RISK AVERSION, ROAD CHOICE AND THE ONE-ARMED BANDIT PROBLEM

Jean-Philippe Chancelier[1], Michel De Lara[2], André de Palma[3]

December 11, 2008

[1]CERMICS, École des ponts, ParisTech, France
[2]
[3]Université de Cergy-Pontoise, École des ponts Paris Tech, CORE and Institut universitaire de France

Credits

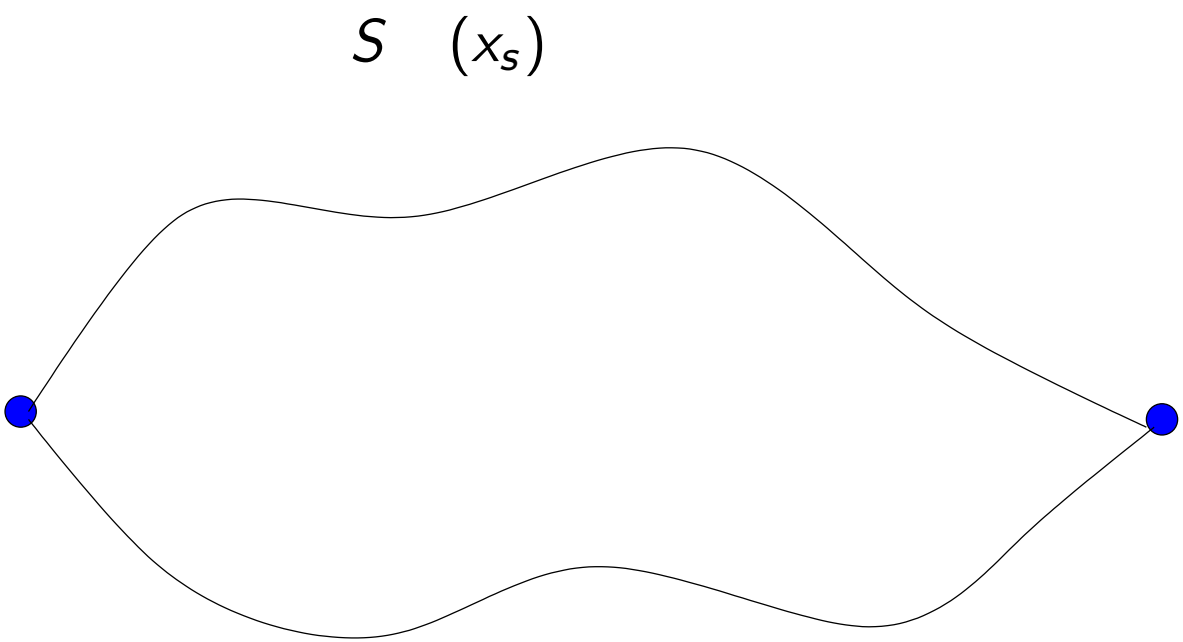
J.-P. Chancelier, M. De Lara, A. de Palma.
Risk aversion, road choice and the one-armed bandit problem.
In *Transportation Science*, 2006 Volume 41, Number 1, February 2007, Pages 1-14.

J.-P. Chancelier, M. De Lara, and A. de Palma.
Risk aversion in expected intertemporal discounted utilities bandit problems.
In *Theory and Decision*, 2008.

Outline of the presentation

Outline of the presentation

A sequential choice between two roads



A sequential choice between two roads

Consider one origin, one destination and two roads in parallel.

At every period $[t, t+1[$ starting at t (for instance a day), denoted by period t in the sequel, the driver selects one of the two roads (safe and random) characterized as follows:

1. The safe road S has constant known travel time x_S.
2. The random road R has a random travel time X_{t+1} realized at the end of period t. X_{t+1} takes $n \geq 2$ values in $\{x_1, \ldots, x_n\}$.

The random road

We assume that

- the values $x_1, \ldots, x_n$ of the random travel time are known to the driver
- and that (without loss of generality):

$$x_1 < \cdots < x_n \quad \text{and} \quad x_1 < x_S < x_n \,. \tag{1}$$

We suppose that

- x_i occurs with probability $\overline{p}_i \in]0,1[$, $i = 1, \ldots, n$, with $\overline{p}_1 + \cdots + \overline{p}_n = 1$,
- but that $\overline{p}_1, \ldots, \overline{p}_n$ are not necessarily known to the driver.

