Flots stochastiques discrets

Elouan Argouarc'h[1,2], François Desbouvries[2], Éric Barat[1], Eiji Kawasaki[1], Thomas Dautremer[1]

[1]Université Paris Saclay, CEA, List
F-91120 Palaiseau, France

[2]Samovar, Telecom SudParis, Institut Polytechnique de Paris
F-91120 Palaiseau, France

{elouan.argouarc'h,eric.barat,eiji.kawasaki,thomas.dautremer}@cea.fr
{elouan.argouarch,francois.desbouvries}@telecom-sudparis.eu

Résumé − Dans cet article nous étendons les modèles de mélanges de Gaussiennes en un modèle probabiliste de mélange (les flots stochastiques discrets, FSD) dont les poids sont des fonctions flexibles définies par le biais de réseau de neurones. Nous montrons que les FSD peuvent être utilisés dans deux classes de problèmes variationnels : celui de l'estimation de densité, et celui de l'inférence variationnelle. À nombre de composantes fixé, le FSD aboutit à un gain de flexibilité important, que nous illustrons par un exemple bi-dimensionnel.

Abstract − In this paper, we propose an extension of Gaussian mixtures models where the constant weights are replaced by a flexible function defined via a Neural Network function. We then show that the corresponding models can be used to tackle two types of variational problems: density estimation and variational inference. We illustrate on a 2D example that this extension is able to capture much finer details that Gaussian mixtures, which is an empirical proof of an increased flexibility.

1 Introduction

Un mélange de Gaussiennes (MG) est une famille paramétrique de lois de probabilité dont la densité de probabilité (ddp) est donnée par $\psi(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)$. Cette famille est paramétrée par $\theta = \{\pi_1, ..., \pi_K,\ \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K\}$ tel que $\mu_k \in \mathbb{R}^d$ et $\Sigma_k \in \mathcal{M}_d(\mathbb{R})$ est une matrice de covariance. Le modèle graphique correspondant est donné par:

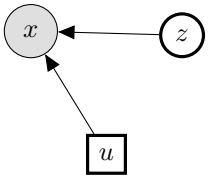


Fig. 1: Modèle graphique MG

où $Z \sim \mathcal{N}(0, \mathrm{I}_d)$ et où $U \sim \mathrm{Cat.}\big(\pi_1, ..., \pi_K\big)$. Ainsi, $\{\pi_1, ..., \pi_K\}$ est un vecteur de probabilités catégorielles tel que $\pi_k = \mathrm{p}\Big(x = \Sigma_k^{1/2} Z + \mu_k | Z = z\Big) \in \big[0, 1\big]$ ne dépend pas de z, et $\sum_{k=1}^{K} \pi_k = 1$. Dans ce papier nous explorons le modèle du FSD qui est une extension du MG et de plus grande expressivité. Le FSD bénéficie d'une fonction de densité calculable et d'un schéma d'échantillonnage explicite, ce qui le positionne comme une alternative au Flot Normalisant [9]. Nous expliquons ainsi comment utiliser un FSD pour résoudre les problèmes d'estimation de densité et d'inférence variationnelle.

2 FSD

Une des solutions possibles pour étendre un MG en un modèle plus expressif consiste à rendre dépendants de x les poids du mélange. On cherche donc à écrire un modèle de la forme $\psi(x) = \sum_{k=1}^{K} \pi_k(x) \mathcal{N}(x; \mu_k, \Sigma_k)$. La fonction ainsi définie est une ddp si $\pi_k(x) \geq 0$ et $\forall z \in \mathbb{R}^d$, $\sum_{k=1}^{K} \pi_k\Big(\Sigma_k^{1/2} z + \mu_k\Big) = 1$. Pour s'assurer que ces conditions sont vérifiées nous utilisons K fonctions paramétrées $\big\{w_1(z), ..., w_K(z)\big\}$ telles que $w_k(z) \geq 0$ et $\sum_{k=1}^{K} w_k(z) = 1$ pour tout $z \in \mathbb{R}^d$. Finalement

$$\psi(x) = \sum_{k=1}^{K} \underbrace{w_k\Big(\Sigma_k^{-1/2}\big(x - \mu_k\big)\Big)}_{\pi_k(x)} \mathcal{N}(x; \mu_k, \Sigma_k) \qquad (1)$$

est une ddp, dont le modèle graphique est donné par la figure 2 :

