The Double Skorohod Map and Real-Time Queues

Steven E. Shreve
Department of Mathematical Sciences
Carnegie Mellon University
www.math.cmu.edu/users/shreve

Joint work with
Łukasz Kruk
John Lehoczky
Kavita Ramanan

Conference on Stochastic Processing Networks
in Honor of J. Michael Harrison
August 29–30, 2009

The two most influential unknown papers of the Twentieth Century

The two most influential unknown papers of the Twentieth Century

- J. M. Harrison & S. R. Pliska (1981) Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Applications* **11**, 215–260.

- J. M. Harrison & S. R. Pliska (1983) A stochastic calculus model of continuous trading: complete markets, *Stochastic Processes and Applications* **15**, 313–316.

Fundamental Theorems of Asset Pricing

Fundamental Theorems of Asset Pricing

Definition

A martingale measure is a probability measure, equivalent to the actual measure, under which all discounted (at the possibly random interest rate) asset price processes are martingales.

Fundamental Theorems of Asset Pricing

Definition

A martingale measure is a probability measure, equivalent to the actual measure, under which all discounted (at the possibly random interest rate) asset price processes are martingales.

Theorem (First Fundamental Theorem)

There exists a martingale measure if and only if a model admits no arbitrage.

Fundamental Theorems of Asset Pricing

Definition

A martingale measure is a probability measure, equivalent to the actual measure, under which all discounted (at the possibly random interest rate) asset price processes are martingales.

Theorem (First Fundamental Theorem)

There exists a martingale measure if and only if a model admits no arbitrage.

Theorem (Second Fundamental Theorem)

Consider a model that admits no arbitrage. The martingale measure is unique if and only if every derivative security can be replicated by trading in the primary assets.

Context

Context

- Discrete-time trading and continuous-time trading.

Context

- Discrete-time trading and continuous-time trading.
- Admissible trading strategies must be self-financing and lead to almost surely nonnegative portfolio values at the final time. Must also satisfy some technical conditions.

Context

- Discrete-time trading and continuous-time trading.
- Admissible trading strategies must be self-financing and lead to almost surely nonnegative portfolio values at the final time. Must also satisfy some technical conditions.
- Semi-martingale asset price processes.

Context

- Discrete-time trading and continuous-time trading.
- Admissible trading strategies must be self-financing and lead to almost surely nonnegative portfolio values at the final time. Must also satisfy some technical conditions.
- Semi-martingale asset price processes.

"We are working dangerously close to the boundaries of our knowledge...." — *J. M. Harrison and S. Pliska*

Consequences

Consequences

- Derivative security pricing no longer restricted to geometric Brownian motion.

Consequences

- Derivative security pricing no longer restricted to geometric Brownian motion.
- No longer tied to Markov assumption.

Consequences

- Derivative security pricing no longer restricted to geometric Brownian motion.
- No longer tied to Markov assumption.
- No longer must asset price processes be continuous.

Further consequences

Further consequences

- Heath-Jarrow-Morton model for interest rate derivatives.

Further consequences

- Heath-Jarrow-Morton model for interest rate derivatives.
- Optimal investment and consumption in a general stochastic process setting.

Further consequences

- Heath-Jarrow-Morton model for interest rate derivatives.
- Optimal investment and consumption in a general stochastic process setting.
- Equilibrium analysis in a general setting.

Further consequences

- Heath-Jarrow-Morton model for interest rate derivatives.
- Optimal investment and consumption in a general stochastic process setting.
- Equilibrium analysis in a general setting.
- Theory of market incompleteness, the case of multiple martingale measures.

Further consequences

- Heath-Jarrow-Morton model for interest rate derivatives.
- Optimal investment and consumption in a general stochastic process setting.
- Equilibrium analysis in a general setting.
- Theory of market incompleteness, the case of multiple martingale measures.

"DNA is not all good. Criminals use it to get out of jail."

—*Stephen Colbert*

Further consequences

- Heath-Jarrow-Morton model for interest rate derivatives.
- Optimal investment and consumption in a general stochastic process setting.
- Equilibrium analysis in a general setting.
- Theory of market incompleteness, the case of multiple martingale measures.

"DNA is not all good. Criminals use it to get out of jail."