Sample space

$$\Omega = \{x_1, \ldots, x_n\}^{\mathbb{N}^*}$$

The coordinates $X_t(\omega) = \omega(t)$, $t \in \mathbb{N}^*$, form the sequence of random travel times on the random road. Let

$$S_{n-1} := \{(p_1, \ldots, p_n) \in \mathbb{R}_+^n\,, \quad p_1 + \cdots + p_n = 1\}$$

$$\text{and} \quad \overline{p} := (\overline{p}_1, \ldots, \overline{p}_n) \in S_{n-1}\,.$$

We define the probability $\mathbb{P}^{\overline{p}}$ on Ω by the marginals

$$\mathbb{P}^{\overline{p}}(X_1 = z_1, \ldots, X_t = z_t) = \prod_{s=1}^{t} [\overline{p}_1 \mathbf{1}_{\{z_s = x_1\}} + \cdots + \overline{p}_n \mathbf{1}_{\{z_s = x_n\}}]\,.$$

The expectation under the probability $\mathbb{P}^{\overline{p}}$ is denoted by $\mathbb{E}^{\overline{p}}$.

Decisions: the road chosen at the beginning of period t

The *decision* $v_t \in \{S, R\}$ is
the road chosen at the beginning of period t.

The observation at the end of period t depends on the information regimes envisaged.

Information could be acquired either by direct observation or *via* some driver information system which may forecast future travel conditions.

Experienced travel time

We shall denote by

$$Y_{t+1} = \Phi(v_t, X_{t+1})\,. \tag{2}$$

the travel time experienced by the driver at the end of period t:

- $Y_{t+1} = x_S$ if he selects the safe road;
- $Y_{t+1} = X_{t+1}$ if he selects the random road.

Thus, the experienced travel time Y_{t+1} depends upon both the decision v_t and X_{t+1}.

Preference model

The preferences of a driver are characterized by

- a utility function V;
- a discount rate $\rho \in [0, 1[$.

We shall call it driver $[V, \rho]$.

The utility function V is strictly decreasing (this is because, in the transportation context, V is decreasing in its argument, the travel time) concave, so that by Eq. (1)

$$V(x_1) > \cdots > V(x_n) \quad \text{and} \quad V(x_1) > V(x_S) > V(x_n)\,.$$

Intertemporal discounted reward maximization

Let the reward $G(v,x)$ be defined by the instantaneous utility resulting from road choice and experienced travel time:

$$G(v,x) := V(\Phi(v,x)), \quad \forall v \in \{R,S\}, \quad \forall x \in \{x_1,\ldots,x_n\}.$$

To a sequence $v(\cdot) = (v_0, v_1, \ldots)$ of decisions
is associated the stochastic intertemporal utility

$$J(v(\cdot)) := \sum_{t=0}^{+\infty} \rho^t G(v_t, X_{t+1}) = \sum_{t=0}^{+\infty} \rho^t V(\Phi(v_t, X_{t+1})) = \sum_{t=0}^{+\infty} \rho^t V(Y_{t+1}).$$

Optimal strategies are given by the maximization of the mathematical expectation (under probability laws specified later) of this random discounted reward.

Four information regimes

1. At the beginning of every period t, *i.e.* before he makes his decision, the visionary driver (υ) knows with certainty the travel time on the random road at the end of period t.
2. The fully informed driver (ϕ) knows $\overline{p} = (\overline{p}_1, \ldots, \overline{p}_n)$;
3. The globally informed driver (γ) does not know $\overline{p}$ (but has a prior π_0 on $\overline{p}$) and, at the beginning of period t, knows all past random travel times $X_1, \ldots, X_t$ unconditional on road choice.
4. The locally informed driver (λ) does not know $\overline{p}$ (but has a prior π_0 on $\overline{p}$) and, at the beginning of period t, knows only $Y_1, \ldots, Y_t$ given by $Y_{s+1} = \Phi(v_s, X_{s+1})$, that is only past random travel times when he has selected the random road.

