# Statistical inference for structured models

Part III: Lepski's principle. Estimation in bifurcating models.

Marc Hoffmann, Université Paris-Dauphine PSL

Spring school SFB 1294, March 2019

# Today's program

- About nonparametric adaptive estimation
  - Lepski's principle: soft heuristics
  - The Goldenshluger-Lepski method without (too much) pain
- Estimation in bifurcating models
  - Age dependent model
  - Size dependent model
  - Estimation in arbitrary BMC models

# Setting

- Goal: estimate a probability distribution $g(t,a)dtda$ from a (IID) drawn

$$Z^N(ds,du) = \mathcal{Z}^N(ds,du) = N^{-1}\sum_{i=1}^{N}\delta_{(T_i,A_i)}(ds,du).$$

- Kernel estimator:

$$\widehat{g}_{\boldsymbol{h}}^N(t,a) = \int_0^T\int_{\mathbb{R}_+} K_{\boldsymbol{h}}(t-s,a-u)\mathcal{Z}^N(ds,du).$$

- We have established, if $g \in \mathcal{H}^{\alpha,\beta}$

$$\mathbb{E}\Big[\big(\widehat{g}_{\boldsymbol{h}^\star}^N(t,a) - g(t,a)\big)^2\Big] \lesssim \mathbb{B}_{\boldsymbol{h}}(g) + \mathbb{V}_{\boldsymbol{h}}^N$$
$$\approx \big(h_1^\alpha + h_2^\beta\big)^2 + \Big(\frac{1}{\sqrt{Nh_1h_2}}\Big)^2$$

# Lepski's principle for two hypotheses

- Simplification: $g(t,a) \equiv g(a) \in \mathcal{H}^{\alpha}$ with $\alpha \in \{\alpha_{\mathrm{min}}, \alpha_{\mathrm{max}}\}$, $\alpha_{\mathrm{min}} < \alpha_{\mathrm{max}}$.
- Let $h^N(\alpha) = \big(N(\log N)^{-1}\big)^{-1/(2\alpha+1)}$.
- Pivotal observable quantity:

$$\left|\widehat{g}_{h^N(\alpha_{\mathrm{min}})}(a) - \widehat{g}_{h^N(\alpha_{\mathrm{max}})}(a)\right| = \left|(K_{h^N(\alpha_{\mathrm{min}})} - K_{h^N(\alpha_{\mathrm{max}})}) \star \mathcal{Z}^N(a)\right|.$$

- To be compared with $N^{-\alpha_{\mathrm{min}}/(2\alpha_{\mathrm{min}}+1)}$.
- Presence of an extra logarithmic factor for the control of stochastic deviations $\rightsquigarrow$ ignored in a first approach.

# Lepski's principle heuristics

- Not a valid proof!
- If $\alpha = \alpha_{\min}$, with overwhelming probability (ignoring log terms)

$$K_{h^N(\alpha_{\min})} \star \mathcal{Z}^N(a) - g(a) \approx N^{-\alpha_{\min}/(2\alpha_{\min}+1)}$$

and

$$\begin{aligned} K_{h^N(\alpha_{\max})} \star \mathcal{Z}^N(a) - g(a) &\approx h^N(\alpha_{\max})^{\alpha_{\min}} + N^{-1/2} h^N(\alpha_{\max})^{-1/2} \\ &= N^{-\alpha_{\min}/(2\alpha_{\max}+1)} + N^{-\alpha_{\max}/(2\alpha_{\max}+1)} \\ &\approx N^{-\alpha_{\min}/(2\alpha_{\max}+1)} \\ &\gg N^{-\alpha_{\min}/(2\alpha_{\min}+1)}. \end{aligned}$$

- Summing-up, if $\alpha = \alpha_{\min}$

$$\boxed{\left| K_{h^N(\alpha_{\min})} \star \mathcal{Z}^N(a) - K_{h^N(\alpha_{\max})} \star \mathcal{Z}^N(a) \right| \gg N^{-\alpha_{\min}/(2\alpha_{\min}+1)}.}$$

# Lepski's principle heuristics

- Conversely, if $\alpha = \alpha_{\max}$, with overwhelming probability (ignoring log terms)

$$K_{h^N(\alpha_{\min})} \star \mathcal{Z}^N(a) - g(a) \approx h^N(\alpha_{\min})^{\alpha_{\max}} + N^{-1/2} h^N(\alpha_{\min})^{-1/2}$$
$$= N^{-\alpha_{\max}/(2\alpha_{\min}+1)} + N^{-\alpha_{\min}/(2\alpha_{\min}+1)}$$
$$\approx N^{-\alpha_{\min}/(2\alpha_{\min}+1)}$$

and

$$K_{h^N(\alpha_{\max})} \star \mathcal{Z}^N(a) - g(a) \approx N^{-\alpha_{\max}/(2\alpha_{\max}+1)} \ll N^{-\alpha_{\min}/(2\alpha_{\min}+1)}.$$

- Summing-up, if $\alpha = \alpha_{\max}$

$$\boxed{\left| K_{h^N(\alpha_{\min})} \star \mathcal{Z}^N(a) - K_{h^N(\alpha_{\max})} \star \mathcal{Z}^N(a) \right| \approx N^{-\alpha_{\min}/(2\alpha_{\min}+1)}.}$$

# Lepski's principle: recap

- $\widehat{g}_h^N(a) = K_h \star \mathcal{Z}^N(a)$.
- $\mathcal{H} = \Big\{ \big(\frac{N}{\log N}\big)^{-1/(2\alpha_{\min}+1)}, \big(\frac{N}{\log N}\big)^{-1/2(\alpha_{\max}+1} \Big\}$.
- Data driven bandwidth: $h_\star^N = h_\star^N(\mathcal{Z}^N)$ solution to

$$h_\star^N = \max\Big\{ h \in \mathcal{H}, \forall \eta \leq h, \big|\widehat{g}_h^N(a) - \widehat{g}_\eta^N(a)\big| \leq C\Big(\frac{\log N}{N\eta}\Big)^{1/2} \Big\}$$

- Final estimator: $\widehat{g}_{h_\star^N}^N(a)$ satisfies the estimate

$$\mathbb{E}\big[\big(\widehat{g}_{h_\star^N}^N(a) - g(a)\big)^2\big] \lesssim \begin{cases} \big(\frac{N}{\log N}\big)^{-\alpha_{\max}/2(\alpha_{\max}+1)} & \text{if} \quad g \in \mathcal{H}^{\alpha_{\max}} \\ \big(\frac{N}{\log N}\big)^{-\alpha_{\min}/2(\alpha_{\min}+1)} & \text{if} \quad g \in \mathcal{H}^{\alpha_{\min}} \end{cases}$$

- Smoothness adaptation over the scale $\mathcal{H}^\alpha$ for $\alpha \in \{\alpha_{\min}, \alpha_{\max}\}$.
- The risk bound inflation by a $\log N$ term is unavoidable.

