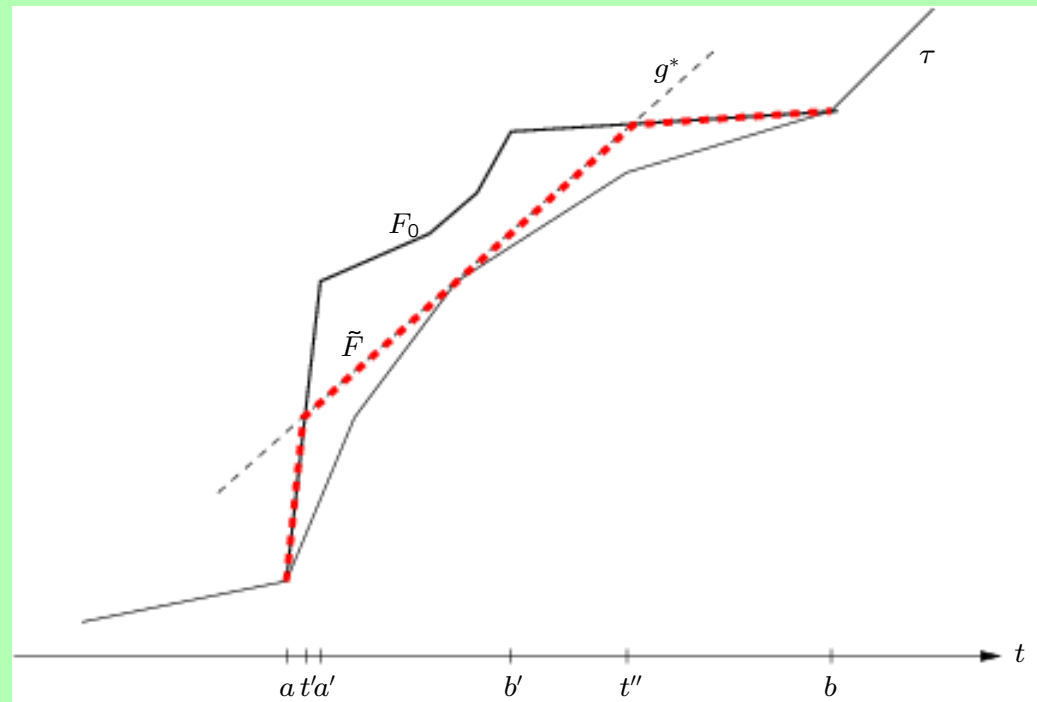


Lemma 2

Assume the taut string $\tau \in T_{L,U}^{a,b}$ is concave in $[a', b']$ and let $F \in T_{L,U}^{a,b}$ be an arbitrary piecewise linear function with $F(t) \geq \tau(t)$ for all $t \in [a', b']$. Then there is a function $\tilde{F} \in T_{L,U}^{a,b}$ with

- $\#L\text{-knots}(\tilde{F}) \leq \#L\text{-knots}(F)$
- $\#U\text{-knots}(\tilde{F}) = 0$ in (a', b')
- $\text{length}(\tilde{F}) \leq \text{length}(F)$

Illustration of Lemma 2



The crucial reduction

Theorem Assume the taut string $\tau \in T_{L,U}^{a,b}$ is concave in $[a', b']$. Then there exist for any function F with minimal number of plateaus a function $\tilde{F} \in T_{L,U}^{a,b}$ with minimal number of plateaus which fulfils additionally

i) \tilde{F} is concave in $[a', b']$

ii) $\tilde{F} \geq \tau$ in $[a', b']$

iii) $\text{length}(\tilde{F}) \leq \text{length}(F)$

iv) $\tilde{F} = \tau$ in $[a', a'']$ and $[b'', b']$

Consequences

1. the optimal function F^* fulfils $i) - iv)$
2. F^* coincides with τ in regions where τ is zig-zag
3. The number of knots of τ can be lessened only in regions where τ several consecutive knots of τ have the same type

The Algorithm for Concave Regions

1. Set $i = 1$, $k = 0$, $\hat{t} = t_1$ and $g_1(t) = \frac{L(t_2) - U(t_1)}{t_2 - t_1}(t - t_1) + U(t_1)$;

2.