Outline of the presentation

The visionary driver

At the beginning of every period t, the visionary driver knows X_{t+1}, *i.e.* gets full information about the realized value of the travel time on the random road R at the end of period t.

The visionary driver optimal strategy $v^{\upsilon}(\cdot)$ consists in maximizing each $G(v_t, X_{t+1})$ since v_t may depend upon the future X_{t+1}.

Obviously, the visionary driver selects

- road R if $X_{t+1} < x_S$,
- road S otherwise.

The fully informed driver

Recall that the fully informed driver knows $\overline{p}$. Thus, he looks for a strategy $v^\phi(\cdot) = (v_0^\phi, v_1^\phi, \ldots)$ which maximizes

$$\mathbb{E}^{\overline{p}}[J(v^\phi(\cdot))] = \sup_{v(\cdot)} \mathbb{E}^{\overline{p}}[J(v(\cdot))] = \sum_{t=0}^{+\infty} \rho^t \mathbb{E}^{\overline{p}}[V(\Phi(v_t, X_{t+1}))]\,,$$

where

$$\begin{aligned} \mathbb{E}^{\overline{p}}[V(\Phi(S, X_{t+1}))] &= V(x_S) \\ \mathbb{E}^{\overline{p}}[V(\Phi(R, X_{t+1}))] &= \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n)\,. \end{aligned}$$

The relevant road

The relevant road is the optimal road given the knowledge of $\overline{p}$.

Definition

For driver $[V, \rho]$, the relevant road is defined as

- the random road R if $V(x_S) < \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n)$;
- the safe road S otherwise.

The optimal strategy v^ϕ for the fully informed driver is to select the relevant road. The optimal expected discounted reward is

$$\mathbb{E}^{\overline{p}}[J(v^\phi(\cdot))] = \frac{1}{1-\rho} \max\{V(x_S), \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n)\}.$$

A common framework for globally and locally informed drivers

Both the globally and locally informed drivers cannot evaluate an expected discounted reward like $\mathbb{E}^{\overline{p}}[J(v(\cdot))]$ since they do not know $\overline{p}$.

However, both have a prior law π_0 over $\overline{p}$: π_0 is a distribution over the simplex S_{n-1}. With π_0, we may define the probability $\mathbb{P}^{\pi_0}$ on Ω by the marginals

$$\mathbb{P}^{\pi_0}(X_1 = z_1, \ldots, X_t = z_t) = \int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n) \prod_{s=1}^{t} [p_1 \mathbf{1}_{\{z_s = x_1\}} + \cdots + p_n \mathbf{1}_{\{z_s = x_n\}}] .$$

A formulation with a state space of probabilities on S_{n-1}

To solve such problems where the element $\overline{p}$ of the simplex S_n is unknown, it is classical to introduce the space $\mathcal{P}(S_{n-1})$ of probabilities on S_{n-1} as the state space.

Let $[\pi]_i$ denote, for $i = 1, \ldots, n$,

$$\forall \pi \in \mathcal{P}(S_{n-1}), \quad [\pi]_i := \int_{S_{n-1}} p_i \pi_0(dp_1 \cdots dp_n),$$

and

$$[\pi] := ([\pi]_1, \ldots, [\pi]_n) \in S_{n-1},$$

A new reward on this state space

For all $v \in \{R, S\}$ and $\pi \in \mathcal{P}(S_{n-1})$, let us define a new reward $\widetilde{G}(v, \pi)$ by

$$\widetilde{G}(v, \pi) := [\pi]_1 G(v, x_1) + \cdots + [\pi]_n G(v, x_n)\,.$$

The reward for the safe road S is

$$\widetilde{G}(S, \pi) = V(x_S)$$

while the reward for the random road R is

$$\widetilde{G}(R, \pi) = [\pi]_1 V(x_1) + \cdots + [\pi]_n V(x_n)\,.$$

Information

The information available to the globally informed driver is the so called history $\mathcal{X}_t$ up to period t:

$$\mathcal{X}_t := \sigma(X_1, \ldots, X_t)$$

is the σ-field generated by past travel times on the random road.