# The Goldenshluger-Lepski method

- Modern formulation of Lepski's principle in terms of oracle inequalities.
- Again, we keep-up with the 1-dimensional case for simplicity.
- We look for $\widehat{h}^{\star} = \widehat{h}^{\star}(\mathcal{Z}^N)$ so that

$$\mathbb{E}\big[\big(\widehat{g}^N_{h^\star}(a) - g(a)\big)^2\big] \lesssim \inf_{h \in \mathcal{H}} \big(\mathbb{B}_h(g)^2 + \mathbb{V}^N_h\big).$$

# The GL method

- Auxiliary oversmoothed estimator

$$\widehat{g}_{h,\eta}(a) = N^{-1} \sum_{i=1}^{N} K_h \star K_\eta(x - A_i), \quad h, \eta \in \mathcal{H}.$$

- $\widehat{g}_{h,\eta}(a) = \widehat{g}_{\eta,h}(a)$.
- $\widehat{g}_{h,\eta}(a) = \widehat{g}_{h+\eta}(a)$ for $K(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$.
- GL principle: for fixed $h \in \mathcal{H}$, we let $\eta$ run through $\mathcal{H}$ and compare $\widehat{g}_h$ to $\widehat{g}_{h,\eta}$

# The GL method: the fundamental quantities

- Construction of the GL estimator

$$\mathfrak{B}_h(\eta) = \left\{ \left| \widehat{g}_\eta(a) - \widehat{g}_{h,\eta}(a) \right| - \chi(\eta) \right\}_+ ,$$

$$\widehat{h} = \text{Argmin}_{h \in \mathcal{H}} \left( \mathfrak{B}_h + \chi(h) \right), \quad \mathfrak{B}_h = \max_{\eta \in \mathcal{H}} \mathfrak{B}_h(\eta),$$

$$\text{GL-estimator} : \widehat{g}_{\widehat{h}}(a).$$

- Handwaving heuristics:
  - $\chi(\eta) \to \infty$ as $\eta \to 0$ appropriate random fluctuation control threshold.
  - $\mathfrak{B}_h(\eta)$ is computable and hopefully close to its expectation in a certain sense
  - Picking $\widehat{h}$ amounts to take something like

$$\widehat{h} \approx \max \left\{ h \in \mathcal{H}, \; \forall \eta \leq h : \left| \widehat{g}_\eta(a) - \widehat{g}_{h,\eta}(a) \right| \lesssim \left( \mathrm{Var}(\widehat{g}_\eta) \right)^{1/2} \right\}.$$

# Soft proof of the GL method

- Step 1: for every $h \in \mathcal{H}$:

$$\boxed{\left|\widehat{g}_{\widehat{h}}(a) - g(a)\right| \leq 2\big(\mathfrak{B}_h + \chi(h)\big) + \left|\widehat{g}_h(a) - g(a)\right| \quad (\star)}$$

- Step 2: Fundamental control of $\mathbb{E}[|\mathfrak{B}_h|^2]$:

$$\begin{aligned}\mathfrak{B}_h = \max_{\eta\in\mathcal{H}} \mathfrak{B}_h(\eta) &= \max_{\eta\in\mathcal{H}} \big\{|\widehat{g}_\eta(a) - \widehat{g}_{h,\eta}| + \chi(\eta)\big\}_+ \\ &\leq \max_{\eta\in\mathcal{H}} \big\{\zeta_\eta - \chi_1(\eta)\big\}_+ + \max_{\eta\in\mathcal{H}} \big\{\zeta_{h,\eta} - \chi_2(\eta)\big\}_+ \\ &\quad + \max_{\eta\in\mathcal{H}} \left|\mathbb{E}[\widehat{g}_\eta(a)] - \mathbb{E}[\widehat{g}_{h,\eta}(a)]\right|.\end{aligned}$$

- $\chi = \chi_1 + \chi_2$, $\zeta_{(h),\eta} = \widehat{g}_{(h),\eta}(a) - \mathbb{E}[\widehat{g}_{(h),\eta}(a)]$.
- Last term:
  $\max_{\eta\in\mathcal{H}} |K_\eta \star g - K_\eta \star K_h \star g|_\infty \lesssim |K|_1 |g - K_h \star g|_\infty$.
- First two stochastic terms: concentration inequalities.

# Soft proof of the GL method

- Concentration inequality

Proposition (Benett, Bernstein)

*$-b \leq Y_i \leq b$ independent r.v. such that $\sum_{i=1}^N \mathbb{E}[Z_i^2] \leq v$. With $\lambda(u) = \sqrt{2vu} + \frac{2}{3}bu$, we have*

$$\mathbb{P}\Big(\sum_{i=1}^N Z_i - \mathbb{E}[Z_i] \geq \lambda(u)\Big) \leq \exp(-u).$$

- Applied to $Y_i = N^{-1}K_\eta(x - A_i)$ or $K_h \star K_\eta(x - A_i)$ yields appropriate $\lambda_{\eta\,(\text{resp.}h)}(u) = \lambda_{N,\eta\,(\text{resp. }h),K,|g|_\infty}(u)$.
- Set finally $\chi_i(\eta) = \lambda(\gamma|\log \eta|)$, $i = 1, 2$, $\gamma > 0$ to be specified.
- The first two stochastic terms are of order $N^{-1}\sum_{\eta \in \mathcal{H}} \eta^{\gamma-1}$.

# Soft proof of the GL method

- We piece all the estimates together, take $\mathbb{E}[(\cdot)^2]$ and $\min_h$:

$$\mathbb{E}\Big[\big(\widehat{g}_{\widehat{h}}(a) - g(a)\big)^2\Big]$$
$$\lesssim \min_{h \in \mathcal{H}} \Big[\mathbb{E}[(\widehat{g}_h(a) - g(a))^2] + \frac{|\log h|}{Nh} +$$
$$+ |K_h \star g - g|_\infty^2\Big] + N^{-1} \sum_{\eta \in \mathcal{H}} \eta^{\gamma - 1}.$$

- Choose $\mathcal{H}$ sufficiently rich to approximate $h_N(\alpha) = N^{-1/(2\alpha+1)}$ while $N^{-1} \sum_{h \in \mathcal{H}} \eta^{\gamma-1} \lesssim$ minimax rate.
- It remains to prove $(\star)$...

# Soft proof of the GL method

- Completely deterministic argument:

$$|\widehat{g}_{\widehat{h}}(a) - g(a)| \leq \left\{|\widehat{g}_{\widehat{h}}(a) - \widehat{g}_{h,\widehat{h}}| - \chi(\widehat{h})\right\}_+ + \chi(\widehat{h})$$
$$+ \left\{|\widehat{g}_{\widehat{h},h} - \widehat{g}_h(a)| - \chi(h)\right\}_+ + \chi(h)$$
$$+ |\widehat{g}_h(a) - g(a)|.$$

- First term in the RHS: $\mathfrak{B}_h(\widehat{h}) + \chi(\widehat{h}) \leq \max_{\eta\in\mathcal{H}} \mathfrak{B}_h(\eta) + \chi(\widehat{h})$
- Second term in the RHS similar: $\leq \max_{\eta\in\mathcal{H}} \mathfrak{B}_{\widehat{h}}(\eta) + \chi(h)$.
- Adding and regrouping, we obtain

$$\mathfrak{B}_{\widehat{h}} + \chi(\widehat{h}) + \mathfrak{B}_h + \chi(h) \leq 2(\mathfrak{B}_h + \chi(h))$$

by construction of $\widehat{h}$.
- $\left|\widehat{g}_{\widehat{h}}(a) - g(a)\right| \leq 2\big(\mathfrak{B}_h + \chi(h)\big) + \left|\widehat{g}_h(a) - g(a)\right|$ $(\star)$ follows.

# Estimation in bifurcating models

- We turn back to our PDE related stochastic models.
- We start with growth-fragmentation models for the simplest observation scheme: we observe

$$\mathcal{Z}^N = \{(\zeta_u, \xi_u), u \in \mathbb{U}_n^{\varrho}\},$$

  where
  - $(\zeta_u, \xi_u) =$ (life length, size at birth) of the individual $u$.
  - $\mathbb{U}_n^{\varrho}$ is a $\varrho$-regular tree of size $N \approx 2^{n\varrho}$.
- The underlying stochastic tools are Markov chains on trees.
- For age-dependent division, the statistical model has a particularily simple structure.

