Les processus d'évolution génétique en filtrage de signaux et en analyse de risques

P. Del Moral

INRIA Centre Bordeaux-Sud Ouest

Séminaire de Stat. et Santé Publique de l'IFR 99, décembre 08

qq-références :

↪ Feynman-Kac formulae. Genealogical and interacting particle systems, Springer (2004), + Réfs

↪ + Doucet, Jasra. SMC Samplers. *JRSS B (2006).*

↪ + Kallel, Rowe, Modelling genetic models with IPS, Theoretical Aspects of Evolutionary Computing (2000)

Outline

Summary

Dynamique adaptative des populations *(temps discret)*

- Population d'individus évoluant en interaction selon des mécanismes de mutation/exploration coupl/'es à des processus de naissance et mort et/ou de branchements spatiaux-temporel.
- **Processus stochastique formalisé :** *(taille fixe N pour simplifier)*

$$\Rightarrow \text{Population} = N \text{ individus } \xi_n^i \text{ dans un espace d'état } E_n$$

  - Indice-Label de l'individu : $i \in \{1, \ldots, N\}$
  - Paramètre temporel : $n \geq 0$ *(temps discret)*.
- **Exemples d'espace d'états :**

  Gènes (séquences adn, allèles), positions, vitesses, accélérations, permutations, sous-espaces de points, codes génétiques, traits d'individus, age, taille, poids, communautés, nationalité, lignes ancestrales, excursions trajectorielles, ...

2 mécanismes d'adaptation naturelle $\Rightarrow$ **Sélection & Mutation**

- $\boxed{\textbf{Sélection}}$ (*=naissance/mort=clonage=branchements*) :

$$\left(\xi_n^1,\ldots,\xi_n^N\right) \in E_n^N \longrightarrow \underbrace{\left(\widehat{\xi}_n^1,\ldots,\widehat{\xi}_n^N\right)}_{\text{individus sélectionnés}} \in E_n^N$$

$\sim$ Potentiel d'adaptation $\quad G_n \ : \ x_n \in E_n \mapsto G_n(x_n) \in [0,1]$

1. Acceptation $\rightsquigarrow \widehat{\xi}_n^i = \xi_n^i$ avec une proba $G_n(\xi_n^i)$
2. *Sinon* saut sur le j-ième $\widehat{\xi}_n^i = \xi_n^j$ avec une proba $\frac{G_n(\xi_n^j)}{\sum_{k=1}^N G_n(\xi_n^k)}$

QQ exemples académiques élémentaires:

$$G_n(x_n) = 1_{A_n}(x_n), \ e^{-\frac{1}{2}\ (y_n-x_n)^2} \text{ avec } y_n \text{ référence fixée}$$

ou encore en trajectoriel $\rightsquigarrow$ sélection de trajectoires

$$x_n = (x_0', \ldots, x_n') \rightsquigarrow G_n(x_n) = 1_{(x_p' \neq x_q' \ , \ \forall p<q<n)}$$

2 mécanismes d'adaptation naturelle $\Rightarrow$ **Sélection & Mutation**

- $\boxed{\textbf{Mutation}}$ (*=exploration=proposition=prédiction*) :

$$\left(\widehat{\xi}_n^1, \ldots, \widehat{\xi}_n^N\right) \in E_n^N \longrightarrow \left(\xi_{n+1}^1, \ldots, \xi_{n+1}^N\right) \in E_{n+1}^N$$

Indépendamment les uns des autres, chaque individu sélectionné $\widehat{\xi}_n^i$ explore l'espace en mimant les transitions

$$X_n = \widehat{\xi}_n^i \rightsquigarrow X_{n+1} = \xi_{n+1}^i$$

d'une chaîne de Markov $X_n \in E_n \rightsquigarrow X_{n+1} = E_{n+1}$.

