Numerical approximation of average Markov decision processes in continuous-time

Jonatha Anselmi; François Dufour; Tomàs Prieto-Rumeau

INRIA Bordeaux Sud Ouest, France

Institut Polytechnique de Bordeaux
INRIA Bordeaux Sud-Ouest
Institut de Mathématiques de Bordeaux, France

UNED, Madrid, Spain

Research supported by grant MTM2016-75497-P from the Spanish Ministerio de Economía y Competitividad.

Outline

1. Continuous-time Markov decision processes (CT-MDPs)
   - Introduction
   - Approach
2. The control model $\mathcal{M}$
   - Parameters
   - Construction
   - Admissible strategies and distribution of the process
   - Optimization problem
   - Assumptions and basic result
3. Approximation of the value function
   - Construction of the approximating finite model $\mathcal{M}_{k,\delta}$
   - Estimation of the value function
   - Approximation with empirical probability measure
4. Numerical example
   - The model
   - Numerical results with empirical distribution

CT-MDPs

Important modeling tool for complex systems.

General class of continuous-time stochastic models: piecewise constant trajectory punctuated by random jumps.

R. Bellman [*Dynamic programming*,1957];

R. Howard [*Dynamic programming and Markov processes*,1960];

Applications

Engineering, medecine, biology, operations research, management science, economics, dependability and safety, ...

Theoretical point of view

- Continuous-time Markov decision processes (CT-MDPs) extensively studied.
- Dynamic programming and linear programming approaches: Policy Iteration Algorithm, Value Iteration Algorithm.
- Existence of optimal policies, smoothness of the value function, sufficiency of specific classes of policies,...

Approaches hardly applicable in practice.
Solving explicitly or numerically a CT-MDP $\rightsquigarrow$ critical issue.

- Calculation of the value function and an optimal policy?
- Solving an MDP $\rightsquigarrow$ numerical methods to get quasi-optimal solutions.

Numerical point of view

- In the discrete-time framework: two groups of methods.
  - MDPs with discrete state and action spaces (large but finite or countable). Stochastic approximation: reinforcement learning, neuro-dynamic programming, simulation based methods,...
  - MDPs with uncountable state and action spaces (Borel space). Approximation by means of MDP with finite state/action spaces.
- In the continuous-time context: few techniques on this topic (P. Dupuis & H. Kushner via the weak convergence theory)

Our objective $\rightsquigarrow$ a method of approximation for the value function of CT-MDP with Borel state space $\mathbf{X}$ and Borel action space $\mathbf{A}$ under the expected long-run average cost criterion.

Overall idea and properties

i) Discretization of the state space.

- The positive part q^+ of the transition rate q of $\mathcal{M}$

$$q^+(dy|x,a) = p(y|x,a)\mu(dy)$$

where μ is a reference probability measure on $\mathbf{X}$.
- Replace μ by a probability measure μ_k with finite support $\mathbf{X}_k$.
- Consider the model with finite state space $\mathbf{X}_k$ and

$$q_k^+(dy|x,a) = p(y|x,a)\mu_k(dy).$$

Error $\rightarrow$ Wasserstein distance between μ and μ_k: $\mathcal{W}(\mu,\mu_k)$.

ii) Discretization of the action sets.

- Replace the action sets $\mathbf{A}$ with a finite subset $\mathbf{A}_\delta$ parametrized by $\delta > 0$.
  Error $\rightarrow$ the Hausdorff distance between $\mathbf{A}$ and $\mathbf{A}_\delta$ which is assumed to be of (small) order $\delta > 0$.

$\mathcal{M}(\mathbf{X}, \mathbf{A}, q)$
General model
$(q^+ \ll \mu)$
Value function $\mathcal{J}^*$

approximation
$\rightsquigarrow$

$\mathcal{M}_{k,\delta}(\mathbf{X}_k, \mathbf{A}_\delta, q_k)$
Finite model
Value function $\mathcal{J}^*_{k,\delta}$

Properties of the approximation procedure:

- Control explicitly the approximation error and to get non asymptotic bounds depending on $\mathcal{W}(\mu, \mu_k)$ and δ

$$|\mathcal{J}^* - \mathcal{J}^*_{k,\delta}| = \mathrm{O}(\mathcal{W}(\mu, \mu_k)) + \mathrm{O}(\delta).$$

- Discretization of $\mathbf{X}$ based on a probabilistic criterion: points are placed in $\mathbf{X}$ according to their importance in the Markov transition kernel. We do not use a geometric criterion as opposed to the usual discretization procedures.
  $\rightsquigarrow$ A direct approach to discretize a not necessarily compact state space $\mathbf{X}$ as opposed to classical procedures that need a *compactification* step.

Construction of an approximation of μ with finite support in the Wasserstein metric.