# Growth-fragmentation: the age dependent model

- The associated deterministic model is

$$\begin{cases} \partial_t g(t,a) + \partial_a g(t,a) + B(a) g(t,a) = 0 \\ g(0,a) = g_0(a), \;\; g(t,0) = 2\int_0^\infty B(a) g(t,a) da. \end{cases}$$

- We are interested in recovering $a \mapsto B(a)$ from data

$$\mathcal{Z}^N = \big\{(\zeta_u, \xi_u), u \in \mathbb{U}_n^\varrho\big\}.$$

- The data $(\xi_u)_{u \in \mathbb{U}_n^\varrho}$ are irrelevant here and we discard them.
- The data $(\zeta_u)_{u \in \mathbb{U}_n^\varrho}$ are independent and identically distributed with common density

$$\mathbb{P}(\zeta_u \in da) = B(a) \exp\Big(-\int_0^a B(u) du\Big) da$$

# Growth-fragmentation: the age dependent model

- The formula can be inverted: if $f_B(a)da = \mathbb{P}(\zeta_u \in da)$, we also have

$$\boxed{B(a) = \frac{f_B(a)}{1 - \int_0^a f_B(u)du}}$$

provided $\int^\infty B = \infty$, an assumption in force from now on.

- Let $N = |\mathbb{U}_n^\varrho| \approx 2^{n\varrho}$. Let

$$\boxed{\widehat{B}_h^N(a) = \frac{N^{-1}\sum_{u\in\mathbb{U}_n^\varrho} K_h(a-\zeta_u)}{\max(N^{-1}\sum_{u\in\mathbb{U}_n^\varrho} \mathbf{1}_{\{\zeta_u\geq a\}}, \varpi_N)}}$$

for some (technical) threshold $\varpi_N \to 0$.

- Numerator eligible to data-driven bandwidth selection according to Lepski's principle $h \rightsquigarrow h_\star^N$.
- Denominator converges to $1 - \int_0^a f_B(u)du$ at rate $N^{-1/2}$ strongly.

# Growth-fragmentation: the age dependent model

- $(H1)$ $\mathcal{B}$ consists of (uniformly) bounded functions such that $\int^{\infty} B = \infty$.

**Theorem**

*Under $(H1)$, for $0 < \alpha_{\min} < \alpha_{max}$, there exists a choice of $\mathcal{H}$ such that*

1. *The GL bandwidth $h^N_\star$ satisfies*

$$\boxed{\mathbb{E}\Big[\big(\widehat{B}^N_{h^N_\star}(a) - B(a)\big)^2\Big] \lesssim \inf_{h\in\mathcal{H}} \big(\mathbb{B}_h(f_B) + \mathbb{V}^N_h\big) + N^{-1}}.$$

2. *Moreover, for every $\alpha \in [\alpha_{\min}, \alpha_{\max}]$:*

$$\boxed{\sup_{B\in\mathcal{B}\cap\mathcal{H}^\alpha} \mathbb{E}\Big[\big(\widehat{B}^N_{h^N_\star}(a) - B(a)\big)^2\Big] \lesssim \Big(\frac{\log N}{N}\Big)^{2\alpha/(2\alpha+1)}}$$

*where $\mathcal{H}^\alpha$ is a (locally around $a$) Hölder ball.*

- *The result is minimax adaptive optimal.*

# Growth-fragmentation: the size dependent model

- We start with a singe cell of size $x_0$.
- For simplicity, the cell grows exponentially according to a constant rate $\tau > 0$:

$$\frac{dX(t)}{dt} = \kappa\big(X(t)\big)dt = \tau X(t)dt.$$

- The mother cell gives rize to two children, at a size dependent rate $x \mapsto B(x)$.
- The two children have initial size $x_1/2$, where $x_1$ is the size of the mother at division.
- They grow independently according to the rate $\tau$ and divide according to the rate $B(x)$.

# Growth-fragmentation: the size dependent model

- We observe
$$\mathcal{Z}^N = \{(\zeta_u, \xi_u), u \in \mathbb{U}_n^{\varrho}\},$$
where
  - $(\zeta_u, \xi_u) =$ (life length, size at birth) of the individual $u$.
  - $\mathbb{U}_n^{\varrho}$ is a $\varrho$-regular tree of size $N \approx 2^{n\varrho}$.
- We look for an analog of the inversion formula $\mathbb{P}(\zeta_u \in da) \leftrightarrow B(a)$ obtained in the age-dependent model.
- The $\xi_u$ and the $\zeta_u$ are not independent – not identically distributed – anymore!
- They however form a Markov chain along branches of the genealogical tree $\mapsto$ bifurcating Markov chain.

# Growth-fragmentation: the size dependent model

- If $u^-$ denotes the parent of $u$, we have

$$\boxed{2\xi_u = \xi_{u^-} \exp\big(\tau\zeta_{u^-}\big).}$$

- $\tau$ is identified via the observation of a single $(\zeta_{u^-}, \xi_{u^-}, \xi_u)$.
- We have

$$\mathbb{P}(\zeta_u \in [t, t+dt] \,|\, \zeta_u \geq t, \xi_u = x) = B(xe^{\tau t})dt$$

that entails the density of the lifetime $\zeta_{u^-}$ conditional on $\xi_{u^-} = x$:

$$\boxed{t \mapsto B(xe^{\tau t}) \exp\Big(-\int_0^t B(xe^{\tau s})ds\Big).}$$

# Growth-fragmentation: the size dependent model

- We can derive a simple and explicit representation for the transition kernel $K_B(x, dx')$ of the underlying Markov chain:

$$K_B(x,x')dx' = \mathbb{P}\big(\xi_u \in dx' \,\big|\, \xi_{u^-} = x\big)$$
$$= \frac{B(2x')}{\tau x'}\mathbf{1}_{\{x' \geq x/2\}} \exp\Big(-\int_{x/2}^{x'} \tfrac{B(2s)}{\tau s} ds\Big) dx'.$$

- The inversion formula is obtained by looking at the equation

$$\int_{x\in\mathbb{R}_+} \nu_B(dx) K_B(x,x')dx' = \nu_B(dx')$$

that charaterises the invariant probability measures $\nu_B(dx) = \nu_B(x)dx$ of $K_B$.

# Growth-fragmentation: the size dependent model

- Expand the invariant measure equation $\nu_B K_B = \nu_B$

$$\nu_B(x') = \int_0^\infty \nu_B(x) K_B(x, x') dx$$
$$= \frac{B(2x')}{\tau x'} \int_0^{2x'} \nu_B(x) \exp\Big( - \int_{x/2}^{x'} \frac{B(2s)}{\tau s} ds \Big) dx$$
$$= \frac{B(2x')}{\tau x'} \int_0^\infty \int_0^\infty \mathbf{1}_{\{x \leq 2x',\, s \geq x'\}} \nu_B(x)\, K_B(x, s)\, ds dx.$$

- This yields the key representation

$$\boxed{\nu_B(x) = \frac{B(2x)}{\tau x} \mathbb{P}_{\nu_B}\big(\xi_{u^-} \leq 2x,\ \xi_u \geq x\big)}$$

with $\mathbb{P}_{\nu_B} = \int_0^\infty \nu_B(dx) \mathbb{P}\big( \cdot \mid \xi_\emptyset = x\big)$.