Do [$\tilde{t} := \min\{t > \hat{t} : g_i(t) = U(t) \text{ and } U(t) \text{ is no discontinuity of } U\}$;

$g(t) := gcm_U^{[\tilde{t}, t_{n+2}]}(t)$;

acquire \hat{t} such that $(\hat{t}, U(\hat{t}))$ is the rightmost knot (including initial point) $\neq (t_{n+2}, U(t_{n+2}))$ of g ;

$i := i + 1, k = 0$;

Do [$\tilde{g}(t) = \frac{\tau(t_{n+2-k}) - U(\hat{t})}{t_{n+2-k} - \hat{t}}(t - \hat{t}) + U(\hat{t})$;

$k := k + 1$;

] ;

While $\exists t \in [\hat{t}, t_{n+2}] : \tilde{g}(t) < \tau(t)$;

$g_i(t) = \tilde{g}(t)$;

] ;

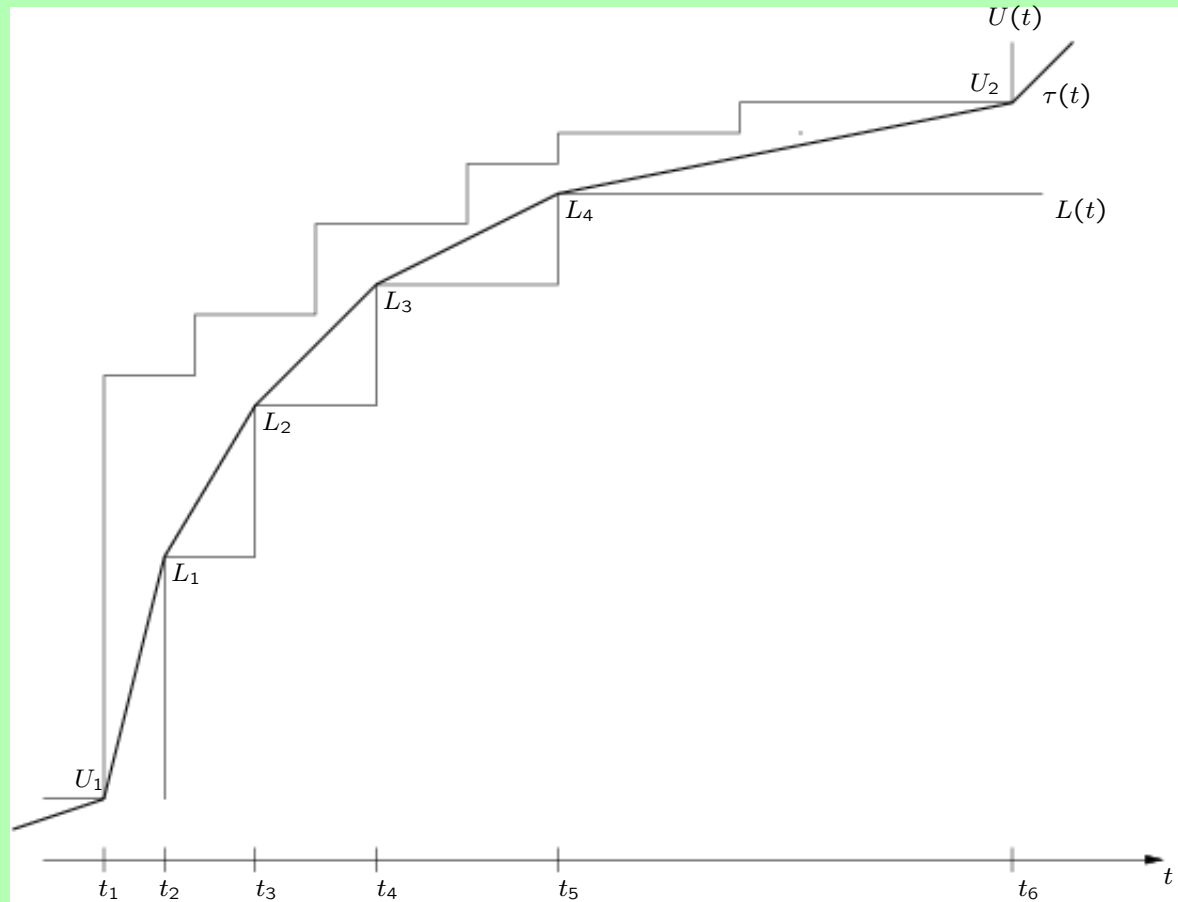
While $\exists t \in [\tilde{t}, t_{n+2}] : g(t) < \tau(t)$;