—*Stephen Colbert*

- Risk-neutral pricing, a consequence of the existence of the martingale measure, can be applied blindly without thinking whether the measure is unique.
  See "Did faulty mathematical models cause the financial fiasco?," *Analytics Magazine*, Spring 2009, available at www.math.cmu.edu/users/shreve.

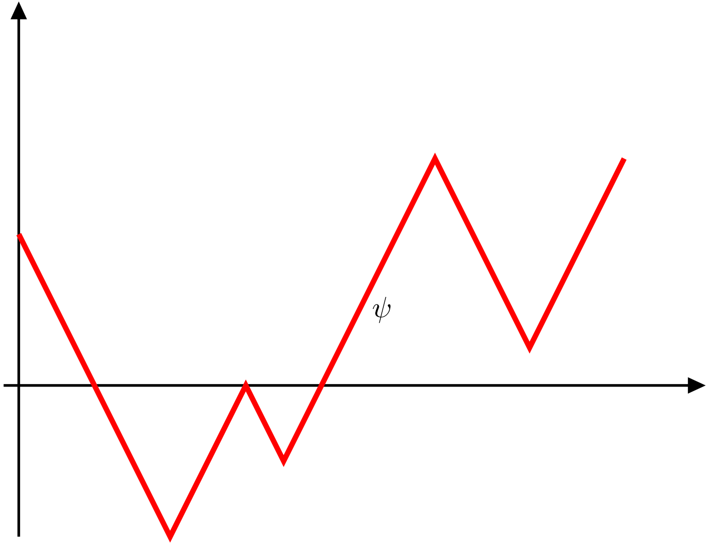
Outline of the Rest of the Talk

Skorohod Map

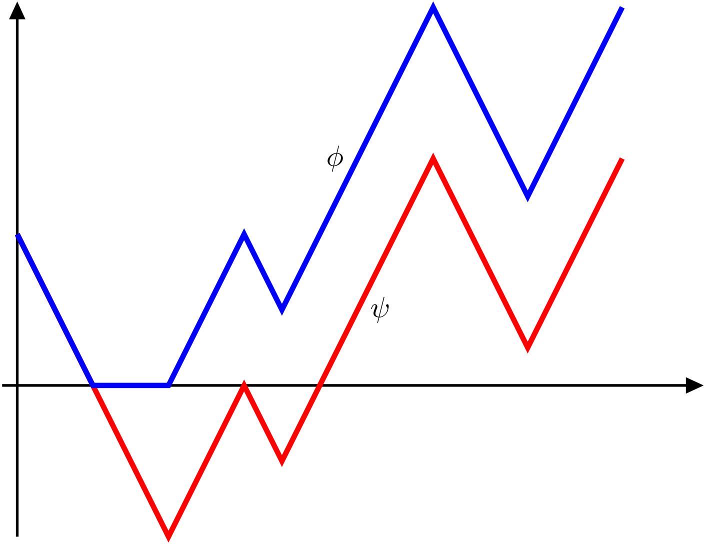
Real-Time Queues

Real-Time Queues with Reneging

1. Skorohod Map

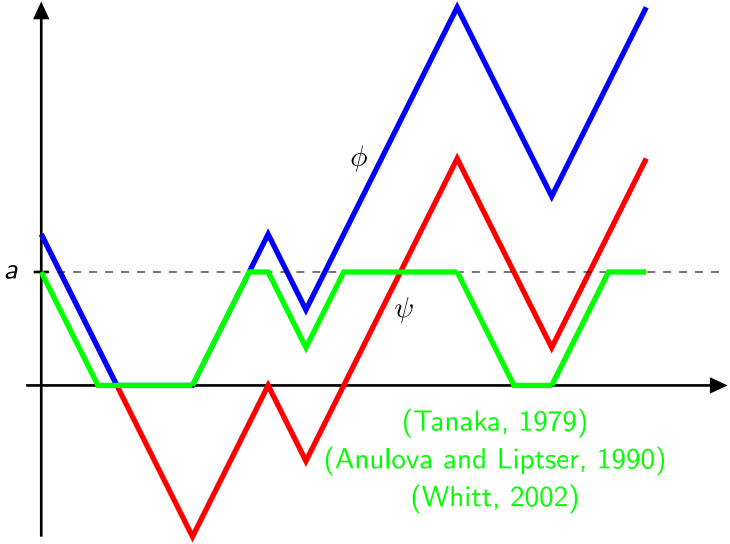


1. Skorohod Map



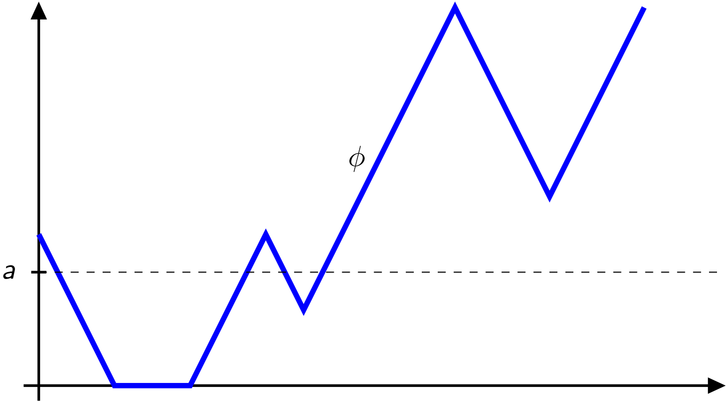
$\phi(t) \triangleq \psi(t) - \inf_{0 \leq s \leq t} \left[\psi(s) \wedge 0\right]$. (Skorokhod, 1961)