Two different approaches:

- One consists in deriving μ_k starting from a covering of $\mathbf{X}$ with small radius. This "deterministic" approach allows controlling the distance $\mathcal{W}(\mu, \mu_k) \rightsquigarrow$ additional computational challenge.
- Another possibility is to use a "random" approximation by considering the empirical probability measure μ_k obtained from k i.i.d. draws from μ.
  The approximation error $\rightsquigarrow$ concentration inequality for the non-asymptotic deviation converging to zero in probability at an exponential speed in the sample size k.

Outline

1. Continuous-time Markov decision processes (CT-MDPs)
   - Introduction
   - Approach
2. The control model $\mathcal{M}$
   - Parameters
   - Construction
   - Admissible strategies and distribution of the process
   - Optimization problem
   - Assumptions and basic result
3. Approximation of the value function
   - Construction of the approximating finite model $\mathcal{M}_{k,\delta}$
   - Estimation of the value function
   - Approximation with empirical probability measure
4. Numerical example
   - The model
   - Numerical results with empirical distribution

We deal with a control model $\mathcal{M} = \{\mathbf{X}, \mathbf{A}, \{\mathbf{A}(x)\}_{x\in\mathbf{X}}, q, c\}$:

- The state space $\mathbf{X}$ is a Borel space.
- The action space $\mathbf{A}$ is a Borel space. Set of feasible actions $\mathbf{A}(x)$ in state $x \in \mathbf{X}$. Admissible state-action set:

$$\mathbf{K} = \{(x, a) \in \mathbf{X} \times \mathbf{A} : a \in \mathbf{A}(x)\} \in \mathcal{B}(\mathbf{X} \times \mathbf{A}).$$

- The transition rate q is a bounded conservative signed kernel on $\mathbf{X}$ given $\mathbf{K}$, i.e.,

$$\hat{q} = \sup_{(x,a)\in\mathbf{K}} \{-q(\{x\}|x, a)\} < \infty,$$

and

$$q(\mathbf{X}|x, a) = 0 \quad \text{for all } (x, a) \in \mathbf{K}$$

- The cost rate is a bounded measurable function $c : \mathbf{K} \to \mathbb{R}$.

The canonical space

$$\Omega = \Big(\bigcup_{n=1}^{\infty} \Omega_n\Big) \bigcup (\mathbf{X} \times (\mathbb{R}_+^* \times \mathbf{X})^\infty)$$

with $\Omega_n = \mathbf{X} \times (\mathbb{R}_+^* \times \mathbf{X})^n \times (\{\infty\} \times \{x_\infty\})^\infty$.

Interpretation of Ω:

Consider $\omega = (x_0, \theta_1, x_1, \theta_2, x_2, \ldots) \in \Omega$

- $x_0 \in \mathbf{X}$ is the initial state.
- Given $n \geq 0$, if $x_n \in \mathbf{X}$ then
  - either $0 < \theta_{n+1} < \infty$, and we interpret θ_{n+1} as the sojourn time in state $x_n \in \mathbf{X}$, while $x_{n+1} \in \mathbf{X}$ is the post-jump location of the process;
  - or $\theta_{n+1} = \infty$; this means that the system has been absorbed by x_n. In this case we set $x_m = x_\infty$ and $\theta_m = \infty$ for all $m > n$. Such sample paths belong to Ω_n.

Introduce the location X_n and the sojourn time Θ_n

$$X_n : \ \Omega \to \mathbf{X}_\infty = \mathbf{X} \cup \{x_\infty\} \text{ by } X_n(\omega) = x_n,$$

$$\Theta_n : \ \Omega \to \overline{\mathbb{R}}_+^* \text{ by } \Theta_n(\omega) = \theta_n, \text{ with } \Theta_0(\omega) = 0,$$

for $\omega = (x_0, \theta_1, x_1, \theta_2, x_2, \ldots) \in \Omega$.

Time of jumps: $T_n = \sum_{i=1}^{n} \Theta_i$, explosion time: $T_\infty = \lim_{n\to\infty} T_n$

The continuous-time process $\{\xi_t\}_{t\geq 0}$ with values in $\mathbf{X}_\infty$ is given by

$$\xi_t(\omega) = \begin{cases} X_n(\omega), & \text{if } T_n(\omega) \leq t < T_{n+1}(\omega) \text{ for } n \in \mathbb{N}, \\ x_\infty, & \text{if } t \geq T_\infty(\omega). \end{cases}$$

Control strategies

- An admissible control policy is a sequence $u = (\pi_n)_{n\in\mathbb{N}}$ where, for any $n \in \mathbb{N}$, π_n is a stochastic kernel on $\mathbf{A}_\infty = \mathbf{A} \cup \{a_\infty\}$ given $\mathbf{H}_n \times \mathbb{R}_+^*$ satisfying

$$\pi_n(\mathbf{A}(x_n)|h_n, t) = 1,$$

for any $h_n = (x_0, \theta_1, x_1, \ldots \theta_n, x_n) \in \mathbf{H}_n$ and $t \in \mathbb{R}_+^*$ and with $\mathbf{A}(x_\infty) = \{a_\infty\}$.