QQ exemples académiques élémentaires:

$$X_n = \text{marche aléa. sur } \mathbb{Z}, \text{ processus } X_n = F_n(X_{n-1}, W_n) \text{ sur } \mathbb{R}^d$$

ou encore en trajectoriel $\rightsquigarrow$ extension locale de trajectoires

$$X_n = (X'_0, \ldots, X'_n) \rightsquigarrow X_{n+1} = ((X'_0, \ldots, X'_n), X'_{n+1})$$

Algorithme d'évolution génétique

$$\left.\begin{array}{c}\xi_n^1\\ \vdots\\ \xi_n^i\\ \vdots\\ \xi_n^N\end{array}\right] \xrightarrow{\text{Sélection}} \left[\begin{array}{ccc}\widehat{\xi}_n^1 & \xrightarrow{\text{Mutation}} & \xi_{n+1}^1\\ \vdots & & \vdots\\ \widehat{\xi}_n^i & \longrightarrow & \xi_{n+1}^i\\ \vdots & & \vdots\\ \widehat{\xi}_n^N & \longrightarrow & \xi_{n+1}^N\end{array}\right]$$

Exemples :

- X_n marche aléatoire sur $\mathbb{Z}$ & $G_n(X_n) = 1_A(X_n)$.
- $X_n = F_n(X_{n-1}, W_n)$ dyn. aléa sur $\mathbb{R}^d$ & $G_n(X_n) = e^{-\frac{1}{2}(Y_n - H_n(X_n))^2}$.
- $X_n = X'_{[T_{n-1}, T_n]}$ excursion inter-niveaux & $G_n(X_n) = 1_{B_n}(X'_{T_n})$.
- $X_n = X'_{[0,n]}$ traject. aléa. sur $\mathbb{Z}^2$ & $G_n(X_n) = 1_{(X'_p \neq X'_q,\ \forall p<q\leq n)}$.
- Modèles trajectoriels $\mathbf{X}_n = X_{[0,n]}$ & $\mathbf{G_n}(\mathbf{X_n}) = G_n(X_n)$.

Summary

(Sélection & Mutation) trajectorielle $\Leftrightarrow$Arbres généalogiques

- **Mutation** $=$ extension locale des trajectoires $\sim$ Proc. trajectoriel

$$X_n = (X'_0, \ldots, X'_n) \rightsquigarrow X_{n+1} = ((X'_0, \ldots, X'_n), X'_{n+1})$$

- **Sélection**=clonage=sauts des trajectoires $\sim$ Potentiel trajectoriel

$$G_n(X_n) = G_n(X'_0, \ldots, X'_n)$$

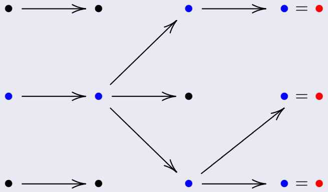
$$\Downarrow$$

Arbre généalogique

- **Individus trajectoriels** $\rightsquigarrow \xi^i_n := \underbrace{\left(\xi^i_{0,n}, \xi^i_{1,n}, \ldots, \xi^i_{n,n}\right)}_{=\text{ligne ancestrale}}$

- **Arbre généalogique** $= \left(\xi^i_{0,n}, \xi^i_{1,n}, \ldots, \xi^i_{n,n}\right)_{1 \leq i \leq N}$

$(N = 3)$



- **Population courante** $\hookrightarrow \eta_n^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_n^i}$ $\leftarrow$ i-ième individu au temps n

- **Arbre généalogique** $\hookrightarrow \frac{1}{N}\sum_{i=1}^{N} \delta_{(\xi_{0,n}^i,\xi_{1,n}^i,\ldots,\xi_{n,n}^i)}$ $\leftarrow$ i-ième ligne ancestrale

- **Arbre généalogique complet** $\hookrightarrow \frac{1}{N}\sum_{i=1}^{N} \delta_{(\xi_0^i,\xi_1^i,\ldots,\xi_n^i)}$

- $\oplus$ **Potentiels moyens [% de succès $(G_n = 1_A)$]** $\hookrightarrow \frac{1}{N}\sum_{i=1}^{N} G_n(\xi_n^i)$

Summary

Confinement $\Leftrightarrow$ absorption $\rightsquigarrow$ Quantum & Diffusion Monte Carlo

X_n marche aléatoire sur $\mathbb{Z}$ & $G_n(X_n) = 1_A(X_n)$.