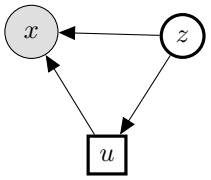


Fig. 2: Modèle graphique FSD

Cette construction aboutit à une variable aléatoire $X = \Sigma_U^{1/2} Z + \mu_U$ où $Z \sim \mathcal{N}\big(0, \mathrm{I}_d\big)$ et $U \sim \mathrm{Cat.}\big(w_1(Z), ..., w_K(Z)\big)$. Ainsi, à z fixé, $\big\{w_1(z), ..., w_K(z)\big\}$ est un vecteur de proba-

bilités catégorielles, mais où $w_k(z) = \mathbb{p}(X = \Sigma_k^{1/2} Z + \mu_k | Z = z)$ dépend de z. De manière évidente, si pour tout $k \in [1, ..., K], w_k(z) = \pi_k$ est une fonction constante, alors on retrouve un MG standard.

Enfin, nous proposons une paramétrisation simple et flexible des $w_k(z)$ grâce à un réseau de neurones. Du fait des contraintes identifiées, nous construisons $\{w_1(z), ..., w_K(z)\}$ comme la sortie d'un modèle de classification à K labels pour le vecteur z. Plus précisement, considérons un réseau de neurones avec L couches cachées telles que:

$$\begin{aligned} &h_1 = \sigma(W_0 z + b_0), \\ &h_{l+1} = \sigma(W_l h_l + b_l) \text{ pour } l = 1, ..., L-1, \\ &\left[\widetilde{w_1}(z), ..., \widetilde{w_K}(z)\right]^T = W_L h_L + b_L, \end{aligned}$$

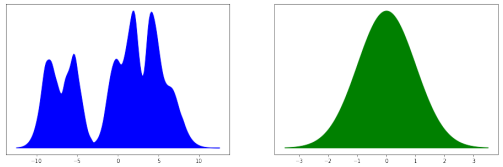
où $W_l \in \mathbb{R}^{n_{l+1} \times n_l}$ et $b_l \in \mathbb{R}^{n_{l+1}}$ pour $l = 0, ..., L$ (avec $n_0 = d$ et $n_{L+1} = K$) sont les paramètres de poids et de biais, et où $\sigma(.)$ est une fonction d'activation. Ce réseau calcule des poids non-normalisés, et pour obtenir un vecteur de probabilités catégorielles, il suffit d'appliquer une normalisation *Softmax* ce qui permet d'obtenir des valeurs positives et de somme 1 : $\left[w_1(z), ..., w_K(z)\right]^T = \textit{Softmax}(\left[\widetilde{w_1}(z), ..., \widetilde{w_K}(z)\right]^T)$. Finalement, cette construction permet de construire une famille de distributions de probabilité paramétrée par $\theta = \{W_0, b_0, ..., W_L, b_L, \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K\}$, dont la ddp (1) est calculable, et pour lesquelles on dispose d'une procédure simple pour obtenir des échantillons.

3 Mécanisme de dé- et re-construction d'une ddp

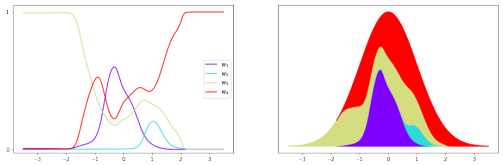
Dans cette section, nous illustrons le mécanisme du FSD sur un exemple unidimensionnel. En particulier, nous montrons comment une distribution Normale, illustrée sur la partie droite de la figure 3a, est découpée en différents éléments de masse, qui sont ensuite déplacés puis recombinés pour former une distribution complexe présentée à gauche de la figure 3a.