The set of admissible control policies is denoted by $\mathcal{U}$.

- Associated to $u = (\pi_n)_{n\in\mathbb{N}} \in \mathcal{U}$, the control process takes values in $\mathcal{P}(\mathbf{A}_\infty)$

$$\pi(da|t) = \sum_{n\in\mathbb{N}} I_{\{T_n<t\leq T_{n+1}\}}\pi_n(da|H_n, t-T_n)+\mathbf{I}_{\{t\geq T_\infty\}}\delta_{a_\infty}(da).$$

For a given $u = (\pi_n)_{n\in\mathbb{N}} \in \mathcal{U}$, the conditional distribution of $(\Theta_{n+1}, X_{n+1})_{n\in\mathbb{N}}$ w.r.t. the past $H_n = (X_0, \Theta_1, X_1, \ldots, \Theta_n, X_n) \rightsquigarrow$

$$\delta_{(\infty,x_\infty)}(\Gamma)\Big[\delta_{X_n}(\{x_\infty\}) + \delta_{X_n}(\mathbf{X})e^{-\Lambda_n(\mathbf{X},H_n,\infty)}\Big] + \delta_{X_n}(\mathbf{X})\int_{\Gamma\cap(\mathbb{R}_+^*\times\mathbf{X})} \lambda_n(dx, H_n, t)e^{-\Lambda_n(\mathbf{X},H_n,t)}dt,$$

for $\Gamma \in \mathcal{B}(\overline{\mathbb{R}}_+^* \times \mathbf{X}_\infty)$ and where

$$\lambda_n(\cdot, H_n, t) = \int_{\mathbf{A}_\infty} q^+(\cdot|X_n, a)\pi_n(da|H_n, t),$$

and

$$\Lambda_n(\mathbf{X}, H_n, t) = \int_0^t \lambda_n(\mathbf{X}, H_n, s)ds.$$

Consider $u = (\pi_n)_{n\in\mathbb{N}} \in \mathcal{U}$ and an initial state $x \in \mathbf{X}$.
There exists a probability $\mathbb{P}^{x,u}$ on $(\Omega, \mathcal{F})$ such that

$$\mathbb{P}^{x,u}\{X_0 = x\} = 1$$

$$\mathbb{P}^{x,u}\{(\Theta_{n+1}, X_{n+1}) \in \Gamma \mid H_n\}$$

$$= \delta_{(\infty,x_\infty)}(\Gamma)\Big[\delta_{X_n}(\{x_\infty\}) + \delta_{X_n}(\mathbf{X})e^{-\Lambda_n(\mathbf{X},H_n,\infty)}\Big]$$

$$+ \delta_{X_n}(\mathbf{X}) \int_{\Gamma\cap(\mathbb{R}_+^*\times\mathbf{X})} \lambda_n(dx, H_n, t)e^{-\Lambda_n(\mathbf{X},H_n,t)}dt,$$

with

$$\lambda_n(\cdot, H_n, t) = \int_{\mathbf{A}_\infty} q^+(\cdot|X_n, a)\pi_n(da|H_n, t).$$

(Jacod, *Multivariate point processes*, 1975).

The expected long-run average cost

$$\mathcal{J}(u,x) := \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}^{x,u}\Big[\int_0^{t\wedge T_\infty}\int_{\mathbf{A}(\xi_s)} c(\xi_s,a)\pi(da|s)ds\Big],$$

and $\mathcal{J}^*(x) := \inf_{u\in\mathcal{U}} \mathcal{J}(x,u)$ for $x \in \mathbf{X}$.

$u^* \in \mathcal{U}$ is average optimal if $\mathcal{J}(x,u^*) = \mathcal{J}^*(x)$ for any $x \in \mathbf{X}$.

The expected α-discounted cost for $\alpha \in (0,1)$ is given by

$$\mathcal{V}_\alpha(u,x) = \mathbb{E}^{x,u}\Big[\int_0^{T_\infty} e^{-\alpha s}\int_{\mathbf{A}(\xi_s)} c(\xi_s,a)\pi(da|s)ds\Big],$$

and $\mathcal{V}^*_\alpha(x) := \inf_{u\in\mathcal{U}} \mathcal{V}_\alpha(x,u)$ for $x \in \mathbf{X}$.

Assumptions

These include a standard Lyapunov condition and continuity-compactness requirements, plus some additional conditions to ensure the existence of a *smooth* solution to the average cost optimality inequalities.