$$\Downarrow$$

$$\eta_n^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi_n^i} \simeq_{N\uparrow\infty} \mathrm{Loi}(X_n \mid \forall 0 \leq p < n \quad X_p \in A) := \eta_n$$

et

$$\prod_{0\leq p<n} \frac{1}{N}\sum_{i=1}^N 1_A\left(\xi_p^i\right) \simeq_{N\uparrow\infty} \mathbb{P}\left(\forall 0 \leq p < n \quad X_p \in A\right)$$

et

$$\frac{1}{N}\sum_{i=1}^N \delta_{\left(\xi_{0,n}^i,\xi_{1,n}^i,\ldots,\xi_{n,n}^i\right)} \simeq_{N\uparrow\infty} \mathrm{Loi}((X_0,\ldots,X_n) \mid \forall 0 \leq p < n \quad X_p \in A)$$

Remarque : idem avec des mutation-sélection données par

$$(\widehat{X}_n \text{ transition locales} \in A) \;\&\; (\widehat{G}_n(x_n) = \mathbb{P}(X_{n+1} \in A \mid X_n = x_n))$$

Comparaisons & Convergence

- *Méthode de Monte Carlo usuelle* $\sim$ *copies iid* X^i *de* X

$$(1) \qquad \frac{1}{N}\sum_{i=1}^{N}\prod_{0\leq p<n} 1_A\left(X_p^i\right) \quad \simeq_{N\uparrow\infty} \quad \mathbb{P}\left(\forall 0\leq p<n \;\; X_p\in A\right)$$

$$(2) \qquad \prod_{0\leq p<n}\frac{1}{N}\sum_{i=1}^{N} 1_A\left(\xi_p^i\right) \quad \simeq_{N\uparrow\infty} \quad \mathbb{P}\left(\forall 0\leq p<n \;\; X_p\in A\right)$$

- *Est. sans biais de variances :* $[P_n=\mathbb{P}\left(\forall 0\leq p<n \;\; X_p\in A\right)\simeq e^{-a\;n}]$

$$\mathrm{Var}(1)=\frac{1}{N}\;P_n\;(1-P_n) \quad \& \quad \mathrm{Var}(2)\leq c\;\frac{n}{N}\;P_n^2$$

de plus (sous qq hyp. de regularité)

$$\sup_{n\geq 0}\mathbb{E}\left(\left|\eta_n^N(f)-\eta_n(f)\right|^p\right)^{1/p}\leq c/\sqrt{N}$$

Marches aléatoires répulsives $\rightsquigarrow$ prune-enriched Rosenbluth method

$X_n = X'_{[0,n]}$ traject. aléa. sur $\mathbb{Z}^2$ & $G_n(X_n) = 1_{(X'_p \neq X'_q,\ \forall p<q\leq n)}$.

$$\Downarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{(\xi^i_{0,n},\xi^i_{1,n},\ldots,\xi^i_{n,n})} \simeq_{N\uparrow\infty} \mathrm{Loi}((X'_0,\ldots,X'_n) \mid \forall p<q<n \ \ X'_p \neq X'_q)$$

et

$$\prod_{0\leq p<n} \frac{1}{N}\sum_{i=1}^{N} 1_{(\xi^i_{k,p}\neq\xi^i_{l,p},\ \forall k<l\leq p)} \simeq_{N\uparrow\infty} \mathbb{P}\left(\forall p<q<n \ \ X'_p \neq X'_q\right)$$

Remarque : idem avec des mutation-sélection données par

$(\widehat{X}'_n$ transition locales sans intersections$)$

&

$(\widehat{G}_n(x_n) = \mathbb{P}(X'_{n+1} \text{ intersecte le passé} \mid X_n = x_n))$

Processus de ruine = excursions/branchements sur des niveaux $\downarrow$

- $X_n = X'_{[T_{n-1}, T_n]}$ excursion inter-niveaux & $G_n(X_n) = 1_{B_n}(X'_{T_n})$.
- T_n=temps d'atteinte du niveau B_n ou d'un ensemble attractif C

$$\Downarrow$$

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{\text{ligne ancestrale}_n(i)} \simeq_{N\uparrow\infty} \text{Loi(excursions de X} \mid B_n \text{ atteint avant } C)$$

et

$$\prod_{0 \le p \le n} (\text{\% de succès au temps p}) \simeq_{N\uparrow\infty} \mathbb{P}(B_n \text{ atteint avant } C)$$

Filtrage de processus $\rightsquigarrow$ filtres particulaires

$X_n = F_n(X_{n-1}, W_n)$ dyn. aléa sur $\mathbb{R}^d$ & $G_n(X_n) = e^{-\frac{1}{2}(Y_n - H_n(X_n))^2}$.