A gauche de la figure 3b sont tracées les fonctions de poids $w_k(z)$. Ces poids sont positifs et de somme 1 pour tout z et nous pouvons réécrire $\mathcal{N}(z; 0, \mathrm{I}_d) = \sum_{k=1}^K w_k(z)\, \mathcal{N}(z; 0, \mathrm{I}_d)$, donc ces fonctions induisent une partition de l'espace latent z. On peut alors voir que les $w_k(z)$ ont pour rôle de découper une distribution Normale en différents éléments de masse, ce qui est mis en évidence à droite de la figure 3b. Cette figure représente la distribution jointe (z, u) où la valeur de z est lue sur l'axe des abcisses et où les valeurs de $u = 1, ..., K$ sont représentées par les différentes couleurs. De manière équivalente, on représente les éléments de masse individuels avec la couleur correspondante sur la première ligne de la figure 3c.

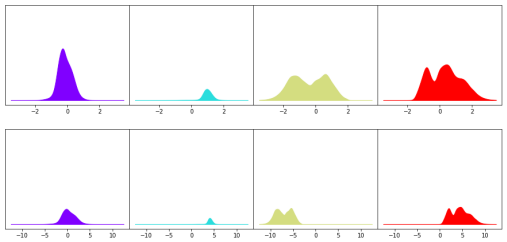
Par la suite, sachant $u = k$, un échantillon normal z est transporté par $x = \mu_k + \Sigma_k^{1/2} z$. Ainsi chaque élément de masse est envoyé dans une région différente de l'espace observé par la transformation linéaire inversible correspondante. Ce méchanisme est illustré par la deuxième ligne de la figure 3c. Finalement tous ces éléments sont recombinés pour former une distribution de probabilité Ψ. Nous affichons ainsi la ddp associée ψ sur la figure 3d.



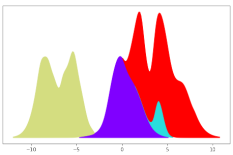
(a) Un FSD transforme une ddp normale (verte) en la ddp Ψ (bleue)



(b) Partition de l'espace latent avec les $w_k(z)$



(c) Eléments de masse transportés via $\Sigma_k^{1/2} z + \mu_k$



(d) Recombinaison

FIG. 3: Mécanisme FSD pour déconstruire/reconstruire $\mathcal{N}(0, \mathrm{I}_d)$ en Ψ

4 Estimation de Densité

4.1 Énoncé du problème

Considérons le problème d'estimation de densité : supposons que l'on dispose d'observations $x_1, ..., x_M$ d'une distribution de probabilité P dont on ne sait pas calculer la ddp $p(x)$. Pour obtenir une estimation de cette ddp, on peut considérer une approximation variationnelle $\Psi_{\theta^\star}$ obtenue par optimisation:

$$\theta^\star = \arg\min_\theta D_{\mathrm{KL}}\big(P || \Psi_\theta\big),$$

où Ψ_θ appartient à une famille paramétrée de distributions de probabilité telle que $\psi_\theta(x)$ soit calculable, et où

$D_{\mathrm{KL}}\big(P||\Psi_\theta\big) = \mathbb{E}_{X\sim P}\left[\log\left(\frac{p(X)}{\psi_\theta(X)}\right)\right]$ est la divergence de Kullback-Leibler [8] (D_{KL}) dans le sens *forward*. La D_{KL} n'est pas symétrique et, pour l'estimation de densité, ce choix spécifique est motivé par le fait que l'entropie de P ne dépend pas de θ et n'intervient pas dans l'optimisation. De plus, pour P arbitraire, cette divergence n'admet pas de forme analytique et donc ne peut pas être optimisée directement. On optimise plutôt une approximation Monte Carlo (MC) de cette divergence, qui est calculée en utilisant les observations de p. Le problème de minimisation de la D_{KL} se réduit donc à un problème de maximum de vraisemblance:

$$\theta^\star = \arg\max_\theta \sum_{i=1}^{M} \log\big(\psi_\theta(x_i)\big).$$

Sous la condition que la ddp ψ_θ est différentiable par rapport à θ, il est possible d'obtenir un minimum local de la log-vraisemblance par une approche basée sur le calcul du gradient.