# Growth-fragmentation: the size dependent model

- We obtain the representation formula

$$B(x) = \frac{\tau x}{2} \frac{\nu_B(x/2)}{\mathbb{P}_{\nu_B}\big(\xi_{u^-} \leq x,\ \xi_u \geq x/2\big)}.$$

- But! We always have $\{\xi_{u^-} \geq x\} \subset \{\xi_u \geq x/2\}$, hence

$$\begin{aligned}\mathbb{P}_{\nu_B}\big(\xi_{u^-} \leq x, \xi_u \geq x/2\big) &= \mathbb{P}_{\nu_B}\big(\xi_u \geq x/2\big) - \mathbb{P}_{\nu_B}\big(\xi_{u^-} \geq x\big) \\ &= \int_{x/2}^{\infty} - \int_{x}^{\infty} \\ &= \int_{x/2}^{x} \nu_B(u)du.\end{aligned}$$

- Remark: the general inversion formula still allows for some room (if $\tau = \tau_u$ is tree-dependent and random for instance)

# Growth-fragmentation: the size dependent model

- In turn, we obtain the final representation

$$\boxed{B(x) = \frac{\tau x}{2}\frac{\nu_B(x/2)}{\int_{x/2}^{x}\nu_B(u)du}}$$

- This yields the kernel-based estimator

$$\boxed{\widehat{B}_h^N(x) = \frac{\tau x}{2}\frac{N^{-1}\sum_{u\in\mathbb{U}_n^{\varrho}} K_h(\xi_u - x/2)}{\max(N^{-1}\sum_{u\in\mathbb{U}_n^{\varrho}}\mathbf{1}_{\{\xi_{u^-}\leq x,\, \xi_u \geq x/2\}}, \varpi_N)}}$$

  for some (technical) threshold $\varpi_N \to 0$.

- The study of the convergence of empirical means is more involved.

# Convergence of empirical means

- **Notation**: $K_B^m\varphi(x) = K_B(K_B^{m-1}\varphi)(x)$ with

$$K_B\varphi(x) = \int_0^\infty \varphi(x')K_B(x,x')dx' = \mathbb{E}\big[\varphi(\xi_u)\,|\,\xi_{u^-} = x\big].$$

- $(H2)$ $\inf_{B\in\mathcal{B}}\inf_x B(x) > 0$.

**Proposition**

*Under $(H1),(H2)$, the invariant probability $\nu_B$ is well defined and there exists $\rho_B < 1$ such that for $\mathbb{V}(x) = 1 + x^2$, we have*

$$\boxed{\sup_{|\varphi|\leq\mathbb{V}}\big|K_B^m\varphi(x) - \langle\varphi,\nu_B\rangle\big| \lesssim \rho_B^m\,\mathbb{V}(x).}$$

# Convergence of empirical means

- Result uniform in $B \in \mathcal{B}$ and $\tau$ over compact sets of $(0, \infty)$.
- Proof: classical, relies on the existence of a Lyapunov function $\mathbb{V}(x) \geq 1$ s.t.

$$K_B \mathbb{V}(x) \leq \lambda \mathbb{V}(x) + C \quad \text{and} \quad \inf_{|x| \leq C} K(x, dx') \geq \lambda \mu(dx')$$

  for some $0 < \lambda < 1$, $C > 0$ and a probability measure $\mu$.
- Enables one to control covariance terms:

$$\mathbb{E}\big[\varphi(\xi_u)\varphi(\xi_v)\big] = \mathbb{E}\big[K_B^{|u|-|u\wedge v|}\varphi(X_{u\wedge v})K_B^{|v|-|u\wedge v|}\varphi(X_{u\wedge v})\big],$$

  $u \wedge v =$ most recent common ancestor between $u$ and $v$.

# Convergence of empirical means

- Two difficulties:
  1. Order of the covariance terms in terms of $\varphi \rightsquigarrow$ usually needs a control in $|\cdot|_2$-norm.
  2. Competition between growth of the binary tree (geometric rate $= 2$) and decorrelation (geometric rate $= \rho_B$).
- Answer 1: Assume for simplicity that $\mathbb{E}[\varphi(\xi_u)] = \mathbb{E}[\varphi(\xi_u)]$ and $|u| \leq |v|$. The last term is bounded above by

$$\mathbb{E}\big[\varphi(\xi_u)\varphi(\xi_v)\big] \lesssim \min\big(\rho_B^{d(u,v)}|\varphi|_\infty^2, \rho_B^{|v|-|u\wedge v|}|\varphi|_\infty|\varphi|_1\big),$$

  $d(u,v) =$ graph distance between $u$ and $v$.
- Answer 2: Sufficient condition: $\rho_B < \frac{1}{2}$.

# Convergence of empirical means

- $(H3)$ We have $\sup_{B\in\mathcal{B}} \rho_B < \frac{1}{2}$.
- Let $\mathcal{M}_{\mathbb{U}_n^\rho}(\varphi) = N^{-1}\sum_{u\in\mathbb{U}_n^\rho}\varphi(\xi_u)$.

## Proposition

*Under $(H_1), (H2), (H3)$, for any initial condition $\mu$, we have*

$$\boxed{\mathbb{E}_\mu\big[\big(\mathcal{M}_{\mathbb{U}_n^\rho}(\varphi) - \langle\varphi,\nu_B\rangle\big)^2\big] \lesssim N^{-1}\big(|\varphi|^2_{L^2(\mu+\nu_B)} + (1+|\mathbb{V}|^2_{L^2(\mu)})|\varphi|_\infty|\varphi|_{\nu_B}}$$

*uniformly in $\mathcal{B}$.*

- This results holds in wider generality for bifurcating Markov chains:
  - Arbitrary deterministic flows between jumps.
  - Random flows (diffusions) between jumps.
  - Test functions on forks: $\varphi(\xi_u) \rightsquigarrow \psi(\xi_u, \xi_{u0}, \xi_{u1})$.

# Nonparametric estimation of $B(x)$

- With the specification $h^N = N^{-1/(2\alpha+1)}$, the variance bound is sufficient to obtain

$$\boxed{\sup_{B \in \mathcal{B} \cap \mathcal{H}^\alpha} \mathbb{E}_\mu\Big[\big(\widehat{B}^N_{h^N}(x) - B(x)\big)^2\Big] \lesssim \varpi_N^{-2} N^{-2\alpha/(2/\alpha+1)}}$$

  for any $\mu(dx') \ll dx'$ locally around $x$.
- The rate is minimax nearly-optimal but non-adaptive!
- In order to extend the result to adaptation, we need concentration properties.
- We need a stringent restriction: uniform geometric ergodicity.

# Uniform geometric ergodicity

- The kernel $K$ is uniformly geometrically ergodic if

$$\left|K_B^m \varphi(x) - \langle \varphi, \nu_B \rangle\right| \lesssim |\varphi|_\infty \rho_B^m.$$

- This amounts to have a bounded Lyapunov function $\mathbb{V}$.
- We have a sufficient (but slightly artificial) condition that implies uniform geometric ergodicity and $(H3)$:
- $(H2')$ $B : (b_{\min}, b_{\max}) \to \mathbb{R}_+$ with $2b_{\min} < b_{\max}$ and

$$\int^{b_{\max}} u^{-1} B(u) du = \infty, \quad \int_{b_{\min}} u^{-1} B(u) du \lesssim 1.$$

- $(H1')$ $\mathcal{B}$ contains continuous and locally bounded functions with appropriate uniformity conditions.