- Processus d'observations :

$$Y_n = H_n(X_n) + V_n \rightsquigarrow \text{avec} \quad V_n \text{ bruits de mesure } \sim \mathcal{N}(0,1)$$

- Fonctions de vraisemblance locale :

$$\mathbb{P}\left(Y_n \in dy \mid X_n = x\right) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(y - H_n(x))^2}\, dy$$

$$\Downarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{\text{ligne ancestrale}_n(i)} \simeq_{N\uparrow\infty} \mathrm{Loi}((X_0,\ldots,X_n) \mid (Y_0,\ldots,Y_n))$$

Summary

Mesures de Feynman-Kac (Fonctions test $f : E_n \to \mathbb{R}$)

- ***X*-explorations** & **G_n-selection**

$$\Longrightarrow \eta_n^N(f) := \frac{1}{N}\sum_{i=1}^{N} f(\xi_n^i) \longrightarrow_{N\uparrow\infty} \eta_n(f) := \frac{\gamma_n(f)}{\gamma_n(1)}$$

avec

$$\gamma_n(f) := \mathbb{E}\left(f_n(X_n) \prod_{0\leq p<n} G_p(X_p)\right)$$

- *Ex. : Confinement* $G_n = 1_A$

$$\gamma_n(1) = \mathbb{P}\left(\forall 0 \leq p < n \quad X_p \in A\right) \quad \& \quad \eta_n = \mathrm{Law}\left(X_n \mid \forall 0 \leq p < n \quad X_p \in A\right)$$

Ctes de normalisation

Formule produit :

$$\gamma_n(1) := \mathbb{E}\left(\prod_{0\leq p<n} G_p(X_p)\right) = \prod_{0\leq p<n} \eta_p(G_p)$$

$$\Downarrow$$

Estimations non biaisées :

$$\gamma_n^N(1) := \prod_{0\leq p<n} \eta_p^N(G_p) \longrightarrow_{N\uparrow\infty} \gamma_n(1) = \prod_{0\leq p<n} \eta_p(G_p)$$

Ex. : Confinement $G_n = 1_A$

$$\prod_{0\leq p<n} \textbf{(proportion de succès tps p)} \simeq \mathbb{P}\left(\forall 0 \leq p < n \quad X_p \in A\right)$$

Processus historique et arbres généalogiques

- **Mutation et selection de trajectoires :**

$$\rightsquigarrow \boxed{X_n := (X'_0, \ldots, X'_n) \in E_n := (E'_0 \times \ldots \times E'_n)}$$

- **Population courante trajectorielle=Lignes ancestrales :**

$$\eta_n^N(f_n) := \frac{1}{N} \sum_{i=1}^{N} f_n\left(\xi^i_{0,n}, \xi^i_{1,n}, \ldots, \xi^i_{n,n}\right) \longrightarrow_{N\uparrow\infty} \eta_n(f_n) := \frac{\gamma_n(f_n)}{\gamma_n(1)}$$

avec

$$\gamma_n(f_n) := \mathbb{E}\left(f_n(X'_0, \ldots, X'_n) \prod_{0\leq p<n} G_p(X'_0, \ldots, X'_p)\right)$$

Ex. : Confinement $G_n = 1_A$

$$\eta_n = \text{Loi}\left((X'_0, \ldots, X'_n) \mid \forall 0 \leq p < n \quad X'_p \in A\right)$$

Arbres généalogiques complets

- $\sim$ **algo. génétique "sans proba. d'acceptation"**

$$\frac{1}{N}\sum_{i=1}^{N} F_n\left(\xi_0^i,\xi_1^i,\ldots,\xi_n^i\right) \longrightarrow_{N\uparrow\infty} \left(\eta_0\otimes\ldots\otimes\eta_n\right)(F_n)$$

avec les mesures produits

$$\left(\eta_0\otimes\ldots\otimes\eta_n\right)(F_n)=\int_{E_0}\ldots\int_{E_n} \eta_0(dx_0)\ldots\eta_n(dx_n)\ F_n(x_0,\ldots,x_n)$$