1. Skorohod Map



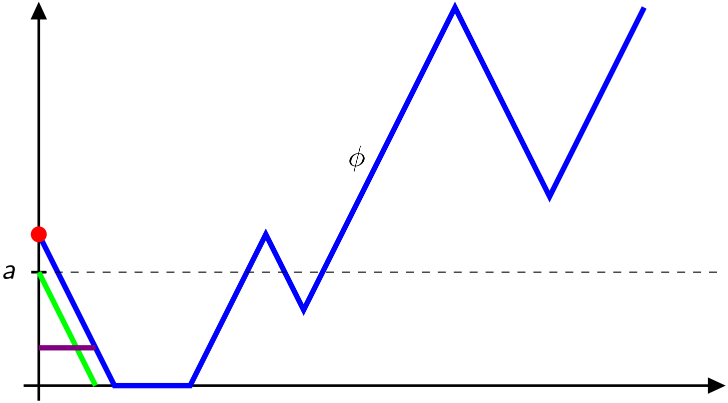
$\phi(t) \triangleq \psi(t) - \inf_{0\leq s\leq t} \left[\psi(s) \wedge 0\right].$ (Skorokhod, 1961)

Formula for Double Skorohod Map



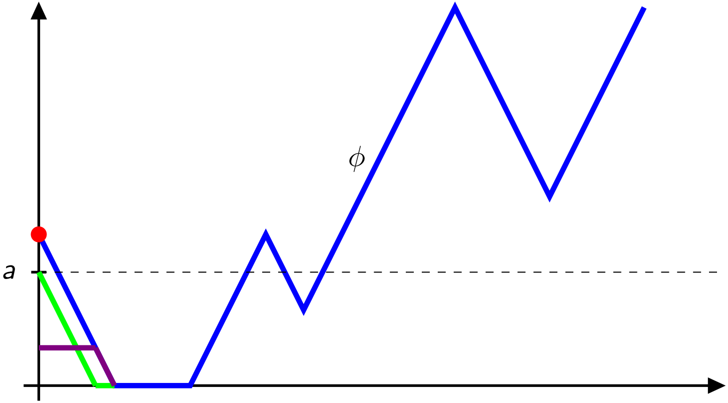
$$\lambda(\phi)(t) \triangleq \phi(t) - \sup_{s\in[0,t]} \left[(\phi(s) - a)^+ \wedge \inf_{u\in[s,t]} \phi(u) \right]$$

Formula for Double Skorohod Map



$$\lambda(\phi)(t) \triangleq \phi(t) - \sup_{s\in[0,t]}\left[\left(\phi(s)-a\right)^{+} \wedge \inf_{u\in[s,t]}\phi(u)\right]$$

Formula for Double Skorohod Map



$$\lambda(\phi)(t) \triangleq \phi(t) - \sup_{s\in[0,t]}\left[\left(\phi(s)-a\right)^{+} \wedge \inf_{u\in[s,t]} \phi(u)\right]$$

Related formulas

Toomey (1998).