The Algorithm for Concave Regions

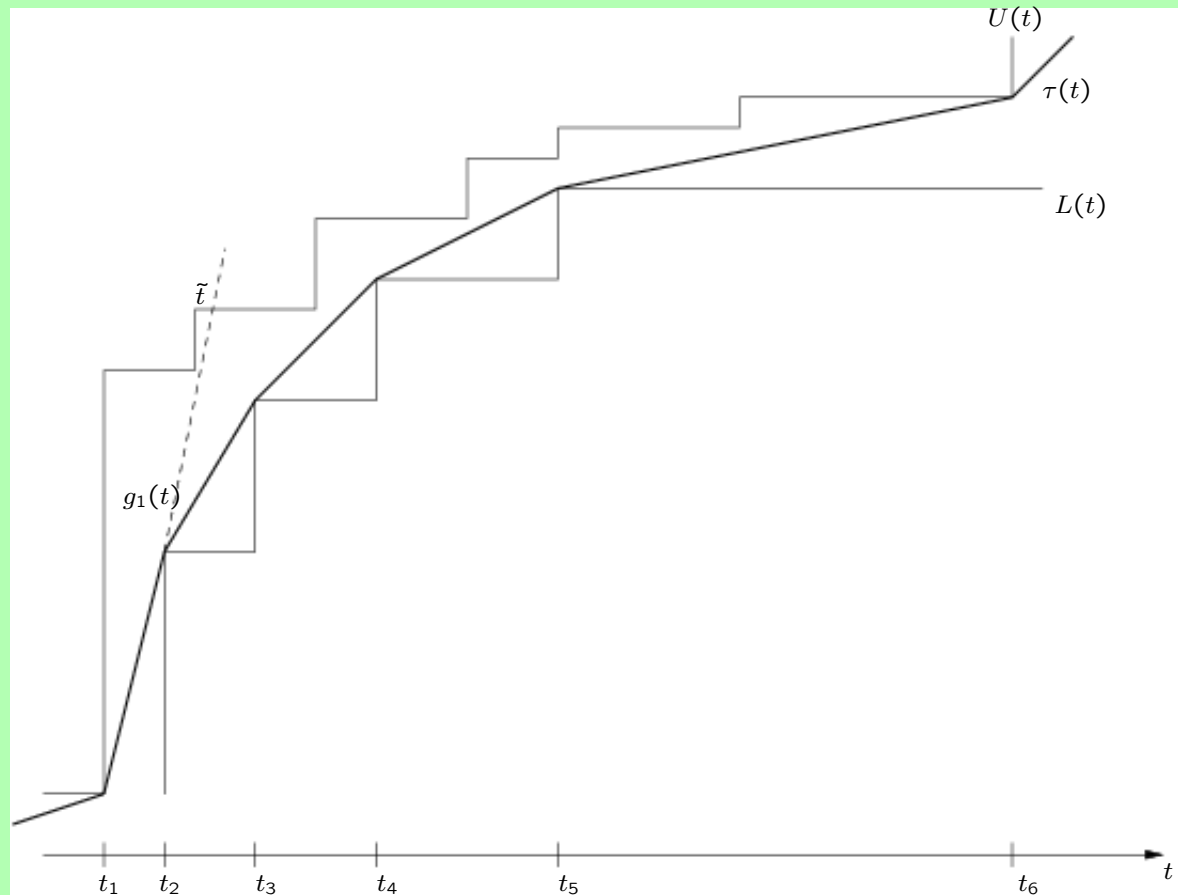
3. For $j = 1$ to $i - 1$ Do [$t'_j = \{t : g_j(t) = g_{j+1}(t)\}$] ;