- **"Avec proba. d'acceptation" $\rightsquigarrow \neq$ Mesures de McKean :**

$$\eta_n=\mathbf{Loi}(\overline{X}_n)\quad \textbf{avec des transitions Markov}\quad \overline{X}_n \overset{\eta_n}{\rightsquigarrow} \overline{X}_{n+1}$$

$$\Updownarrow$$

$$\boxed{\textbf{Algorithme Génétique = Interprétation champ moyen de } \overline{X}_n}$$

Résumé-Conclusions

$\boxed{\textbf{Algorithme Génétique}}$

[X_n-**mutations** $\oplus$ G_n-**adaptation/selection/branchement**]

$\Downarrow$ & $\Uparrow$

$\boxed{\textbf{Mesures de Feynman-Kac}}$

[X_n-**Markov de référence.** $\oplus$ G_n-**fonctions de potentiel**]

$\Updownarrow$

!!! Algorithme génétique $\neq$ Métaheuristique (cf. Wikipedia définition) !!!!!

Algorithmes stochastiques "équivalents" :

- Algorithmes génétiques, processus de branchements.
- Méthodes de Monte Carlo séquentielles.
- Algorithmes de "population Monte Carlo", Diffusion Monte Carlo (DMC), Quantum Monte Carlo (QMC), ...
- Botanique des processus d'adaptation $\sim \neq$ domaines d'applications :

  Bootstrapping, selection, pruning-enrichment, reconfiguration, cloning, go with the winner, spawning, condensation, grouping, rejuvenations, harmony searches, biomimetics, ...

Summary

Flot (non linéaires) de mesures de probabilités

- Flots non linéaires : $[\Phi_n : \text{Proba} \mapsto \text{Proba}]$

$$\eta_n = \Phi_n\left(\eta_{n-1}\right)$$

- $\exists$ Famille de transitions Markov $K_{n,\eta}$ (pas unique!)

$$\begin{aligned}\eta_n(dy) &= \Phi_n\left(\eta_{n-1}\right)(dy) \\ &= \left(\eta_{n-1}K_{n,\eta_{n-1}}\right)(dy) := \int \eta_{n-1}(dx)\; K_{n,\eta_{n-1}}(x, dy)\end{aligned}$$

Ex. : Mesures de Feynman-Kac

$$\eta_n \xrightarrow{\text{Correction}} \widehat{\eta}_n = \Psi_{G_n}(\eta_n) \xrightarrow{\text{Prédiction}} \eta_{n+1} = \widehat{\eta}_n M_{n+1}$$

Transf. de Boltzmann-Gibbs : $G_n(x)$=vraisemblance/qualité du site x

$$\Psi_{G_n}(\eta_n)(dx) := \frac{1}{\eta_n(G_n)} \; G_n(x) \; \eta_n(dx) = \eta_n S_{n,\eta_n}(dx)$$

avec le transport

$$S_{n,\eta_n}(x, dy) := G_n(x) \; \delta_x(dy) + (1 - G_n(x)) \;\; \Psi_n(\eta_n)(dy)$$

$$\Downarrow$$

$$\eta_{n+1} = \eta_n S_{n,\eta_n} M_{n+1}$$

Chaines de Markov $\overline{X}_n$ non linéaires t.q. $\eta_n = \mathrm{Loi}(\overline{X}_n)$

- **Transitions locales :**

$$\mathbb{P}(\overline{X}_n \in dx_n \mid \overline{X}_{n-1}) = K_{n,\eta_{n-1}}(\overline{X}_{n-1}, dx_n)$$

- **Mesure de McKean (processus canonique) :**

$$\mathbb{P}_n(d(x_0,\ldots,x_n)) = \eta_0(dx_0)\ K_{1,\eta_0}(x_0,dx_1)\ldots K_{n,\eta_{n-1}}(x_{n-1},dx_n)$$

Interprétation particulaire de type champ moyen

- Chaine Markov $\xi_n = (\xi_n^1,\ldots,\xi_n^N) \in E_n^N$ t.q.

