École Nationale Supérieure de Techniques Avancées
examen du cours SOD333
"Filtrage bayésien et approximation particulaire"

vendredi 25 octobre 2019, 13:30 à 16:00

Problème

L'objectif de ce problème est de revisiter l'algorithme d'échantillonnage résiduel multinomial dans le cas particulier de poids binaires, et d'étudier ensuite un théorème central limite, avec une expression si possible explicite de la variance asymptotique, pour l'algorithme SIR

- soit avec rééchantillonnage résiduel multinomial (vu en cours),
- ou bien avec rééchantillonnage résiduel sans remise (décrit et étudié dans la deuxième partie du problème),

dans ce cas particulier de poids binaires. Pour fixer le cadre, on considère un mélange fini de distributions de probabilité

$$\eta = \sum_{i=1}^{N} w_i \, m_i \ ,$$

avec des poids binaires associés à un sous–ensemble $I \subset \{1, \cdots, N\}$, c'est–à–dire que

$$w_i = \begin{cases} 0 \ , & \text{si } i \notin I \\ \dfrac{1}{|I|} \ , & \text{si } i \in I \end{cases}$$

de sorte que, en pratique

$$\eta = \frac{1}{|I|} \sum_{i \in I} m_i \ .$$

On se propose d'approcher la distribution de mélange η par une combinaison linéaire convexe η_N de N masses de Dirac, en utilisant l'algorithme d'échantillonnage résiduel multinomial, ou bien un algorithme d'échantillonnage résiduel sans remise.

On commence par établir une identité qui sera utile pour la suite, indépendemment de l'algorithme d'échantillonnage utilisé.

(i) **Montrer que**

$$\mathrm{var}(\phi,\eta) - \frac{1}{|I|}\sum_{i\in I}\mathrm{var}(\phi,m_i) = \frac{1}{|I|}\sum_{i\in I}|\langle m_i,\phi\rangle|^2 - |\frac{1}{|I|}\sum_{i\in I}\langle m_i,\phi\rangle|^2 ,$$

pour toute fonction mesurable bornée ϕ. Le terme à droite de l'égalité peut s'interpréter comme la variance des moyennes intra–composantes affectées de poids uniformes.

Échantillonnage résiduel multinomial

(ii) **En reprenant les définitions vues en cours, donner l'expression des variables N_i et q_i pour tout $i = 1,\cdots,N$, de la variable N_0 et de la distribution de mélange m_0, dans ce cas particulier de poids binaires.**

(iii) **Décrire l'algorithme d'échantillonnage résiduel multinomial dans ce cas particulier de poids binaires. Expliquer comment sont générées les N variables aléatoires qui apparaissent dans la combinaison linéaire convexe η_N de N masses de Dirac.**

(iv) **Donner l'expression de la variance de l'erreur d'estimation pour l'algorithme d'échantillonnage résiduel multinomial dans ce cas particulier de poids binaires. Comment cette variance se compare–t–elle avec la variance de l'erreur d'estimation pour l'algorithme d'échantillonnage multinomial ?**

Échantillonnage résiduel sans remise

Dans ce cas particulier de poids binaires, l'algorithme d'échantillonnage résiduel multinomial consiste à affecter de manière déterministe à chaque composante du mélange le même nombre de représentants, puis à compléter la population par échantillonnage multinomial selon la distribution de mélange initiale. Cette stratégie introduit un aléa excessif, puisque rien n'interdit dans la phase d'échantillonnage multinomial que certaines composantes recoivent plus que un représentant supplémentaire, alors qu'en principe toutes les composantes du mélange ont la même importance. Le comportement désirable serait que toutes les composantes du mélange recoivent au plus un (zéro ou un) représentant supplémentaire.

Concrètement, pour toute composante $i \in I$ le nombre N_* de représentants affectés à l'issue de la première passe est donné par la division euclidienne

$$\frac{N}{|I|} = N_* + q_* \qquad \text{avec} \qquad 0 \le q_* < 1 .$$

On en déduit que $N_* \, |I|$ représentants ont déjà été affectés à l'issue de la première passe, et il reste donc $N_0 = N - N_* \, |I| = q_* \, |I|$ représentants à affecter, à raison de zéro ou un représentant supplémentaire pour chaque composante du mélange. Au final, pour toute composante $i \in I$ le nombre total de représentants affectés est $N_i = N_* + \varepsilon_i$ avec $\varepsilon_i \in \{0, 1\}$, sous la contrainte