The information available to the locally informed driver consists only of experienced travel times up to period t:

$$\mathcal{Y}_t := \sigma(Y_1, \ldots, Y_t) = \sigma(\Phi(v_0, X_1), \ldots, \Phi(v_{t-1}, X_t)),$$

This σ-field informs only on past travel times when the driver has selected the random road.

Problem statement

Denote by $\mathcal{I}_t$ the information $\mathcal{X}_t$ or $\mathcal{Y}_t$, according to the context.
For every period t, the decision v_t is measurable with respect to $\mathcal{I}_t$: we shall denote this by

$$v_t \preceq \mathcal{I}_t \,.$$

For the globally or locally informed driver, the prior law π_0 differs from $\delta_{\overline{p}}$ since $\overline{p}$ is not known.
Thus, the globally or locally informed driver looks for a strategy $v(\cdot) = (v_0, v_1, \ldots)$ to maximize the expectation of $J(v(\cdot))$ under probability $\mathbb{P}^{\pi_0}$, where the decision v_t at the beginning of period $[t, t+1[$ depend upon the information $\mathcal{I}_t$ available at this period:

$$\sup_{v_t \preceq \mathcal{I}_t, t \geq 0} \mathbb{E}^{\pi_0}[J(v(\cdot))] \,.$$

The globally informed driver: problem statement

The information available to the globally informed driver is the so called history $\mathcal{X}_t$ up to period t: $\mathcal{X}_t := \sigma(X_1, \ldots, X_t)$ is the σ-field generated by past travel times on the random road.

Thus, the globally informed driver looks for a strategy $v^\gamma(\cdot) = (v_0^\gamma, v_1^\gamma, \ldots)$ to maximize the expectation of $J(v(\cdot))$ under probability $\mathbb{P}^{\pi_0}$, where the decision v_t at the beginning of period $[t, t+1[$ depend upon the information $\mathcal{X}_t$ available at this period:

$$\mathbb{E}^{\pi_0}[J(v^\gamma(\cdot))] = \sup_{v_t \preceq \mathcal{X}_t, t \geq 0} \mathbb{E}^{\pi_0}[J(v(\cdot))]\,.$$

The posterior law of $\overline{p}$ knowing history

For $i = 1, \ldots, n$, let M_t^i be one plus the number of periods in which x_i has been realized:

$$M_t^i := 1 + \sum_{s=1}^{t} \mathbf{1}_{\{X_s = x_i\}} .$$

Let us also define a distribution $\widehat{\pi}_t^\gamma$ on S_{n-1} by

$$\widehat{\pi}_t^\gamma(dp_1 \cdots dp_n) := \frac{\pi_0(dp_1 \cdots dp_n) p_1^{M_t^1 - 1} \cdots p_n^{M_t^n - 1}}{\int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n) p_1^{M_t^1 - 1} \cdots p_n^{M_t^n - 1}} .$$

The state is $\widehat{\pi}_t^\gamma$, interpreted as the posterior law of $\overline{p}$ knowing history $\mathcal{X}_t$.
Observe that the state $\widehat{\pi}_t^\gamma$ varies independently of the road chosen.

The globally informed driver: optimal strategies

Proposition

The optimal globally informed driver

1. *Selects the safe road if and only if*

$$[\widehat{\pi}_t^{\gamma}]_1 V(x_1) + \cdots + [\widehat{\pi}_t^{\gamma}]_n V(x_n) \leq V(x_S)\,.$$

2. *Always selects the relevant road after a random finite number of periods, if its prior π_0 is a beta law.*

The locally informed driver

The locally informed driver has the same information as the globally informed driver up to the point where the former leaves the random road.

The information available to the locally informed driver consists only of experienced travel times up to period t:

$$\mathcal{Y}_t := \sigma(Y_1, \ldots, Y_t) = \sigma(\Phi(v_0, X_1), \ldots, \Phi(v_{t-1}, X_t)), \qquad (3)$$

This σ-field informs only on past travel times when the driver has selected the random road.

The difficulty comes from the fact that now information $\mathcal{Y}_t$ depends upon past decisions $v_0, \ldots, v_{t-1}$.