Let ψ be piecewise constant. The double reflection in $[0,a]$ of ψ is

$$\inf_{s\in(0,t]} \sup_{u\in(s,t]} \Big[\big(a+\psi(t)-\psi(s)\big) \vee \big(\psi(t)-\psi(u)\big)\Big] \vee \sup_{u\in(0,t]} \Big[\big(\psi(t) \vee \big(\psi(t)-\psi(u)\big)\Big].$$

Related formulas

Cooper, Schmidt and Serfozo (2001).
H is a signed measure on $[0, \infty)$ and

$$X(t) = \sup_{s \in [0,t]} \inf_{u \in [s,t]} \left[x I_{\{s=u=o\}} + H(u,t] - a I_{\{s=u>0\}} \right],$$

Then X is the double reflection in $[-a, 0]$ of the bounded-variation function $t \mapsto (x + H(0,t])$.

Related formulas

Ganesh, O'Connell and Wischik (2004).
Let ψ be a bounded-variation function. The double reflection in $[0, a]$ of ψ is

$$\left(\psi(t) \vee \inf_{s\in[0,t]} \left[N(s,t) \wedge \left(M(s,t)+a\right)\right]\right) \wedge \inf_{s\in[0,t]} \left[N(s,t) \vee \left(M(s,t)+a\right)\right],$$

where

$$\begin{aligned} M(s,t) &= \psi(t) - \sup_{u\in[s,t]} \psi(u), \\ N(s,t) &= \psi(t) - \inf_{u\in[s,t]} \psi(u). \end{aligned}$$

2. Real-Time Queues

Single station, renewal arrival process.

Heavy traffic assumption: For some $\gamma \neq 0$, $\rho^{(n)} = 1 - \frac{\gamma}{\sqrt{n}}$.

Workload process: $W^{(n)}(t)$

Scaled workload process: $\widehat{W}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} W^{(n)}(nt)$

Theorem (Kingman (1961), Iglehart/Whitt (1970))

$$\widehat{W}^{(n)} \Rightarrow W^*,$$

where W^* is a Brownian motion with drift $-\gamma$, reflected at the origin so as to always be nonnegative.

Lead Times

- $L_1^{(n)}, L_2^{(n)}, \ldots$ – IID positive random variables. The lead times.
- $G(y)$ – Cumulative distribution function.

$$\mathbb{P}\left\{\frac{L_j^{(n)}}{\sqrt{n}} \leq y\right\} = G(y)$$

Customers are assigned lead times upon arrival, and lead times decrease at rate 1 per unit time thereafter. Delay grows like $\sqrt{n}$, so we must let lead times also grow like $\sqrt{n}$.

Lead Times

- $L_1^{(n)}, L_2^{(n)}, \ldots$ – IID positive random variables. The lead times.
- $G(y)$ – Cumulative distribution function.

$$\mathbb{P}\left\{ \frac{L_j^{(n)}}{\sqrt{n}} \leq y \right\} = G(y)$$

Customers are assigned lead times upon arrival, and lead times decrease at rate 1 per unit time thereafter. Delay grows like $\sqrt{n}$, so we must let lead times also grow like $\sqrt{n}$.

Earliest Deadline First (EDF) – Always serve customer with smallest lead time. Ties do not matter. Use pre-emption.

Lead Times

- $L_1^{(n)}, L_2^{(n)}, \ldots$ – IID positive random variables. The lead times.
- $G(y)$ – Cumulative distribution function.

$$\mathbb{P}\left\{ \frac{L_j^{(n)}}{\sqrt{n}} \le y \right\} = G(y)$$

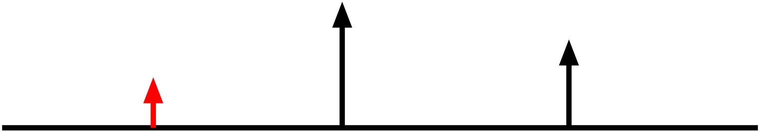
Customers are assigned lead times upon arrival, and lead times decrease at rate 1 per unit time thereafter. Delay grows like $\sqrt{n}$, so we must let lead times also grow like $\sqrt{n}$.

Earliest Deadline First (EDF) – Always serve customer with smallest lead time. Ties do not matter. Use pre-emption.

Problem: Determine the heavy traffic limit of the distribution of lead times of customers in queue.