$$\sum_{i \in I} N_i = \sum_{i \in I} (N_* + \varepsilon_i) = N_* \, |I| + \sum_{i \in I} \varepsilon_i = N \ ,$$

soit

$$\sum_{i \in I} \varepsilon_i = q_* \, |I| \ . \qquad (\star)$$

Compte tenu des identités

$$N = N_* \, |I| + q_* \, |I| = N_* \, |I| + N_0 \qquad \text{avec} \qquad N_0 = q_* \, |I| \ ,$$

et

$$\eta = \frac{1}{|I|} \ \sum_{i \in I} m_i = \frac{1}{N} \ \sum_{i \in I} \frac{N}{|I|} \ m_i = \frac{1}{N} \ \sum_{i \in I} (N_* + q_*) \ m_i \ ,$$

l'approximation proposée consiste à simuler

- pour tout $i \in I$, un $(N_* + 1)$–échantillon $(\xi^{i,0}, \xi^{i,1}, \cdots, \xi^{i,N_*})$ distribué selon m_i,
- une collection $(\varepsilon_i \, , \, i \in I)$ de variables aléatoires *échangeables** à valeurs binaires 0 ou 1, vérifiant la contrainte $(\star)$: il suffit par exemple de construire un sous–ensemble aléatoire $J \subset I$ de taille N_0, par tirage uniforme sans remise dans le sous–ensemble I, et de poser $\varepsilon_i = 1_{(i \, \in \, J)}$ pour tout $i \in I$,

toutes les variables aléatoires étant simulées de manière indépendantes, et à poser

$$\eta_N = \frac{1}{N} \ \sum_{i \in I} \sum_{j=1}^{N_*} \delta_{\xi^{i,j}} + \frac{1}{N} \ \sum_{i \in I} \varepsilon_i \ \delta_{\xi^{i,0}} = \frac{1}{N} \ \sum_{i \in I} [\sum_{j=1}^{N_*} \delta_{\xi^{i,j}} + \varepsilon_i \ \delta_{\xi^{i,0}}] \ .$$

En particulier

$$\langle \eta_N, \phi \rangle = \frac{1}{N} \ \sum_{i \in I} [\sum_{j=1}^{N_*} \phi(\xi^{i,j}) + \varepsilon_i \ \phi(\xi^{i,0})] \ ,$$

*Des variables aléatoires $(X_1, \cdots, X_n)$ sont dites *échangeables* si pour toute permutation σ de l'ensemble $\{1, \cdots, n\}$ la loi jointe de $(X_{\sigma(1)}, \cdots, X_{\sigma(n)})$ est identique à la loi jointe de $(X_1, \cdots, X_n)$, ou autrement dit la loi jointe de $(X_{\sigma(1)}, \cdots, X_{\sigma(n)})$ ne dépend pas de la permutation σ de l'ensemble $\{1, \cdots, n\}$. En particulier, les variables aléatoires $(X_1, \cdots, X_n)$ sont identiquement distribuées, mais elles ne sont pas nécessairement indépendantes.

et par différence

$$\begin{aligned}\langle \eta_N - \eta, \phi \rangle \;&=\; \frac{1}{N}\;\sum_{i\in I}\sum_{j=1}^{N_*}[\phi(\xi^{i,j}) - \langle m_i, \phi\rangle] + \frac{1}{N}\;\sum_{i\in I}[\varepsilon_i\;\phi(\xi^{i,0}) - q_*\;\langle m_i, \phi\rangle] \\ &=\; \frac{1}{N}\;\sum_{i\in I}\sum_{j=1}^{N_*}[\phi(\xi^{i,j}) - \langle m_i, \phi\rangle] + \frac{1}{N}\;\sum_{i\in I}\varepsilon_i\;[\phi(\xi^{i,0}) - \langle m_i, \phi\rangle] \\ &\qquad\qquad + \frac{1}{N}\;\sum_{i\in I}(\varepsilon_i - q_*)\;\langle m_i, \phi\rangle\;,\end{aligned}$$

pour toute fonction mesurable bornée ϕ.