Assumption A.

(A1) *The cost function c is L_c-Lipschitz continuous on $\mathbf{K}$.*

(A2) *The multifunction $\Psi : x \to \mathbf{A}(x)$ is compact-valued and L_Ψ-Lipschitz continuous with respect to the Hausdorff distance.*

(A3) *For any $v \in \mathbb{C}(\mathbf{X})$ we have $qv \in \mathbb{C}(\mathbf{K})$.*

(A4) *There exists $L_Q > 0$ satisfying that $(L_\Psi + 1)L_Q < 1$ such that for every L_v-Lipschitz continuous function $v \in \mathbb{L}(\mathbf{X})$ we have $Qv \in \mathbb{L}(\mathbf{K})$, with Lipschitz constant $L_Q L_v$ where*

$$Q(dy|x,a) = \frac{1}{\hat{q}} q(dy|x,a) + \delta_x(dy)$$

Remarks:

(A1)-(A3) standard continuity-compactness conditions (existence results).

(A4) additional hypothesis to control the approximation errors.

Assumption B. There is a function $w : \mathbf{X} \to [1, \infty)$ with the following properties.

(B1) $w \in \mathbb{L}(\mathbf{X})$ *and there exist constants* $\rho > 0$ *and* $\gamma \geq 0$ *with* $qw(x, a) \leq -\rho w(x) + \gamma$ *for any* $(x, a) \in \mathbf{K}$.

(B2) *There exists* $x_0 \in \mathbf{X}$ *such that the relative difference of the discounted value function* $h_\alpha(x) := \mathcal{V}^*_\alpha(x) - \mathcal{V}^*_\alpha(x_0)$ *satisfies*

$$\sup_{\alpha > 0} \|h_\alpha\|_w < \infty.$$

Remarks: (B1)-(B2) standard technical conditions. Sufficient conditions for (B2):

- uniform ergodic properties
- drift and monotonicity conditions
- communication properties and hitting time of the process

Theorem (of existence)

Suppose that the control model $\mathcal{M}$ satisfies Assumptions A and B.

(i) *There exist a constant g^* and functions $v_1, v_2 \in \mathbb{L}_w(\mathbf{X})$ that are solutions of the average optimality inequalities:*

$$g^* \geq \inf_{a \in \mathbf{A}(x)} \left\{ c(x,a) + \int_{\mathbf{X}} v_1(y) q(dy|x,a) \right\}$$

$$g^* \leq \inf_{a \in \mathbf{A}(x)} \left\{ c(x,a) + \int_{\mathbf{X}} v_2(y) q(dy|x,a) \right\}$$

for every $x \in \mathbf{X}$.

(ii) *The optimal average cost of $\mathcal{M}$ is constant and $g^* = \mathcal{J}^*(x)$ for all $x \in \mathbf{X}$.*

(iii) *Any $f \in \mathbb{F}$ attaining the infimum in the first inequality is average optimal for $\mathcal{M}$, and such f indeed exists.*

Outline

1. Continuous-time Markov decision processes (CT-MDPs)
   - Introduction
   - Approach
2. The control model $\mathcal{M}$
   - Parameters
   - Construction
   - Admissible strategies and distribution of the process
   - Optimization problem
   - Assumptions and basic result
3. Approximation of the value function
   - Construction of the approximating finite model $\mathcal{M}_{k,\delta}$
   - Estimation of the value function
   - Approximation with empirical probability measure
4. Numerical example
   - The model
   - Numerical results with empirical distribution

Assumption C. There exist a probability measure $\mu \in \mathcal{P}_1(\mathbf{X})$ and a nonnegative function p defined on $\mathbf{X} \times \mathbf{K}$ such that:

(C1) For all $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$ we have

$$q^+(B|x,a) = \int_B p(y|x,a)\mu(dy).$$

(C2) There is some $L_p > 0$ such that the function $p(\cdot|x,\cdot)$ is L_p-Lipschitz continuous on $\mathbf{X} \times \mathbf{A}(x)$ for any $x \in \mathbf{X}$.

(C3) For some positive constants $\hat{p}$ and L_{wp} we have

$$p(y|x,a) \leq \hat{p}w(x),$$

$$\left|w(y)p(y|x,a) - w(z)p(z|x,a)\right| \leq L_{wp}w(x)d_{\mathbf{X}}(y,z).$$

(C4) For each $\delta > 0$, there is a finite set $\mathbf{A}_\delta(x) \subseteq \mathbf{A}(x)$ such that the multifunction defined on $\mathbf{X}$ by $x \mapsto \mathbf{A}_\delta(x)$ is Borel-measurable and $d_H(\mathbf{A}(x), \mathbf{A}_\delta(x)) \leq \delta w(x)$.