# Concentration properties

- Let $\Sigma_n(\varphi) = |\varphi|_2^2 + \min_{1 \le \ell \le n-1} \big( |\varphi|_1^2 2^{\ell} + |\varphi|_\infty^2 2^{-\ell} \big)$

**Theorem**

*Work under $(H1'), (H2'), (H3)$ and $(H4)$. For $\delta \gtrsim N^{-1}|\varphi|_\infty$, we have*

$$\mathbb{P}\big(\mathcal{M}_{\mathbb{U}_n^\rho}(\varphi) - \langle \varphi, \nu_B \rangle \ge \delta\big) \le \exp\Big( - C_B \frac{N\delta^2}{\Sigma_n(\varphi) + |\varphi|_\infty \delta} \Big)$$

*with* $\sup_{B \in \mathcal{B}} C_B < \infty$.

- The result extends to
  - More general BMC models (under uniform geometric ergodicity).
  - Test functions on forks: $\varphi(\xi_u) \rightsquigarrow \psi(\xi_u, \xi_{u0}, \xi_{u1})$.

# Adaptive estimation

- Theorem
  *Under (H1′), (H2′), (H3) and (H4), for* $0 < \alpha_{\min} < \alpha_{max}$*, there exists a choice of* $\mathcal{H}$ *and a specification of* $\mathbb{V}_h^N$ *such that*
  1. *The GL bandwidth* $h_\star^N$ *satisfies*

$$\mathbb{E}\big[\big(\widehat{B}^N_{h^N_\star}(a) - B(a)\big)^2\big] \lesssim \inf_{h\in\mathcal{H}} \big(\mathbb{B}_h(\nu_B) + \mathbb{V}_h^N\big) + N^{-1}.$$

  2. *Moreover, for every* $\alpha \in [\alpha_{\min}, \alpha_{\max}]$*:*

$$\sup_{B\in\mathcal{B}\cap\mathcal{H}^\alpha} \mathbb{E}\big[\big(\widehat{B}^N_{h^N_\star}(a) - B(a)\big)^2\big] \lesssim \Big(\frac{\log N}{N}\Big)^{2\alpha/(2\alpha+1)}$$

  *where* $\mathcal{H}^\alpha$ *is a (locally around a) Hölder ball.*

- The result is minimax adaptive optimal.
- Remaining open question: extension to non uniformly geometrically ergodic Markov kernels.

# Supplementary material

- We numerically illustrate the performances of the previous estimator
- The numerics is based on another approximation scheme, by wavelet kernel projection estimators
- The algorithm differs, but the theory is the same.
- We further elaborate on arbitrary Binary Markov Chains models.

# Numerical illustration

- We consider a perturbation of the baseline splitting rate $\widetilde{B}(x) = x/(5-x)$ over the range $x \in \mathcal{S} = (0,5)$ of the form

$$\boxed{B(x) = \widetilde{B}(x) + \mathfrak{c}\, T\big(2^j(x - \tfrac{7}{2})\big)}$$

  with $(\mathfrak{c}, j) = (3,1)$ or $(\mathfrak{c}, j) = (9,4)$, and where $T(x) = (1+x)\mathbf{1}_{\{-1 \le x < 0\}} + (1-x)\mathbf{1}_{\{0 \le x \le 1\}}$ is a tent shaped function.
- The trial splitting rate with parameter $(\mathfrak{c}, j) = (9,4)$ is more localized around $7/2$ and higher than the one associated with parameter $(\mathfrak{c}, j) = (3,1)$.
- For a given $B$, we simulate $M = 100$ Monte Carlo trees up to the generation $n = 15$ with $\tau = 2$.

# Numerical illustration

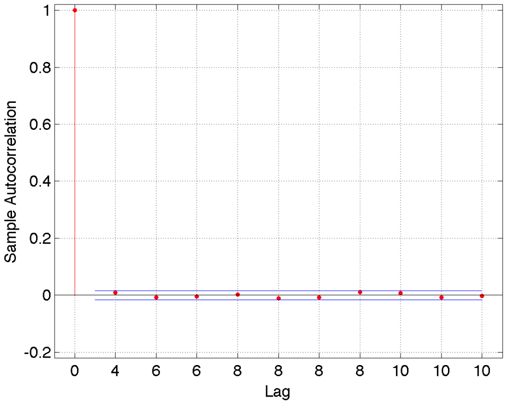


Figure: *Sample autocorrelation of ordered* $(\xi_{u0}, |u| = n-1)$ *for* $n = 15$. *Note: due to the binary tree structure the lags are* $\{4, 6, 6, \ldots\}$. *As expected, we observe a fast decorrelation.*

# Numerical illustration

- Here, we implement an alternative adaptive procedure via a projection estimator

$$K_h \star B(x) \rightsquigarrow \int K_h(x,y)B(y)dy$$

with

$$K_h(x,y) = \sum_k \varphi_{h,k}(x)\varphi_{h,k}(y),$$

where the $\varphi_{h,k}(x) = h^{-1/2}\varphi(h^{-1}x - k)$ (on a dyadic scale $h^{-1} = 2^j$) generate a regular multiresolution analysis associated to a scaling function $\varphi$.

- The adaptve bandwidth is replaced here by wavelet thresholding, taking advantage of the multiresolution structure.
- The underlying theory is close and the required probabilistic properties of the models tools are the same!

# Numerical illustration

- We implement the estimator $\widehat{B}_N$ using the Matlab wavelet toolbox.
- We use compactly supported Daubechies wavelets of order 8 up to maximal level $J := \frac{1}{2}\log_2(N/\log N)$.
- We choose the threshold proportional to $\sqrt{\log|\mathbb{T}_n|/|\mathbb{T}_n|}$, $\mathbb{T}_n =$ the whole tree up to generation $n$.
- We calibrate the constant to 10 or 15 for two trial splitting rates (mainly by visual inspection).
- We evaluate $\widehat{B}_n$ on a regular grid over $[1.5, 4.8]$ with mesh $\Delta x = N^{-1/2}$.

# Numerical illustration

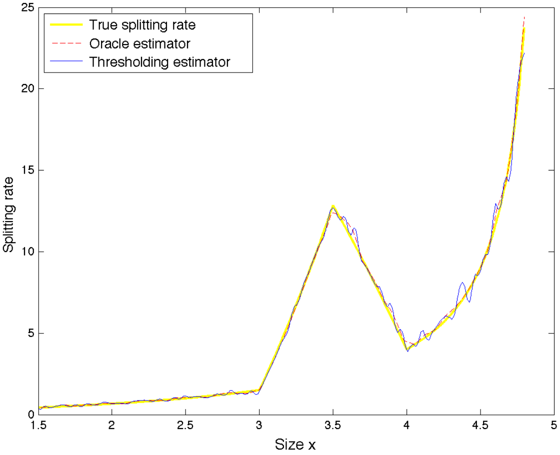


Figure: *Large spike: reconstruction of the trial splitting rate $B$ specified by $(\mathfrak{c}, j) = (3, 1)$ over $[1.5, 4.8]$ based on one sample $(\xi_u, u \in \mathbb{T}_n)$ for $n = 15$ (i.e. $\frac{1}{2}|\mathbb{T}_n| = 32\ 767$).*

# Numerical illustration

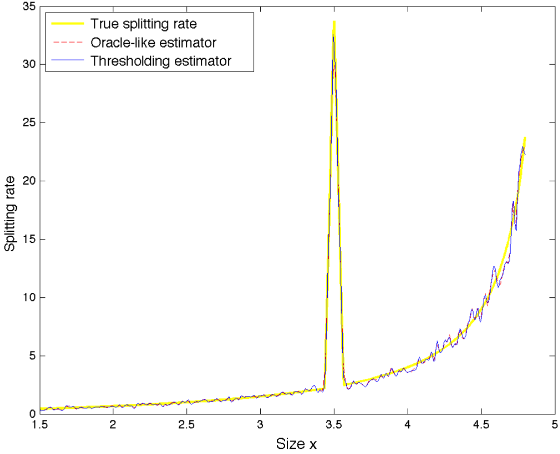


Figure: *High spike: reconstruction of the trial splitting rate $B$ specified by $(\mathfrak{c}, j) = (9, 4)$ over $\mathcal{D} = [1.5, 4.8]$ based on one sample $(\xi_u, u \in \mathbb{T}_n)$ for $n = 15$ (i.e. $\frac{1}{2}|\mathbb{T}_n| = 32\ 767$).*

# Estimation in arbitrary BMC models

- We review some generic results for nonparametric estimation in arbitrary BMC models.
- We slightly depart from the previous appproach, but the methodology is essentially the same.