$$\eta_n^N := \frac{1}{N}\sum_{1\leq i\leq N}\delta_{\xi_n^i}\ \simeq_{N\uparrow\infty}\ \eta_n$$

- Transitions locales "approchées" $(\forall 1 \leq i \leq N)$

$$\xi_{n-1}^i \rightsquigarrow \xi_n^i \sim K_{n,\eta_{n-1}^N}(\xi_{n-1}^i, dx_n)$$

Figure schématique : $\xi_n \in E_n^N \rightsquigarrow \xi_{n+1} \in E_{n+1}^N$

$$\begin{array}{ccc}
\xi_n^1 & \xrightarrow{K_{n+1,\eta_n^N}} & \xi_{n+1}^1 \\
\vdots & & \vdots \\
\xi_n^i & \longrightarrow & \xi_{n+1}^i \\
\vdots & & \vdots \\
\xi_n^N & \longrightarrow & \xi_{n+1}^N
\end{array}$$

Idée intuitive :

$$\begin{aligned}
\eta_n^N \simeq_{N\uparrow\infty} \eta_n \quad & \Longrightarrow \quad K_{n+1,\eta_n^N} \simeq_{N\uparrow\infty} K_{n+1,\eta_n} \\
& \Longrightarrow \quad \xi_n^i \text{ copies i.i.d. de } \overline{X}_n
\end{aligned}$$

& les mesures de McKean trajectorielles :

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{\left(\xi_0^i,\xi_1^i,\ldots,\xi_n^i\right)} \longrightarrow_{N\uparrow\infty} \eta_0 \times K_{1,\eta_0} \times \ldots \times K_{n,\eta_{n-1}}$$

QQ Avantages

- Modèle de champ moyen=linéarisation/perturbation stoch. :

$$\eta_n^N = \eta_{n-1}^N K_{n,\eta_{n-1}^N} + \frac{1}{\sqrt{N}}\; W_n^N$$

avec $W_n^N \simeq W_n$ Champs gaussiens centrés et $\perp$.

- $\eta_n = \eta_{n-1} K_{n,\eta_{n-1}}$ stable $\Rightarrow$ non propagation des erreurs locales

$\Longrightarrow$ control uniforme des erreurs globales / temps

- "Pas besoin" d'étudier la cv à l'équilibre de modèles MCMC.
- Grille/Maillage stochastique adaptatif.
- Non linéarité du syst. $\rightsquigarrow$ interactions bénéfiques.
- Algo. naturel et facile a simuler, etc.

Theorie "asymptotique" TCL,PGD, PDM,...(n,N). QQ exemples :

- Mesure McKean particulaire

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{(\xi_0^i,\ldots,\xi_n^i)} \simeq_N \mathrm{Loi}(\overline{X}_0,\ldots,\overline{X}_n) \quad \text{et} \quad \eta_n^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_n^i} \simeq_N \eta_n$$

- Proc. empiriques : $\sup_{n\geq 0}\sup_{N\geq 1}\sqrt{N}\,\mathbb{E}(\|\eta_n^N-\eta_n\|_{\mathcal{F}_n}^p)<\infty$
- Inégalités de concentration uniformes :

$$\sup_{n\geq 0}\mathbb{P}(|\eta_n^N(f_n)-\eta_n(f_n)|>\epsilon)\leq c\ \exp-(N\epsilon^2)/(2\ \sigma^2)$$

- Propagations du chaos : $\mathbb{P}_{n,q}^N := \mathrm{Loi}(\xi_n^1,\ldots,\xi_n^q)$

$$\mathbb{P}_{n,q}^N \simeq \eta_n^{\otimes q} + \frac{1}{N}\ \partial^1\mathbb{P}_{n,q} + \ldots + \frac{1}{N^k}\ \partial^k\mathbb{P}_{n,q} + \frac{1}{N^{k+1}}\ \partial^{k+1}\mathbb{P}_{n,q}^N$$

avec $\sup_{N\geq 1}\|\partial^{k+1}\mathbb{P}_{n,q}^N\|_{\mathrm{tv}}<\infty$ et $\sup_{n\geq 0}\|\partial^1\mathbb{P}_{n,q}\|_{\mathrm{tv}}\leq c\ q^2$.