Parameters of $\mathcal{M}_{k,\delta}$

The elements of the control model $\mathcal{M}_{k,\delta}$ are given by

$$\{\mathbf{X}, \mathbf{A}, \{\mathbf{A}_\delta(x)\}_{x\in\mathbf{X}}, q_k, c\},$$

with q_k defined as follows:

$$q_k(B|x,a) = \int_B p(y|x,a)\mu_k(dy) - \int_{\mathbf{X}} p(y|x,a)\mu_k(dy)\delta_x(B),$$

for $B \in \mathcal{B}(\mathbf{X})$ and $(x,a) \in \mathbf{K}_\delta$ where

$$\mathbf{K}_\delta = \{(x,a) \in \mathbf{X}\times\mathbf{A} : a \in \mathbf{A}_\delta(x)\} \in \mathcal{B}(\mathbf{X}\times\mathbf{A}).$$

The model $\mathcal{M}_{k,\delta}$ is finite: $\mathbf{X}_k$ (support of μ_k) is an absorbing set $\rightsquigarrow$ starting from any initial state in $\mathbf{X}$, after the first transition the state of the system belongs to $\mathbf{X}_k$.

Control policies and value function for $\mathcal{M}_{k,\delta}$.

- An admissible control $u = (\pi_n)_{n\in\mathbb{N}}$ where π_n is a stochastic kernel on $\mathbf{A}_\infty$ given $\mathbf{H}_n \times \mathbb{R}_+^*$: $\pi_n(\mathbf{A}_\delta(x_n)|h_n, t) = 1$.
- The set of admissible control policies $\mathcal{U}_\delta \subset \mathcal{U}$.

There exists $\mathbb{P}_{k,\delta}^{x,u}$ on $(\Omega, \mathcal{F})$ modelling the CT-MDP $\mathcal{M}_{k,\delta}$.

- The expected long-run average cost of the control policy $u \in \mathcal{U}_\delta$ for the initial state $x \in \mathbf{X}$ is given by

$$\mathcal{J}_{k,\delta}(x,u) := \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}_{k,\delta}^{x,u}\Big[\int_0^{t\wedge T_\infty}\int_{\mathbf{A}_\delta(\xi_s)} c(\xi_s, a)\pi(da|s)ds\Big]$$

- The value function of the average cost control problem is

$$\mathcal{J}_{k,\delta}^*(x) := \inf_{u\in\mathcal{U}_\delta} \mathcal{J}_{k,\delta}(x,u) \quad \text{for } x \in \mathbf{X}.$$

Bound for the approximation error:

Theorem

Suppose that Assumptions A–C hold. For any $x \in \mathbf{X}$, $\delta > 0$, *and* $k \geq 1$ *such that* $\mathcal{W}(\mu, \mu_k) \leq \frac{\rho}{2(L_{wp}+L_p)}$, *we have*

$$\sup_{x \in \mathbf{X}} \left| g^* - \mathcal{J}^*_{k,\delta}(x) \right| \leq \mathcal{C}\delta + \mathcal{D}\mathcal{W}(\mu, \mu_k),$$

where $\mathcal{C} = \left[L_c + \|v_1\|_w \mu(w) L_p \right] \frac{2\gamma}{\rho}$,
$\mathcal{D} = \left(\left[\mathcal{L}_1 + L_p \|v_1\|_w \right] \vee \left[\mathcal{L}_2 + L_p \|v_2\|_w \right] \right) \frac{2\gamma}{\rho}$,
$\mathcal{L}_{1/2} = \|v_{1/2}\|_w (L_{wp} + \hat{p} L_w) + L_{v_{1/2}} \hat{p}$.

We now specialize the main result to the case when the measure μ_k is given by the empirical probability measure drawn from the measure μ.
In other words, μ_k is a random probability measure defined on the probability space $(\mathbf{X}^\infty, \mathcal{B}(\mathbf{X})^\infty, \mathbb{P}_\mu)$ defined by

$$\mu_k(\zeta) = \frac{1}{n} \sum_{i=1}^{k} \delta_{\zeta_i} \in \mathcal{P}(\mathbf{X})$$

where $\zeta = (\zeta_1, \zeta_2, \ldots) \in \mathbf{X}^\infty$.