### Definition

A bifurcating Markov chain is a family $(X_u)_{u\in\mathbb{T}}$ of random variables with value in $(\mathcal{S},\mathfrak{S})$ such that $X_u$ is $\mathcal{F}_{|u|}$-measurable for every $u\in\mathbb{T}$ and

$$\mathbb{E}\big[\prod_{u\in\mathbb{G}_m} g_u(X_u,X_{u0},X_{u1})\big|\mathcal{F}_m\big] = \prod_{u\in\mathbb{G}_m}\mathcal{P}g_u(X_u)$$

for every $m\geq 0$ and $(g_u)_{u\in\mathbb{G}_m}$, where
$\mathcal{P}g(x)=\int_{\mathcal{S}\times\mathcal{S}} g(x,y,z)\mathcal{P}(x,dy\,dz)$

# Estimation in arbitrary BMC models

- We consider a BMC $(X_u, u \in \mathbb{T})$ that we observe on $\mathbb{T}_n$, with

$$\mathbb{T} = \bigcup_{m \in \mathbb{N}} \mathbb{G}_m, \;\; \mathbb{G}_m = \{0,1\}^m, \;\; (\mathbb{G}_0 = \emptyset).$$

- We thus have a regular tree with $\varrho = 1$ and $N = 2^{n+1} - 1$.
- Several objects of interest:
  - The transition of the tagged-branch chain or mean transition.
  - The transition of the BMC itself.
  - The invariant (probability) measure of the mean transition.

# The tagged-branch chain

- The tagged-branch chain $(Y_m)_{m\geq 0}$: $Y_0 = X_\emptyset$ and for $m \geq 1$,

$$Y_m = X_{\emptyset\epsilon_1\cdots\epsilon_m},$$

$(\epsilon_m)_{m\geq 1}$ IID Bernoulli with parameter $1/2$, independent of $(X_u)_{u\in\mathbb{T}}$.

- Transition (mean transition)

$$\mathcal{Q} = (\mathcal{P}_0 + \mathcal{P}_1)/2,$$

obtained from the marginals $\mathcal{P}_0(x, dy) = \int_{z\in\mathcal{S}} \mathcal{P}(x, dy\, dz)$ and $\mathcal{P}_1(x, dz) = \int_{y\in\mathcal{S}} \mathcal{P}(x, dy\, dz)$.

# Digest

- Guyon (2007) proves that if $(Y_m)_{m\geq 0}$ is ergodic with invariant measure $\nu$, then

$$\frac{1}{|\mathbb{G}_n|}\sum_{u\in\mathbb{G}_n} g(X_u) \to \int_{\mathcal{S}} g(x)\nu(dx)$$

  holds almost-surely as $n \to \infty$.

- We also have

$$\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} g(X_u, X_{u0}, X_{u1}) \to \int_{\mathcal{S}} \mathcal{P}g(x)\nu(dx)$$

  almost-surely as $n \to \infty$.

- These results are appended with central limit theorems.

# Toward statistical inference

- $\mathcal{D} \subseteq \mathcal{S}$ that will be later needed for statistical purposes.
- Mean transition $\mathcal{Q} = \frac{1}{2}(\mathcal{P}_0 + \mathcal{P}_1)$.

# Assumptions

- **Assumption (D)** The family $\{\mathcal{Q}(x, dy), x \in \mathcal{S}\}$ is dominated:

$$\mathcal{Q}(x, dy) = \mathcal{Q}(x, y)\mathfrak{n}(dy) \;\text{ for every }\; x \in \mathcal{S},$$

for some $\mathcal{Q} : \mathcal{S}^2 \to [0, \infty)$ such that

$$|\mathcal{Q}|_{\mathcal{D}} = \sup_{x \in \mathcal{S}, y \in \mathcal{D}} \mathcal{Q}(x, y) < \infty.$$

- **Assumption (UE)** $\mathcal{Q}$ admits a unique invariant probability measure $\nu$ and there exist $R > 0$ and $0 < \rho < 1/2$ such that

$$\left|\mathcal{Q}^m g(x) - \nu(g)\right| \le R|g|_\infty \, \rho^m, \quad x \in \mathcal{S}, \quad m \ge 0,$$

# Variance definitions

- For $g : \mathcal{S}^d \to \mathbb{R}$, define $\Sigma_{1,1}(g) = |g|_2^2$ and for $n \geq 2$,

$$\Sigma_{1,n}(g) = |g|_2^2 + \min_{1 \leq \ell \leq n-1} \left( |g|_1^2 2^{\ell} + |g|_\infty^2 2^{-\ell} \right). \tag{1}$$

- Define also $\Sigma_{2,1}(g) = |\mathcal{P} g^2|_1$ and for $n \geq 2$,

$$\Sigma_{2,n}(g) = |\mathcal{P} g^2|_1 + \min_{1 \leq \ell \leq n-1} \left( |\mathcal{P} g|_1^2 2^{\ell} + |\mathcal{P} g|_\infty^2 2^{-\ell} \right). \tag{2}$$

# One-step deviations

**Theorem**

*Under* **(D)** *and* **(UE)**, *for every $n \geq 1$:*

**(i)** *For any $\delta > 0$ such that $\delta \geq 4R|g|_\infty |\mathbb{G}_n|^{-1}$, we have*

$$\mathbb{P}\Big(\frac{1}{|\mathbb{G}_n|} \sum_{u \in \mathbb{G}_n} g(X_u) - \nu(g) \geq \delta\Big) \leq \exp\left(\frac{-|\mathbb{G}_n|\delta^2}{\kappa_1 \Sigma_{1,n}(g) + \kappa_2 |g|_\infty \delta}\right).$$

**(ii)** *For any $\delta > 0$ such that $\delta \geq 4R(1-2\rho)^{-1}|g|_\infty |\mathbb{T}_n|^{-1}$, we have*

$$\mathbb{P}\Big(\frac{1}{|\mathbb{T}_n|} \sum_{u \in \mathbb{T}_n} g(X_u) - \nu(g) \geq \delta\Big) \leq \exp\left(\frac{-|\mathbb{T}_n|\delta^2}{\kappa_3 \Sigma_{1,n}(g) + \kappa_4 |g|_\infty \delta}\right).$$

# Two-steps deviations

## Theorem

*Under* **(D)** *and* **(UE)**, *for every $n \geq 2$:*

**(i)** *For any $\delta > 0$ such that $\delta \geq 4R|\mathcal{P}g|_\infty |\mathbb{G}_n|^{-1}$, we have*

$$\mathbb{P}\Big(\frac{1}{|\mathbb{G}_n|}\sum_{u\in\mathbb{G}_n} g(X_u, X_{u0}, X_{u1}) - \nu(\mathcal{P}g) \geq \delta\Big) \leq \exp\Big(\frac{-|\mathbb{G}_n|\delta^2}{\kappa_1 \Sigma_{2,n}(g) + \kappa_2 |g|_\infty \delta}\Big)$$