(v) **Par définition, les variables aléatoires $(\varepsilon_i\,,\, i \in I)$ sont échangeables, à valeurs binaires 0 ou 1, et vérifient la contrainte $(\star)$. Montrer que nécessairement**

$$\mathbb{E}[\varepsilon_i] = q_* \qquad \textbf{et} \qquad \mathbb{E}\,|\varepsilon_i - q_*|^2 = q_*\;(1-q_*)\;,$$

pour tout $i \in I$, et

$$\mathbb{E}[\,(\varepsilon_i - q_*)\;(\varepsilon_j - q_*)\,] = -\frac{q_*\;(1-q_*)}{|I|-1} \leq 0\;,$$

pour tout $i, j \in I$ tel que $i \neq j$.

(vi) **Montrer que**

$$\mathbb{E}\langle \eta_N, \phi\rangle = \langle \eta, \phi\rangle$$

pour toute fonction mesurable bornée ϕ, c'est–à–dire que l'approximation est sans biais.

(vii) **En exploitant l'indépendance des différentes variables aléatoires, en conditionnant par rapport à la tribu $\mathcal{E}_N$ engendrée par les variables aléatoires $(\varepsilon_1, \cdots, \varepsilon_N)$, et compte tenu que $\varepsilon_i^2 = \varepsilon_i$ pour tout $i \in I$, montrer que**

$$\begin{aligned}\mathbb{E}[\,|\,\langle \eta_N - \eta, \phi\rangle\,|^2 \mid \mathcal{E}_N] \;&=\; \frac{1}{N^2}\;\sum_{i\in I}(N_* + \varepsilon_i)\;\mathrm{var}(\phi, m_i) \\ &\qquad + \frac{1}{N^2}\;\sum_{i,j\in I}(\varepsilon_i - q_*)\;(\varepsilon_j - q_*)\;\langle m_i, \phi\rangle\;\langle m_j, \phi\rangle\;,\end{aligned}$$

pour toute fonction mesurable bornée ϕ.

Clairement, la fonction $x \mapsto x - \lfloor x \rfloor$ définie sur $[0, \infty)$ et à valeurs dans $[0, 1)$ est périodique de période 1 et discontinue pour les valeurs entières de x, mais la fonction $x \mapsto c(x) = (x - \lfloor x \rfloor)\,(1 - (x - \lfloor x \rfloor))$ définie sur $[0, \infty)$ et à valeurs dans $[0, \frac{1}{4}]$ est périodique de période 1 et continue, et on vérifie que

$$q_* = \frac{N}{|I|} - \lfloor \frac{N}{|I|} \rfloor \qquad \text{et} \qquad q_* \,(1 - q_*) = c(\frac{N}{|I|}) \ .$$

(viii) **En utilisant les identités établies à la question (v), montrer que**

$$N \,\mathbb{E} \,|\, \langle \eta_N - \eta, \phi \rangle \,|^2 = \frac{1}{|I|} \sum_{i \in I} \mathrm{var}(\phi, m_i)$$
$$+ \frac{1}{N} \, c(\frac{N}{|I|}) \, [\sum_{i \in I} |\, \langle m_i, \phi \rangle \,|^2 - \frac{1}{|I| - 1} \sum_{i,j \in I, i \neq j} \langle m_i, \phi \rangle \, \langle m_j, \phi \rangle \,] \ ,$$

pour toute fonction mesurable bornée ϕ.

(ix) **En utilisant l'identité**

$$\sum_{i,j \in I, i \neq j} \langle m_i, \phi \rangle \, \langle m_j, \phi \rangle = |\, \sum_{i \in I} \langle m_i, \phi \rangle \,|^2 - \sum_{i \in I} |\, \langle m_i, \phi \rangle \,|^2 \ ,$$

montrer que

$$\sum_{i \in I} |\, \langle m_i, \phi \rangle \,|^2 - \frac{1}{|I| - 1} \sum_{i,j \in I, i \neq j} \langle m_i, \phi \rangle \, \langle m_j, \phi \rangle = \frac{|I|^2}{|I| - 1} \, W_N$$

avec

$$W_N = \frac{1}{|I|} \sum_{i \in I} |\, \langle m_i, \phi \rangle \,|^2 - |\, \frac{1}{|I|} \sum_{i \in I} \langle m_i, \phi \rangle \,|^2 \ ,$$

pour toute fonction mesurable bornée ϕ.