$$F(t) = \begin{cases} g_1(t) & t_1 \leq t \leq t'_1 \\ g_2(t) & t'_1 < t \leq t'_2 \\ \vdots & \\ g_i(t) & t'_{i-1} < t \leq t_{n+2} \end{cases}$$

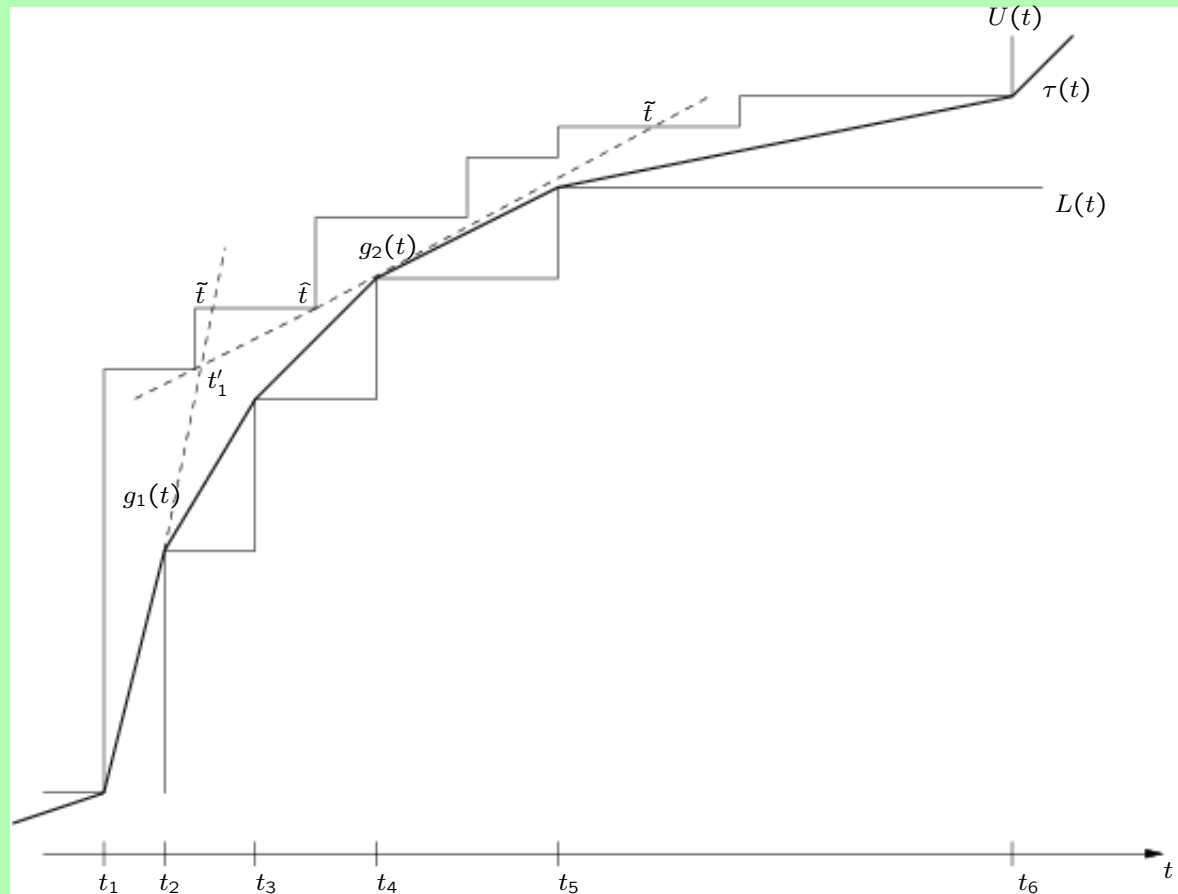
An Example



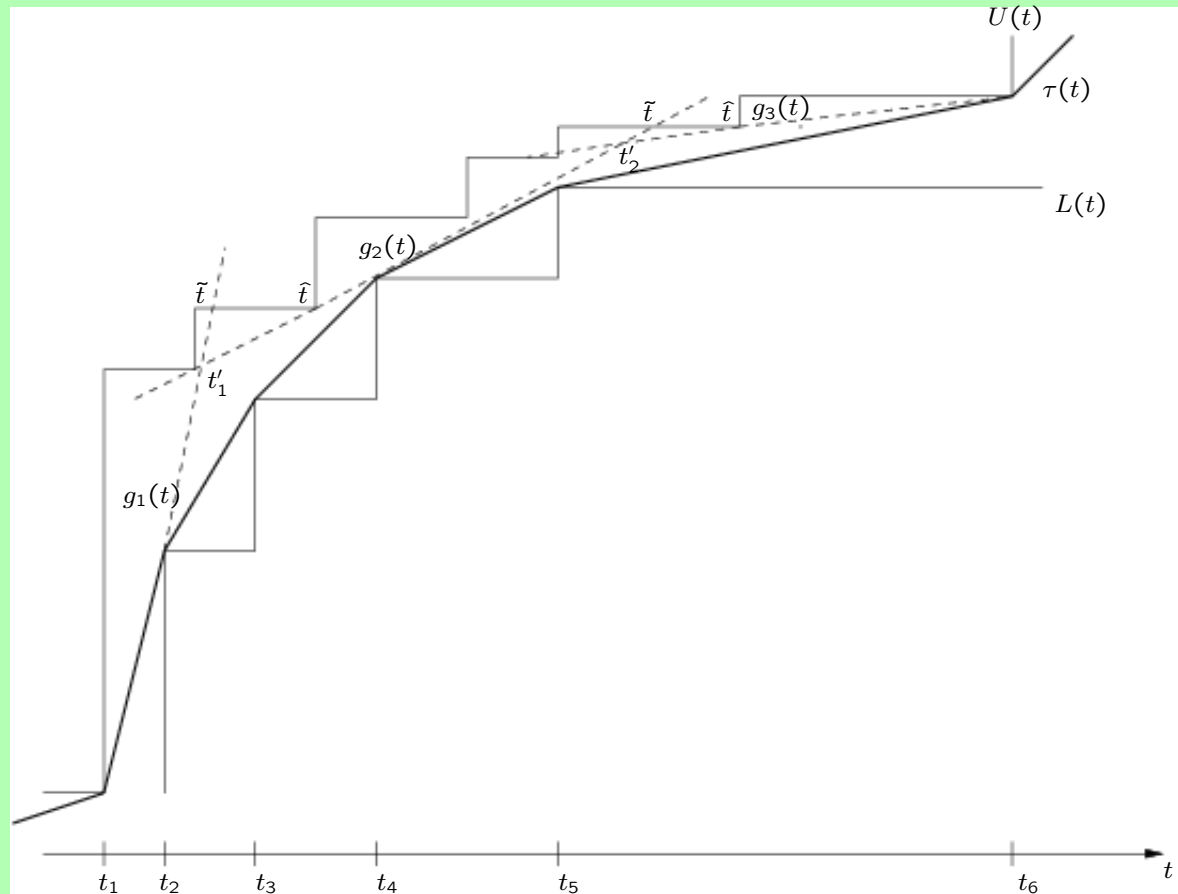
An Example



An Example



An Example



An Example