Proposition (Boissard, 2011)

Given $\mu \in \mathcal{P}_{\exp}(\mathbf{X})$ and $\epsilon > 0$, there exist positive constants C_ϵ and D_ϵ such that

$$\mathbb{P}_\mu\{\mathcal{W}(\mu, \mu_k(\zeta)) > \epsilon\} \leq C_\epsilon \exp\{-D_\epsilon k\} \quad \textit{for all } k \geq 1.$$

Theorem

Suppose that the control model $\mathcal{M}$ satisfies Assumptions A–C with $\mu \in \mathcal{P}_{\exp}(\mathbf{X})$. Fix an initial state $x \in \mathbf{X}$ and let $\epsilon > 0$ be some given precision. There exist $\delta > 0$, and positive constants $C(\epsilon)$ and $D(\epsilon)$ such that

$$\mathbb{P}_\mu\{\sup_{x\in\mathbf{X}} |\mathcal{J}^*(x) - \mathcal{J}^*_{k,\delta}(x)| > \epsilon\} \le C(\epsilon)e^{-D(\epsilon)k} \quad \textit{for all } k \ge 1.$$

Outline

1. Continuous-time Markov decision processes (CT-MDPs)
   - Introduction
   - Approach
2. The control model $\mathcal{M}$
   - Parameters
   - Construction
   - Admissible strategies and distribution of the process
   - Optimization problem
   - Assumptions and basic result
3. Approximation of the value function
   - Construction of the approximating finite model $\mathcal{M}_{k,\delta}$
   - Estimation of the value function
   - Approximation with empirical probability measure
4. Numerical example
   - The model
   - Numerical results with empirical distribution

Parameters of $\mathcal{M}$

Consider the system with the following elements:

- The state space is $\mathbf{X} = [0, C]$ for some $C > 0$.
- The action space is some interval $\mathbf{A} = [a_m, a_M]$. We suppose that $\mathbf{A}(x) = \mathbf{A}$ for all $x \in \mathbf{X}$.
- The transition kernel is given by

$$q(B|x,a) = \int_{B\cap(x,C]} 2(y-x)dx + a\mathbf{I}_B(0) - (a+(C-x)^2)\mathbf{I}_B(x)$$

  for any $0 \leq x \leq C$, $a_m \leq a \leq a_M$, and $B \in \mathcal{B}(\mathbf{X})$.
- The cost rate function defined on $\mathbf{X} \times \mathbf{A}$ by $c(x,a) = (1-x)(10-a)$.

By considering $\mu = \eta\delta_0 + \dfrac{1-\eta}{C}\lambda$, for $0 < \eta < 1$ and where λ is the Lebesgue measure on $(0, C]$, we can show that

$$q^+(B|x,a) = \int_B p(y|x,a)\mu(dy)$$

with

$$p(y|x,a) = \begin{cases} \frac{2C}{1-\eta}(y-x) & \text{for } 0 < y \leq C, \\ a/\eta & \text{for } 0 = y \end{cases}$$

Let $\mathbf{A}_\delta$ to be $k+1$ equally spaced points in $\mathbf{A}$

- If $a_m > (1+C)(2C+1)$ then Assumptions A–C are satisfied.
- $C = 1$ and $\mathbf{A} = [7, 8]$

Stochastic approximation of $\mu = \eta\delta_0 + \dfrac{1-\eta}{C}\lambda$

For $\mu_k = \eta\delta_0 + \dfrac{1-\eta}{k}\displaystyle\sum_{j=1}^{k}\delta_{x_j}$, where $\{x_j\}$ is a sample of size k of the uniform distribution $(0, C]$.