**(ii)** *For any $\delta > 0$ such that $\delta \geq 4(nR|\mathcal{P}g|_\infty + |g|_\infty)|\mathbb{T}_{n-1}|^{-1}$, we have*

$$\mathbb{P}\Big(\frac{1}{|\mathbb{T}_{n-1}|}\sum_{u\in\mathbb{T}_{n-1}} g(X_u, X_{u0}, X_{u1}) - \nu(\mathcal{P}g) \geq \delta\Big)$$

$$\leq \exp\Big(\frac{-n^{-1}|\mathbb{T}_{n-1}|\delta^2}{\kappa_1 \Sigma_{2,n-1}(g) + \kappa_2 |g|_\infty \delta}\Big).$$

# Statistical inference

- From now on $(\mathcal{S}, \mathfrak{S}) = \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ and $\mathcal{D} \subset \mathcal{S}$ compact interval
- **Assumption (S)** The family $\{\mathcal{P}(x, dy\,dz), x \in \mathcal{S}\}$ is dominated w.r.t. the Lebesgue measure:

$$\mathcal{P}(x, dy\,dz) = \mathcal{P}(x, y, z)dy\,dz \;\text{ for every }\; x \in \mathcal{S}$$

for some $\mathcal{P} : \mathcal{S}^3 \to [0, \infty)$ such that

$$|\mathcal{P}|_{\mathcal{D}} = \sup_{(x,y,z) \in \mathcal{D}^3} |\mathcal{P}(x, y, z)| < \infty.$$

# Statistical inference (cont.)

- For some $n \geq 1$, we observe $(X_u)_{u \in \mathbb{T}_n}$
- Under **(D), (S)**, with $\mathfrak{n}(dy) = dy$, we have
  - $\mathcal{P}(x, dy\, dz) = \mathcal{P}(x, y, z) dy\, dz$
  - $\mathcal{Q}(x, dy) = \mathcal{Q}(x, y) dy$
  - $\nu(dx) = \nu(x) dx$
- Goal: estimate nonparametrically $x \rightsquigarrow \nu(x)$, $(x, y) \rightsquigarrow \mathcal{Q}(x, y)$ and $(x, y, z) \rightsquigarrow \mathcal{P}(x, y, z)$ for $x, y, z \in \mathcal{D}$.

# Nonparametric estimation of $\nu(x)$

- For a $\sigma$-regular wavelet basis, we approximate the representation

$$\nu(x) = \sum_{\lambda \in \Lambda} \nu_\lambda \psi_\lambda^1(x), \quad \nu_\lambda = \langle \nu, \psi_\lambda^1 \rangle$$

by

$$\widehat{\nu}_n(x) = \sum_{|\lambda| \leq J} \widehat{\nu}_{\lambda,n} \psi_\lambda^1(x),$$

with

$$\widehat{\nu}_{\lambda,n} = \mathcal{T}_{\lambda,\eta}\Big( \frac{1}{|\mathbb{T}_n|} \sum_{u \in \mathbb{T}_n} \psi_\lambda^1(X_u) \Big).$$

- $\mathcal{T}_{\lambda,\eta}(x) = x \mathbf{1}_{|x| \geq \eta}$ threshold operator (with $\mathcal{T}_{\lambda,\eta}(x) = x$ for the low frequency part.
- $\widehat{\nu}_n$ is specified by the maximal resolution level $J$ and the threshold $\eta$.

## Theorem

*Under* **(D)** *and* **(UE)** *with* $\mathfrak{n}(dx) = dx$, *specify* $\widehat{\nu}_n$ *with*

$$J = \log_2 \frac{|\mathbb{T}_n|}{\log |\mathbb{T}_n|} \quad \text{and} \quad \eta = c\sqrt{\log |\mathbb{T}_n|/|\mathbb{T}_n|}$$

*for some* $c > 0$. *For every* $\pi \in (0, \infty]$, $s \in (1/\pi, \sigma]$ *and* $p \geq 1$, *for large enough* $n$ *and* $c$, *the following estimate holds*

$$\left(\mathbb{E}\left[\|\widehat{\nu}_n - \nu\|_{L^p(\mathcal{D})}^p\right]\right)^{1/p} \lesssim \left(\frac{\log |\mathbb{T}_n|}{|\mathbb{T}_n|}\right)^{\alpha_1(s,p,\pi)},$$

*with* $\alpha_1(s, p, \pi) = \min\left\{\frac{s}{2s+1}, \frac{s+1/p-1/\pi}{2s+1-2/\pi}\right\}$, *up to a constant that depends on* $s, p, \pi, \|\nu\|_{\mathcal{B}^s_{\pi,\infty}(\mathcal{D})}$, $\rho$, $R$ *and* $|\mathcal{Q}|_{\mathcal{D}}$ *and that is continuous in its arguments.*

- The estimator $\widehat{\nu}_n$ is *smooth-adaptive* in the following sense: for every $s_0 > 0$, $0 < \rho_0 < 1/2$, $R_0 > 0$ and $\mathcal{Q}_0 > 0$, define the sets $\mathcal{A}(s_0) = \{(s, \pi), s \geq s_0, s_0 \geq 1/\pi\}$ and

$$\mathcal{Q}(\rho_0, R_0, \mathcal{Q}_0) = \{\mathcal{Q} \text{ such that } \rho \leq \rho_0, R \leq R_0, |\mathcal{Q}|_{\mathcal{D}}, \leq \mathcal{Q}_0\},$$

where $\mathcal{Q}$ is taken among mean transitions for which **(UE)** holds. Then, for every $C > 0$, there exists $c^\star = c^\star(\mathcal{D}, p, s_0, \rho_0, R_0, \mathcal{Q}_0, C)$ such that $\widehat{\nu}_n$ specified with $c^\star$ satisfies

$$\sup_n \sup_{(s,\pi)\in\mathcal{A}(s_0)} \sup_{\nu,\mathcal{Q}} \Big(\frac{|\mathbb{T}_n|}{\log|\mathbb{T}_n|}\Big)^{p\alpha_1(s,p,\pi)} \mathbb{E}\big[\|\widehat{\nu}_n - \nu\|^p_{L^p(\mathcal{D})}\big] < \infty$$

where the supremum is taken among $(\nu, \mathcal{Q})$ such that $\nu\mathcal{Q} = \nu$ with $\mathcal{Q} \in \mathcal{Q}(\rho_0, R_0, \mathcal{Q}_0)$ and $\|\nu\|_{\mathcal{B}^s_{\pi,\infty}(\mathcal{D})} \leq C$.

# Nonparametric estimation of the mean transition $\mathcal{Q}(x, y)$

- First estimate
$$f_{\mathcal{Q}}(x,y) = \nu(x)\mathcal{Q}(x,y)$$
of the distribution of $(X_{u^-}, X_u)$ (when $\mathcal{L}(X_\emptyset) = \nu$) by
$$\widehat{f}_n(x,y) = \sum_{|\lambda| \leq J} \widehat{f}_{\lambda,n} \psi_\lambda^2(x,y),$$
with
$$\widehat{f}_{\lambda,n} = \mathcal{T}_{\lambda,\eta}\Big(\frac{1}{|\mathbb{T}_n^\star|} \sum_{u \in \mathbb{T}_n^\star} \psi_\lambda^2(X_{u^-}, X_u)\Big),$$
$(\mathbb{T}_n^\star = \mathbb{T}_n \setminus \mathbb{G}_0.)$
- Estimate $\mathcal{Q}(x, y)$ via
$$\widehat{\mathcal{Q}}_n(x,y) = \frac{\widehat{f}_n(x,y)}{\max\{\widehat{\nu}_n(x), \varpi\}} \tag{3}$$
for some $\varpi > 0$.
- Thus $\widehat{\mathcal{Q}}_n$ is specified by $J$, $\eta$ and $\varpi$.