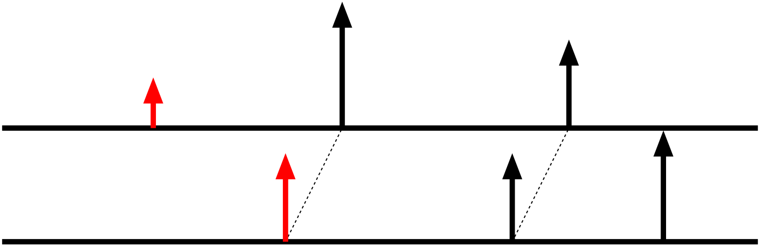
Dynamics of lead times under EDF

$F^{(n)}(t)$ – Largest lead time of any customer who has ever been in service by time t, the frontier.



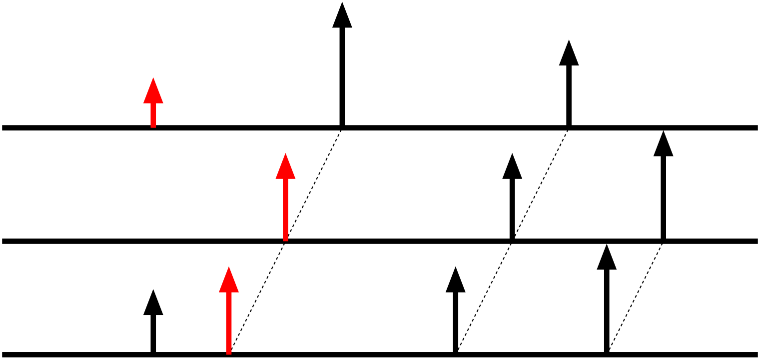
Dynamics of lead times under EDF

$F^{(n)}(t)$ – Largest lead time of any customer who has ever been in service by time t, the frontier.



Dynamics of lead times under EDF

$F^{(n)}(t)$ – Largest lead time of any customer who has ever been in service by time t, the frontier.



Workload and arrived-work measures

Let B be a Borel subset of $\mathbb{R}$. Define

$$\mathcal{W}^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work associated with customers in} \\ \text{queue at time } t \text{ with lead times in } B. \end{array} \right\}$$

$$\mathcal{V}^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work associated with customers arrived} \\ \text{by time } t \text{ with lead times in } B, \text{ whether} \\ \text{or not customer is still present at time } t. \end{array} \right\}$$

Workload and arrived-work measures

Let B be a Borel subset of $\mathbb{R}$. Define

$$\mathcal{W}^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work associated with customers in} \\ \text{queue at time } t \text{ with lead times in } B. \end{array} \right\}$$

$$\mathcal{V}^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work associated with customers arrived} \\ \text{by time } t \text{ with lead times in } B, \text{ whether} \\ \text{or not customer is still present at time } t. \end{array} \right\}$$

Scaled processes

$$\widehat{\mathcal{W}}^{(n)}(t)(B) \triangleq \frac{1}{\sqrt{n}} \mathcal{W}^{(n)}(nt)(\sqrt{n}\,B),$$

$$\widehat{\mathcal{V}}^{(n)}(t)(B) \triangleq \frac{1}{\sqrt{n}} \mathcal{V}^{(n)}(nt)(\sqrt{n}\,B),$$

$$\widehat{F}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} F^{(n)}(nt).$$

Limiting lead-time distribution

Lemma (Crushing)

$$\widehat{\mathcal{W}}^{(n)}\big(-\infty, \widehat{F}^{(n)}\big] \Rightarrow 0.$$

Limiting lead-time distribution

Lemma (Crushing)

$$\widehat{\mathcal{W}}^{(n)}\big(-\infty, \widehat{F}^{(n)}\big] \Rightarrow 0.$$

Corollary

For every $y \in \mathbb{R}$,

$$\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - \widehat{\mathcal{V}}^{(n)}(t)\big(y \vee \widehat{F}^{(n)}(t), \infty\big) \Rightarrow 0.$$

Limiting lead-time distribution

Lemma (Crushing)

$$\widehat{\mathcal{W}}^{(n)}\big(-\infty, \widehat{F}^{(n)}\big] \Rightarrow 0.$$