The locally informed driver as a one-armed bandit problem

For $i = 1, \ldots, n$, let N_t^i be one plus the number of periods x_i has been observed:

$$N_t^i := 1 + \sum_{s=1}^{t} \mathbf{1}_{\{Y_s = x_i\}} .$$

Let us also define a distribution $\widehat{\pi}_t^\lambda$ on S_{n-1} by

$$\widehat{\pi}_t^\lambda(dp_1 \cdots dp_n) := \frac{\pi_0(dp_1 \cdots dp_n) p_1^{N_t^1 - 1} \cdots p_n^{N_t^n - 1}}{\int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n) p_1^{N_t^1 - 1} \cdots p_n^{N_t^n - 1}} .$$

Observe that the state $\widehat{\pi}_t^\lambda$ varies only when the random road is selected: this characterizes bandit problems where one job evolves only if selected.

Gittins index rule

The locally informed driver optimal strategies are expressed by means of $\widehat{\pi}_t^\lambda$ and of the so called Gittins indexes μ_S and μ_R given below. When

$$\mu_S \geq \mu_R(\widehat{\pi}_t^\lambda)\,,$$

the locally informed optimal driver selects the safe road at period t, and conversely.

Gittins indexes

- The index μ_S of the safe road is the constant reward

$$\mu_S(\pi) = \widetilde{G}(S, \pi) = V(x_S), \quad \forall \pi \in \mathcal{P}(S_{n-1}).$$

- The index μ_R of the random road is the following supremum over stopping times $\tau > 0$:

$$\mu_R(\pi) := \sup_{\tau > 0} \frac{\mathbb{E}^\pi[\sum_{t=0}^{\tau-1} \rho^t \widetilde{G}(R, \widehat{\pi}_t^\gamma)]}{\mathbb{E}^\pi[\sum_{t=0}^{\tau-1} \rho^t]}.$$

Optimal index strategy

Proposition

The optimal locally informed driver

1. *selects the safe road if and only if*

$$\mu_R(\widehat{\pi}_t^{\lambda}) \leq V(x_S)\,.$$

2. *sticks to the safe road, once he has selected it (a locally informed driver never switches from the safe road to the random road).*

The second assertion is rather intuitive since once a driver switches to a safe road, he does not update his information, so that there is no reason to shift back to the random road.

Outline of the presentation

Locally versus globally informed driver

The road choice optimal behaviors of the locally and of the globally informed drivers are different but related in a way specified by Proposition below.

Proposition

1. *If an optimal locally informed driver selects the safe road for a given prior, he would also do so if he was globally informed with the same prior.*
2. *If an optimal globally informed driver selects the random road from the first period up to a period t, he would do the same if he was locally informed and facing the same realizations of random travel times on the random road.*

Comparison of optimal expected discounted rewards

We can now compare the optimal expected discounted rewards for the four information regimes envisaged.

Under the probability law $\mathbb{P}^{\overline{P}}$ which drives the realizations of random times on the random road, we have the following inequalities for optimal expected discounted rewards:

$$\mathbb{E}^{\overline{P}}[J(v^{\upsilon}(\cdot))] \geq \mathbb{E}^{\overline{P}}[J(v^{\phi}(\cdot))] \geq \max(\mathbb{E}^{\overline{P}}[J(v^{\gamma}(\cdot))], \mathbb{E}^{\overline{P}}[J(v^{\lambda}(\cdot))]) .$$

Risk aversion and locally informed drivers

Proposition

Consider two drivers with common belief π_0 and one more risk averse than the other.

Assume that, at the beginning, the more risk averse driver selects the random road based on π_0. Then, so does the less risk averse driver and, as long as the more risk averse driver selects the random road, so does also the less risk averse driver.

Moreover, the mean number of periods spent selecting the random road decreases with the degree of absolute risk aversion.

Outline of the presentation

For this numerical illustration, we restrict to the case $n = 2$. The random road R has a random travel time X_{t+1} realized at the end of period t which takes value in $\{x_-, x_+\}$. We assume that

$$x_- < x_S < x_+ \ ,$$

We suppose that x_- occurs with probability $\overline{p} \in]0, 1[$.

For numerical experiments we have used $x^- = 10/60$, $x_s = 20/60$ and $x^+ = 22/60$, discount rate $\rho = 1/1.08$ and the following CARA utility function

$$V_\theta(x) = \frac{1 - e^{\theta x}}{\theta} \ .$$

The parameter θ is the Arrow-Pratt degree of absolute risk aversion $-V''_\theta / V'_\theta$.