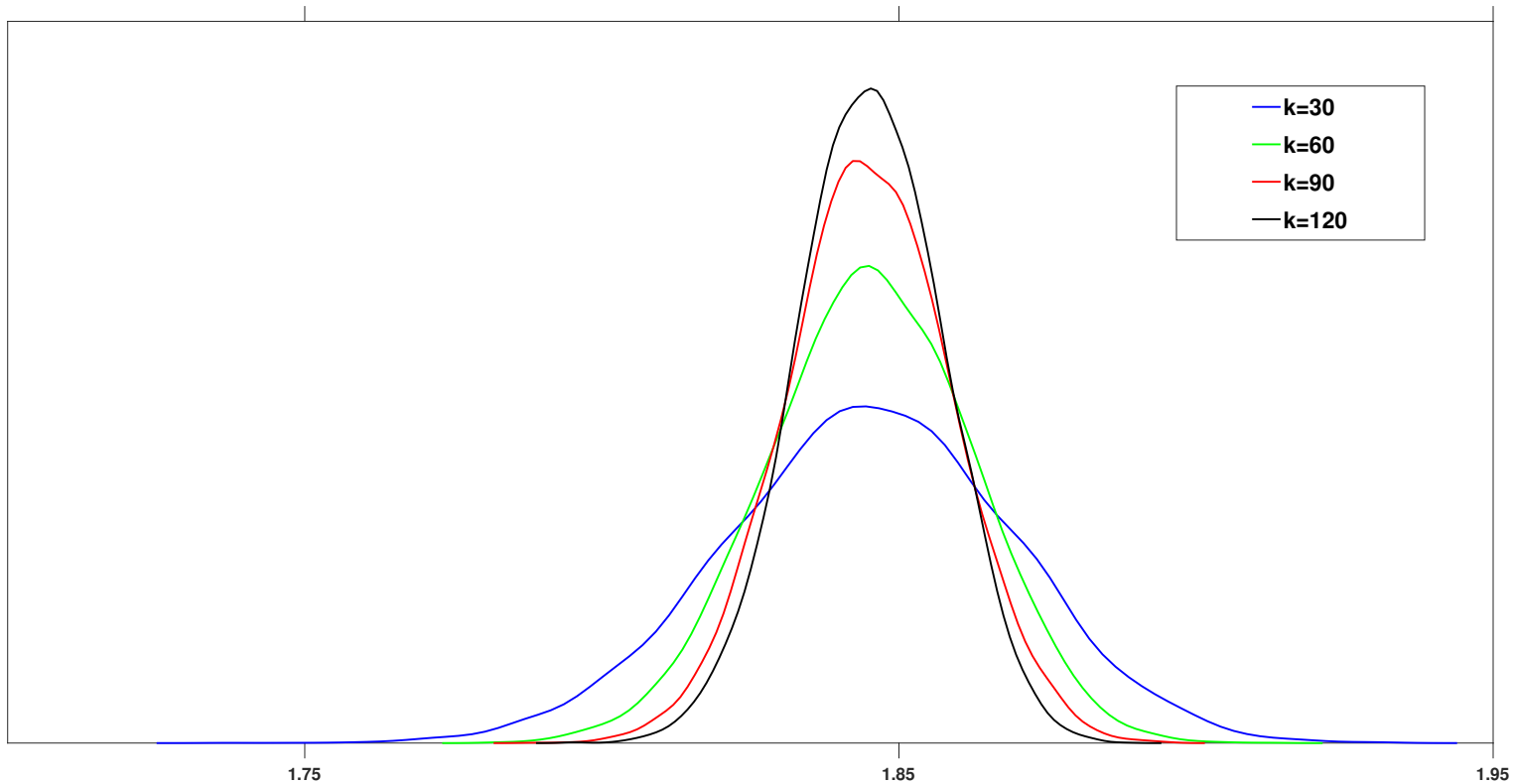
- 10 000 samples of size k of the uniform distribution in $(0, C]$.
- For each sample, $\mathcal{M}_{k,\delta}$ is solved by using the PIA.
  $\rightsquigarrow$ 10 000 realizations of the optimal cost $g^*_{k,\delta}$:

	$k = 30$	$k = 60$	$k = 90$	$k = 120$
Mean	1.8449	1.8450	1.8449	1.8450
Std. Dev.	0.0243	0.0172	0.0139	0.0122

Table: Estimation of the optimal average cost $g^*_{k,\delta}$

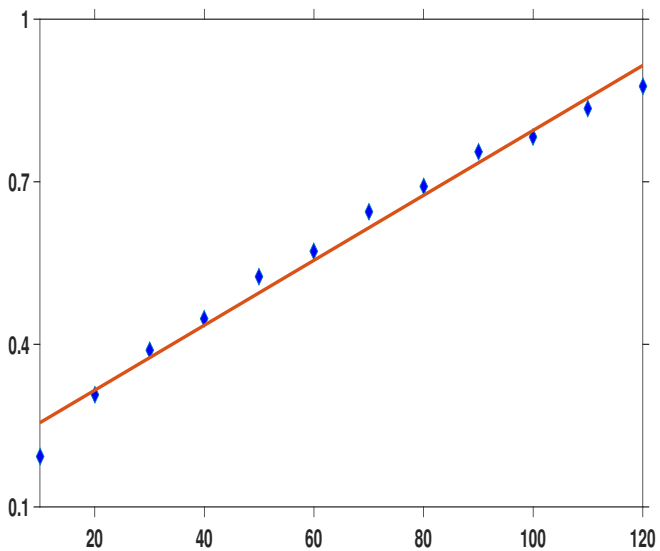
The estimations are very stable (the mean is practically the same for all values of k) and they become more concentrated as k grows (the standard deviation decreases).

The density estimators (based on normal kernels) of the 10 000 approximation of the optimal costs for the different values of k:

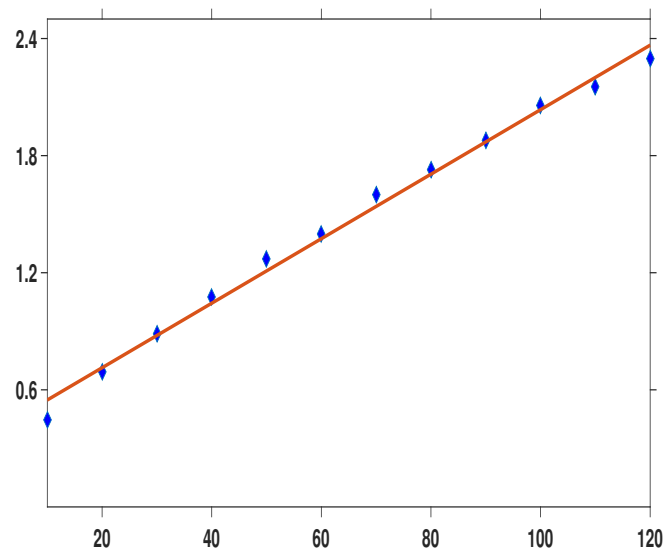


The estimations $\rightsquigarrow$ more accurate and concentrated as k grows.