Corollary

For every $y \in \mathbb{R}$,

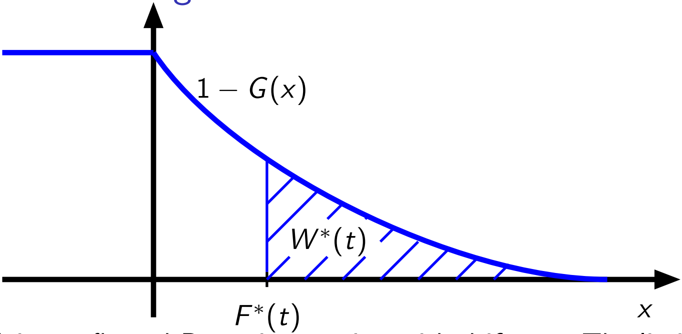
$$\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - \widehat{\mathcal{V}}^{(n)}(t)\big(y \vee \widehat{F}^{(n)}(t),\infty\big) \Rightarrow 0.$$

Theorem

For all $y \in \mathbb{R}$,

$$\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) \Rightarrow H(y) \triangleq \int_y^\infty \big(1 - G(x)\big)\, dx.$$

Evolution of limiting workload measure



$W^*(t)$ is a reflected Brownian motion with drift $-\gamma$. The limiting scaled frontier is

$$F^*(t) = H^{-1}\big(W^*(t)\big).$$

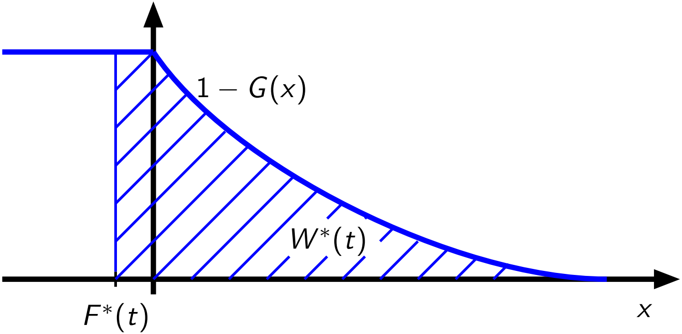
The limit of the measure-valued workload process $\widehat{\mathcal{W}}^{(n)}(t)$ has density $\big(1 - G(x)\big) I_{\{x \geq F^*(t)\}}$. We call this limiting measure-valued process

$$\mathcal{W}^*(t).$$

(Doytchinov, Lehoczky, Shreve (2000))

3. Real-Time Queues with Reneging

Customers are late in the limiting system when $F^*(t)$ is negative.



$$F^*(t) < 0 \Longleftrightarrow W^*(t) > H(0) = \int_0^\infty (1 - G(x))\, dx.$$

Theorem

If customers renege when their lead times reach zero, then the limiting scaled workload process is a

doubly reflected Brownian motion on $[0, H(0)]$ *with drift* $-\gamma$*.*

The limiting scaled workload measure is as before.

Ingredients of the proof

- $\mathcal{M}$ – The set of finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- $D_{\mathcal{M}}[0, \infty)$ – The set of cádlág functions taking values in $\mathcal{M}$. A sample path of the workload process, either scaled or unscaled, is an element of $D_{\mathcal{M}}[0, \infty)$.

Ingredients of the proof

- $\mathcal{M}$ – The set of finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- $D_{\mathcal{M}}[0,\infty)$ – The set of cádlág functions taking values in $\mathcal{M}$. A sample path of the workload process, either scaled or unscaled, is an element of $D_{\mathcal{M}}[0,\infty)$.
- $\Lambda\colon D_{\mathcal{M}}[0,\infty) \to D_{\mathcal{M}}[0,\infty)$

$$\Lambda(\mu)(t)(-\infty,y] \triangleq \left[\mu(t)(-\infty,y] - \sup_{0\le s\le t}\left(\mu(s)(-\infty,0] \wedge \inf_{s\le u\le t}\mu(u)(\mathbb{R})\right)\right]^+ .$$

Ingredients of the proof

- $\mathcal{M}$ – The set of finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- $D_{\mathcal{M}}[0,\infty)$ – The set of cádlág functions taking values in $\mathcal{M}$. A sample path of the workload process, either scaled or unscaled, is an element of $D_{\mathcal{M}}[0,\infty)$.
- $\Lambda \colon D_{\mathcal{M}}[0,\infty) \to D_{\mathcal{M}}[0,\infty)$

$$\begin{aligned}&\Lambda(\mu)(t)(-\infty,y]\\ &\quad\triangleq \left[\mu(t)(-\infty,y] - \sup_{0\le s\le t}\left(\mu(s)(-\infty,0] \wedge \inf_{s\le u\le t}\mu(u)(\mathbb{R})\right)\right]^{+}.\end{aligned}$$