(x) **En déduire que la variance de l'erreur d'estimation pour l'algorithme d'échantillonnage sans remise vérifie**

$$N \,\mathbb{E} \,|\, \langle \eta_N - \eta, \phi \rangle \,|^2 = \mathrm{var}(\phi, \eta) - [\, 1 - \frac{|I|}{|I| - 1} \, \frac{|I|}{N} \, c(\frac{N}{|I|}) \,] \, W_N \ ,$$

avec

$$W_N = \frac{1}{|I|} \sum_{i \in I} |\, \langle m_i, \phi \rangle \,|^2 - |\, \frac{1}{|I|} \sum_{i \in I} \langle m_i, \phi \rangle \,|^2 \ ,$$

pour toute fonction mesurable bornée ϕ. Comment cette variance se compare–t–elle avec la variance de l'erreur d'estimation pour l'algorithme d'échantillonnage multinomial ?

Variance asymptotique pour les variantes de l'algorithme SIR

On considère les distributions normalisées définies par les relations récurrentes

$$\mu_{k-1} \longrightarrow \eta_k = \mu_{k-1} Q_k \longrightarrow \mu_k = g_k \cdot \eta_k$$

avec la condition initiale $\mu_0 = g_0 \cdot \eta_0$, où la notation $\cdot$ désigne le produit projectif. Le problème est caractérisé par

- la distribution de probabilité initiale η_0,
- les noyaux de probabilités de transition Q_k, pour tout $k = 1, \cdots, n$,
- les fonctions de sélection binaires g_k à valeurs 0 ou 1, pour tout $k = 0, 1, \cdots, n$.

On cherche une approximation de la distribution normalisée μ_k sous la forme d'une distribution empirique pondérée de la forme

$$\mu_k^N = \sum_{i=1}^{N} w_k^i \, \delta_{\xi_k^i} \ ,$$

avec les poids

$$w_k^i = \begin{cases} 0 \ , & \text{if } i \notin I_k^N \\ \dfrac{1}{|I_k^N|} \ , & \text{if } i \in I_k^N \end{cases}$$

et avec

$$I_k^N = \{ i = 1, \cdots, N \ : \ g_k(\xi_k^i) \neq 0 \} \ ,$$

de sorte que, en pratique

$$\mu_k^N = \frac{1}{|I_k^N|} \sum_{i \in I_k^N} \delta_{\xi_k^i} \ .$$

Clairement, la distribution de probabilité

$$\mu_{k-1}^N Q_k = \frac{1}{|I_{k-1}^N|} \sum_{i \in I_{k-1}^N} m_k^i \qquad \text{avec} \qquad m_k^i(dx') = Q_k(\xi_{k-1}^i, dx')$$

apparaît seulement sous la forme d'un mélange fini, et on se propose d'étudier son approximation par une combinaison linéaire convexe η_k^N de N masses de Dirac en utilisant

- soit le rééchantillonnage résiduel multinomial,
- ou bien le rééchantillonnage résiduel sans remise.

On a vu en cours que l'approximation particulaire avec rééchantillonnage résiduel multinomial converge dans $\mathbb{L}^1$ (et en fait dans tous les espaces $\mathbb{L}^p$) à la vitesse $1/\sqrt{N}$. On admet que cette propriété est également vraie pour l'approximation particulaire avec rééchantillonnage résiduel sans remise.

(xi) **Donner l'expression de la variance des moyennes intra–composantes affectées de poids uniformes, définie à la question (i).**

(xii) **En utilisant les expressions obtenues dans la première partie, étudier la limite de la variance conditionnelle**

$$\mathbb{E}[\,|\langle \eta_k^N - \mu_{k-1}^N Q_k, \phi\rangle|^2 \mid \mathcal{F}_{k-1}^N]\ ,$$

en probabilité quand $N \uparrow \infty$, pour l'approximation particulaire avec rééchantillonnage résiduel multinomial. Comment cette limite se compare–t–elle avec la limite de la variance conditionnelle pour l'approximation particulaire avec rééchantillonnage multinomial ?

(xiii) **En utilisant les expressions obtenues dans la deuxième partie, étudier la limite de la variance conditionnelle**

$$\mathbb{E}[\,|\langle \eta_k^N - \mu_{k-1}^N Q_k, \phi\rangle|^2 \mid \mathcal{F}_{k-1}^N]\ ,$$

en probabilité quand $N \uparrow \infty$, pour l'approximation particulaire avec rééchantillonnage résiduel sans remise. Comment cette limite se compare–t–elle avec la limite de la variance conditionnelle pour l'approximation particulaire avec rééchantillonnage multinomial, et avec la limite de la variance conditionnelle pour l'approximation particulaire avec rééchantillonnage résiduel multinomial ?