## Theorem

*Under* **(D)** *and* **(UE)** *with* $\mathfrak{n}(dx) = dx$, *specify* $\widehat{\mathcal{Q}}_n$ *with*

$$J = \tfrac{1}{2} \log_2 \frac{|\mathbb{T}_n|}{\log |\mathbb{T}_n|} \quad \text{and} \quad \eta = c\sqrt{(\log |\mathbb{T}_n|)^2 / |\mathbb{T}_n|}$$

*for some* $c > 0$ *and* $\varpi > 0$. *For every* $\pi \in [1, \infty]$, $s \in (2/\pi, \sigma]$ *and* $p \geq 1$, *for large enough* $n$ *and* $c$ *and small enough* $\varpi$, *the following estimate holds*

$$\Big(\mathbb{E}\big[\|\widehat{\mathcal{Q}}_n - \mathcal{Q}\|^p_{L^p(\mathcal{D}^2)}\big]\Big)^{1/p} \lesssim \Big(\frac{(\log |\mathbb{T}_n|)^2}{|\mathbb{T}_n|}\Big)^{\alpha_2(s,p,\pi)}, \tag{4}$$

*with* $\alpha_2(s, p, \pi) = \min\big\{\frac{s}{2s+2}, \frac{s/2+1/p-1/\pi}{s+1-2/\pi}\big\}$, *provided* $m(\nu) = \inf_{x \in \mathcal{D}} \nu(x) \geq \varpi > 0$ *and up to a constant that depends on* $s, p, \pi, \|\mathcal{Q}\|_{\mathcal{B}^s_{\pi,\infty}(\mathcal{D}^2)}$, $m(\nu)$ *and that is continuous in its arguments.*

- This rate is moreover (nearly) optimal: define $\varepsilon_2 = s\pi - (p - \pi)$. We have

$$\inf_{\widehat{\mathcal{Q}}_n} \sup_{\mathcal{Q}} \Big(\mathbb{E}\big[\|\widehat{\mathcal{Q}}_n - \mathcal{Q}\|^p_{L^p(\mathcal{D}^2)}\big]\Big)^{1/p} \gtrsim \begin{cases} |\mathbb{T}_n|^{-\alpha_2(s,p,\pi)} & \text{if} \quad \varepsilon_2 > 0 \\ \Big(\frac{\log|\mathbb{T}_n|}{|\mathbb{T}_n|}\Big)^{\alpha_2(s,p,\pi)} & \text{if} \quad \varepsilon_2 \leq 0. \end{cases}$$

  where the infimum is taken among all estimators of $\mathcal{Q}$ based on $(X_u)_{u\in\mathbb{T}_n}$ and the supremum is taken among all $\mathcal{Q}$ such that $\|\mathcal{Q}\|_{\mathcal{B}^s_{\pi,\infty}(\mathcal{D}^2)} \leq C$ and $m(\nu) \geq C'$ for some $C, C' > 0$.
- The calibration of the threshold $\varpi$ needed to define $\widehat{\mathcal{Q}}_n$ requires an *a priori* bound on $m(\nu)$.
- The $(\log|\mathbb{T}_n|)^2$ comes from the slow term in the deviations inequality and from the wavelet thresholding procedure.

## Nonparametric estimation of the transition $\mathcal{P}(x,y,z)$

- First estimate the density

$$f_{\mathcal{P}}(x,y,z) = \nu(x)\mathcal{P}(x,y,z)$$

of the distribution of $(X_u, X_{u0}, X_{u1})$ (when $\mathcal{L}(X_\emptyset) = \nu$) by

$$\widehat{f}_n(x,y,z) = \sum_{|\lambda| \leq J} \widehat{f}_{\lambda,n}\psi_\lambda^3(x,y,z),$$

with

$$\widehat{f}_{\lambda,n} = \mathcal{T}_{\lambda,\eta}\Big(\frac{1}{|\mathbb{T}_{n-1}|}\sum_{u \in \mathbb{T}_{n-1}} \psi_\lambda^3(X_u, X_{u0}, X_{u1})\Big),$$

- Next estimate the density $\mathcal{P}$ by

$$\widehat{\mathcal{P}}_n(x,y,z) = \frac{\widehat{f}_n(x,y,z)}{\max\{\widehat{\nu}_n(x), \varpi\}} \tag{5}$$

for some threshold $\varpi > 0$.

- Thus the estimator $\widehat{\mathcal{P}}_n$ is specified by $J$, $\eta$ and $\varpi$.

## Theorem

*Under* **(D), (UE), (S)**. *Specify* $\widehat{\mathcal{P}}_n$ *with*

$$J = \frac{1}{3}\log_2 \frac{|\mathbb{T}_n|}{\log|\mathbb{T}_n|} \quad \text{and} \quad \eta = c\sqrt{(\log|\mathbb{T}_n|)^2/|\mathbb{T}_n|}$$

*for some* $c > 0$ *and* $\varpi > 0$. *For every* $\pi \in [1, \infty]$, $s \in (3/\pi, \sigma]$ *and* $p \geq 1$, *for large enough* $n$ *and* $c$ *and small enough* $\varpi$, *the following estimate holds*

$$\left(\mathbb{E}\left[\|\widehat{\mathcal{P}}_n - \mathcal{P}\|^p_{L^p(\mathcal{D}^3)}\right]\right)^{1/p} \lesssim \left(\frac{(\log|\mathbb{T}_n|)^2}{|\mathbb{T}_n|}\right)^{\alpha_3(s,p,\pi)}, \tag{6}$$

*with* $\alpha_3(s,p,\pi) = \min\left\{\frac{s}{2s+3}, \frac{s/3+1/p-1/\pi}{2s/3+1-2/\pi}\right\}$, *provided* $m(\nu) \geq \varpi > 0$ *and up to a constant that depends on* $s, p, \pi, \|\mathcal{P}\|_{\mathcal{B}^s_{\pi,\infty}(\mathcal{D}^3)}$ *and* $m(\nu)$ *and that is continuous in its arguments.*

- This rate is moreover (nearly) optimal: define $\varepsilon_3 = \frac{s\pi}{3} - \frac{p-\pi}{2}$. We have

$$\inf_{\widehat{\mathcal{P}}_n} \sup_{\mathcal{P}} \Big(\mathbb{E}\big[\|\widehat{\mathcal{P}}_n - \mathcal{P}\|^p_{L^p(\mathcal{D}^3)}\big]\Big)^{1/p} \gtrsim \begin{cases} |\mathbb{T}_n|^{-\alpha_3(s,p,\pi)} & \text{if} \quad \varepsilon_3 > 0 \\ \Big(\dfrac{\log|\mathbb{T}_n|}{|\mathbb{T}_n|}\Big)^{\alpha_3(s,p,\pi)} & \text{if} \quad \varepsilon_3 \leq 0, \end{cases}$$

where the infimum is taken among all estimators of $\mathcal{P}$ based on $(X_u)_{u\in\mathbb{T}_n}$ and the supremum is taken among all $\mathcal{P}$ such that $\|\mathcal{P}\|_{\mathcal{B}^s_{\pi,\infty}(\mathcal{D}^3)} \leq C$ and $m(\nu) \geq C'$ for some $C, C' > 0$.