Set

$$\mathcal{U}^{(n)}(t) \triangleq \Lambda(\mathcal{W}^{(n)})(t).$$

Define

$$\begin{aligned}U^{(n)}(t) &\triangleq \mathcal{U}^{(n)}(\mathbb{R})(t)\\ &= W^{(n)}(t) - \sup_{0\le s\le t}\left[\mathcal{W}^{(n)}(s)(-\infty,0] \wedge \inf_{s\le u\le t} W^{(n)}(u)\right].\end{aligned}$$

The doubly-reflected Brownian motion U^*

Scale and pass to the limit:

$$U^*(t) \quad = \quad W^*(t) - \sup_{0 \le s \le t} \left[\mathcal{W}^*(s)(-\infty, 0] \wedge \inf_{s \le u \le t} W^*(u) \right].$$

The doubly-reflected Brownian motion U^*

Scale and pass to the limit:

$$U^*(t) \quad = \quad W^*(t) - \sup_{0 \le s \le t} \left[\mathcal{W}^*(s)(-\infty, 0] \wedge \inf_{s \le u \le t} W^*(u) \right].$$

But in the limit, we have

$$\mathcal{W}^*(s)(-\infty, 0] = \big(W^*(s) - H(0)\big)^+.$$

Therefore,

$$U^*(t) \quad = \quad W^*(t) - \sup_{0 \le s \le t} \left[\big(W^*(s) - H(0)\big)^+ \wedge \inf_{s \le u \le t} W^*(u) \right].$$

The doubly-reflected Brownian motion U^*

Scale and pass to the limit:

$$U^*(t) \quad = \quad W^*(t) - \sup_{0 \le s \le t} \left[\mathcal{W}^*(s)(-\infty, 0] \wedge \inf_{s \le u \le t} W^*(u) \right].$$

But in the limit, we have

$$\mathcal{W}^*(s)(-\infty, 0] = \big(W^*(s) - H(0) \big)^+.$$

Therefore,

$$U^*(t) \quad = \quad W^*(t) - \sup_{0 \le s \le t} \left[\big(W^*(s) - H(0) \big)^+ \wedge \inf_{s \le u \le t} W^*(u) \right].$$

Recall the double-reflection map for a scalar-valued process

$$\lambda(\phi)(t) \triangleq \phi(t) - \sup_{s \in [0,t]} \left[\big(\phi(s) - a \big)^+ \wedge \inf_{u \in [s,t]} \phi(u) \right].$$

The measures $\mathcal{W}_R^{(n)}$ and $\mathcal{U}^{(n)} = \Lambda(\mathcal{W}^{(n)})$

Let $D^{(n)}$ be the work that arrives to the reneging system ahead of the frontier and later reneges.

Lemma (Comparison)

For the unscaled processes, we have

$$0 \le U^{(n)}(t) - W_R^{(n)}(t) \le D^{(n)}(t).$$

For the scaled processes

$$\widehat{U}^{(n)}(t) \triangleq \frac{U^{(n)}(nt)}{\sqrt{n}}, \quad \widehat{W}_R^{(n)}(t) \triangleq \frac{W_R^{(n)}(nt)}{\sqrt{n}}, \quad \widehat{D}^{(n)}(t) \triangleq \frac{D^{(n)}(nt)}{\sqrt{n}},$$

we have the comparison

$$0 \le \widehat{U}^{(n)}(t) - \widehat{W}_R^{(n)}(t) \le \widehat{D}^{(n)}(t).$$

The measures $\mathcal{W}_R^{(n)}$ and $\mathcal{U}^{(n)} = \Lambda(\mathcal{W}^{(n)})$

Lemma (Crushing)
$\widehat{D}^{(n)} \Rightarrow 0.$

Theorem (Limit of reneging system)
$\widehat{\mathcal{W}}_R^{(n)} - \widehat{\mathcal{U}}^{(n)} \Rightarrow 0$, *or equivalently,* $\widehat{\mathcal{W}}_R^{(n)} \Rightarrow \Lambda(\widehat{\mathcal{W}}^*)$.