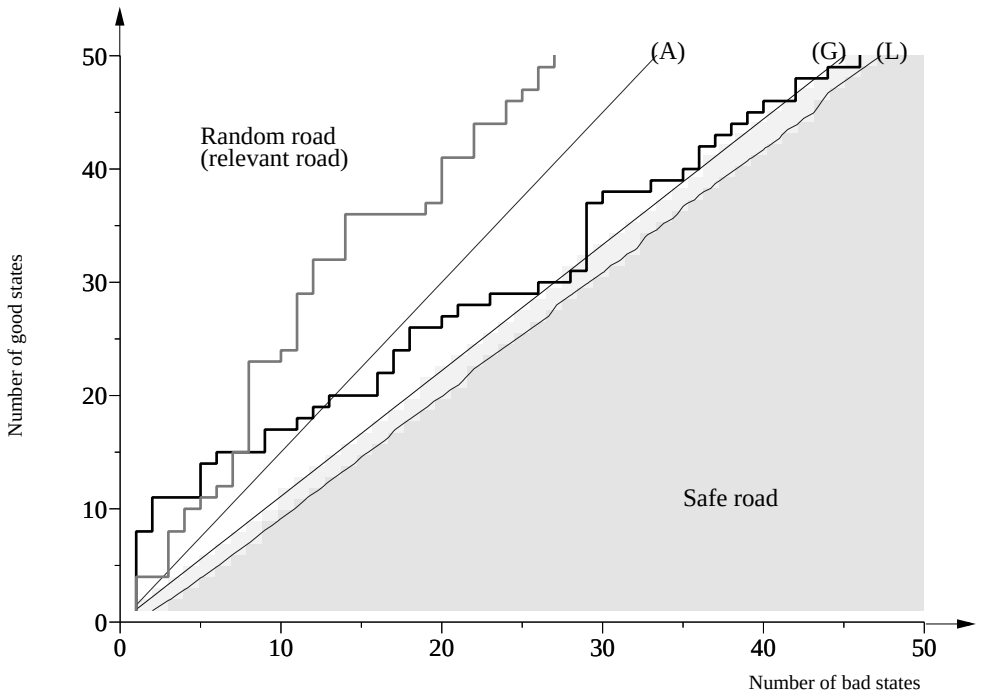


Figure: Two trajectories of posterior laws when the relevant road is the random road. On the horizontal axis is counted the number of bad states (high travel time), while the vertical axis corresponds to the number of good states (low travel time). The (L) curve is the switching curve for the locally informed driver. The (G) curve is the switching curve for the globally informed driver. The (A) curve has slope $\overline{p}$.