- To check empirically the order of convergence, we have estimated $\mathbb{P}\{|g^* - g^*_{k,\delta}| > \epsilon\}$ by the empirical probability (denoted by p_k), that $|g^*_{k,\delta} - \overline{g^*_{k,\delta}}| > \epsilon$ for $\epsilon = 0.01$; 0.02 and $k = 10, 20, \ldots, 120$.
- We then perform a linear regression: $-\log p_k \sim \beta_0 + \beta_1 k$:



(a) Error $\epsilon = 0.01$.



(b) Error $\epsilon = 0.02$.

Figure: Regression line of $-\log p_k$.

Finally, we got the following estimations

$$\mathbb{P}\{|g^* - g^*_{k,\delta}| > 0.01\} \simeq 0.8224 \cdot e^{-0.006k}$$

and

$$\mathbb{P}\{|g^* - g^*_{k,\delta}| > 0.02\} \simeq 0.682 \cdot e^{-0.0165k}.$$

- The *non asymptotic bound* given by our results is tight (its order is indeed attained)
- The estimation of the involved constants is possible.
- The multiplicative constant $C(\epsilon)$, in particular, takes a *reasonably* low value.

Deterministic approximation of $\mu = \eta\delta_0 + \dfrac{1-\eta}{C}\lambda$

For $\mu_k = \eta\delta_0 + \dfrac{1-\eta}{k}\displaystyle\sum_{j=1}^{k}\delta_{x_j}$, where $\{x_i\}$ defined by $x_j = \frac{jC}{k}$.

- It can be shown that $\mathcal{W}(\mu, \mu_k) = \frac{(1-\eta)C}{2k}$.
- We expect that $\left|g^* - g^*_{k,\delta}\right| = \mathrm{O}(1/k)$.

For $k = 50, \ldots, 1200$ (with steps of size 50) we have solved $\mathcal{M}_{k,\delta}$. We observe that the approximations $g^*_{k,\delta}$ become very stable as k grows.

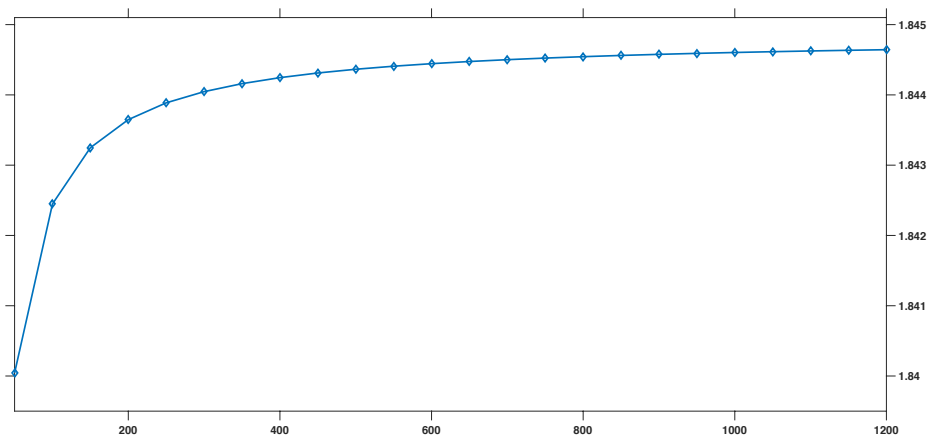


Figure: Optimal average cost $g^*_{k,\delta}$ of $\mathcal{M}_k$.

To check empirically the order of convergence of $g^*_{k,\delta}$ w.r.t. k, we have performed a linear regression analysis of the form $g^*_{k,\delta} \sim \beta_0 + \beta_1 \frac{1}{k}$.
We have obtained $\hat{\beta}_0 = 1.8448$ and $\hat{\beta}_1 = -0.0048$ with residuals satisfying

$$\max_k \left| g^*_{k,\delta} - \hat{\beta}_0 - \hat{\beta}_1/k \right| \leq 4 \cdot 10^{-6}.$$

- The approximations $g^*_{k,\delta}$ and the regression line $\hat{\beta}_0 + \hat{\beta}_1/k$ almost overlap.
- The non asymptotic bound is tight (its order is indeed attained).

Thank You