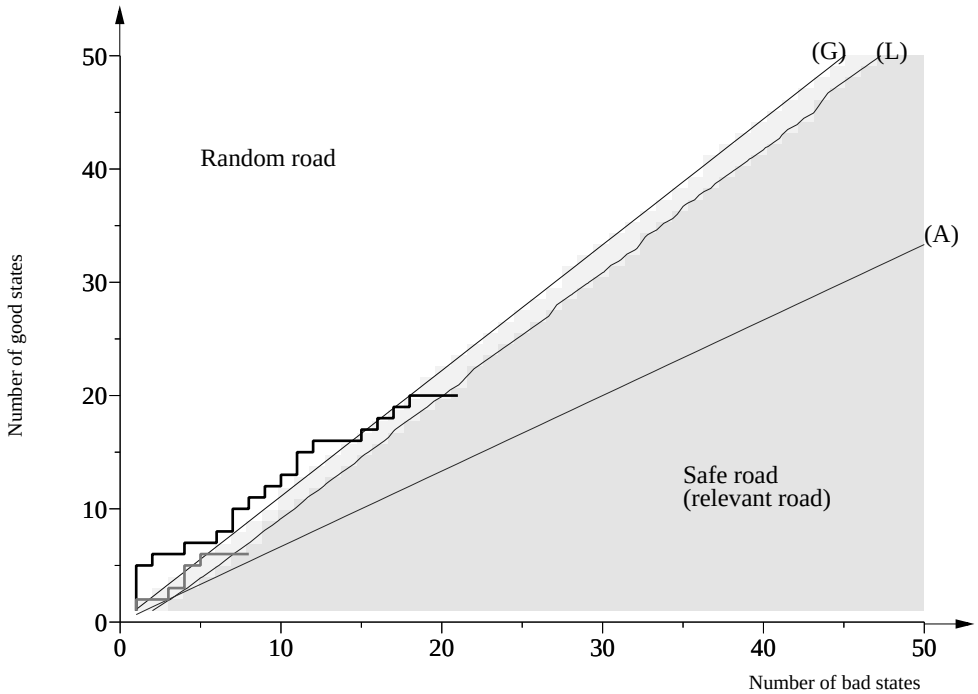


Figure: Two trajectories of posterior laws when the relevant road is the safe road. On the horizontal axis is counted the number of bad states (high travel time), while the vertical axis corresponds to the number of good states (low travel time). The (L) curve is the switching curve for the locally informed driver. The (G) curve is the switching curve for the globally informed driver. The (A) curve has slope $\overline{p}$.

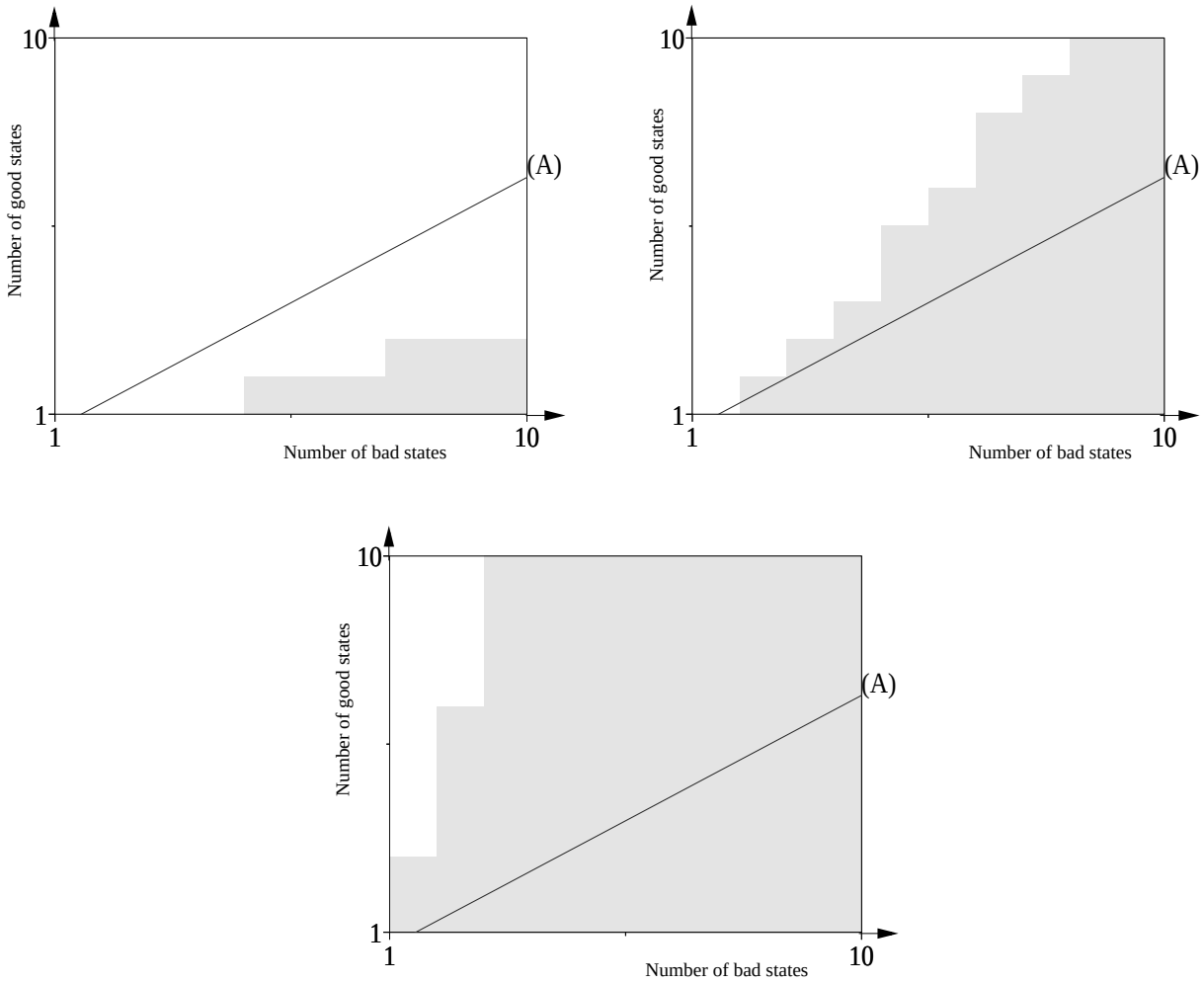


Figure: Gittins rules for increasing values of absolute risk aversion θ. In the grey zone, the locally informed driver optimaly selects the safe road.

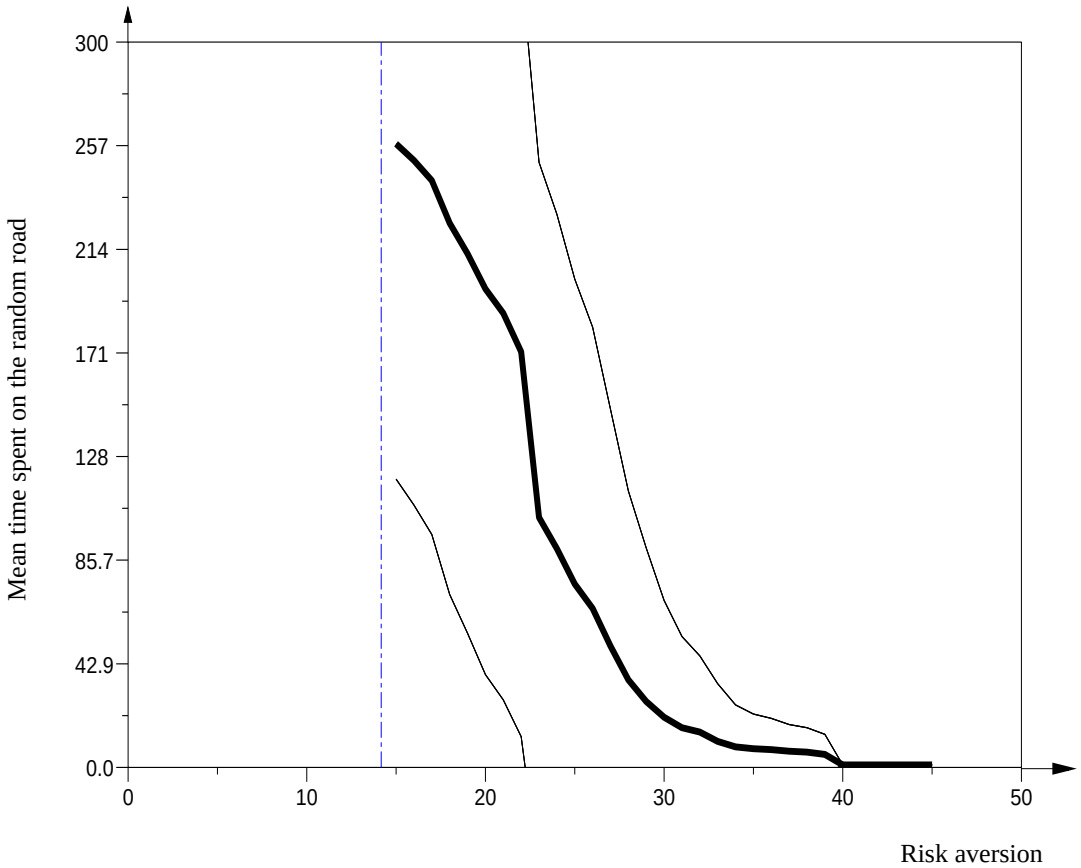


Figure: Mean number of periods spent on the random road ($\pm$ one standard deviation) as a function of absolute risk aversion θ