4.2 Utilisation du FSD en estimation de densité

Par construction, la densité (1) d'un FSD est une fonction différentiable par rapport à ses paramètres, sous la condition que la fonction w_k soit différentiable (c'est le cas si l'on considère la paramétrisation par réseau de neurones proposée précédemment). Il est donc possible de calculer les gradients de la log-vraisemblance :

$$\sum_{i=1}^{M} \log\left(\sum_{k=1}^{K} w_k\Big(\Sigma_k^{-1/2}\big(x_i-\mu_k\big)\Big)\mathcal{N}(x_i;\mu_k,\Sigma_k)\right)$$

dans un environnement de différentiation automatique tel que Pytorch ou Tensorflow, et de procéder à une optimisation par gradients.

Sur la figure 4, on illustre tout d'abord la flexibilité du FSD comparativement à un MG dans le cas d'un problème d'estimation de densité. Alors, grâce à l'utilisation d'une fonction flexible de réseau de neurones qui définit les fonctions $\{w_1(z),...,w_K(z)\}$, le FSD peut capturer des variations beaucoup plus fines, et est ainsi un meilleur modèle pour P que ne l'est un MG. Sur cette figure, le modèle de FSD est obtenu en utilisant pour initialisation un MG obtenu par EM [4]; cette approche permet de rendre plus rapide la convergence du FSD vers P. Pour ce faire, il suffit d'initialiser $W_L = 0$ et $b_L = [\pi_1,...,\pi_K]^T$.

5 Inférence Variationnelle

5.1 Énoncé du problème

Considérons maintenant le problème d'inférence variationnelle [2][10][5][7] : supposons que l'on dispose d'une ddp

FIG. 4: Un FSD (3ème col.) peut approcher la distribution associée à une image (1ère col.) à partir d'échantillons (2ème col.) - comparé à un MG obtenu via EM (4ème col.).

$p(x)$ (possiblement non-normalisée), mais que l'on ne dispose pas de procédure d'échantillonnage simple. Une approche permettant d'obtenir des échantillons approximativement distribués selon P est de considérer une distribution variationnelle $\Psi_{\theta^\star}$ obtenue par optimisation :

$$\theta^\star = \arg\min_\theta D_{\mathrm{KL}}\big(\Psi_\theta||P\big),$$

où Ψ_θ appartient à une famille paramétrée de lois de probabilités telle que Ψ_θ puisse être échantillonnée facilement, et où $D_{\mathrm{KL}}\big(\Psi_\theta||P\big) = \mathbb{E}_{X\sim\Psi_\theta}\left[\log\left(\frac{\psi_\theta(X)}{p(X)}\right)\right]$ est la D_{KL} dans le sens *reverse*. Dans cette section, on suppose que $p(x)$ est différentiable. Pour $p(x)$ arbitraire, la D_{KL} n'admet pas de forme analytique et ne peut pas être calculée ni optimisée directement. Néanmoins, le choix du sens *reverse* permet d'utiliser des échantillons de Ψ_θ pour écrire et optimiser une approximation MC de cette divergence. Il est tout de même nécessaire de calculer le gradient par rapport à θ de cette approximation MC, ce qui peut se révéler complexe car la loi par rapport à laquelle est calculée l'espérance qui définit la D_{KL} dépend également de θ. Par conséquent, en considérant une approximation MC, les échantillons utilisés dépendent de θ, ce qui doit être pris en compte dans le calcul du gradient.

5.2 Un estimateur différentiable de la DKL par Rao-Blackwellisation (RB)

Si l'on écrivait une approximation MC naïve de la D_{KL} reverse, nous obtiendrions:

$$D_{\mathrm{KL}}\big(\Psi_\theta||P\big) \approx \frac{1}{M}\sum_{\substack{i=1\\ x_i\sim\psi_\theta}}^{M} \log\left(\frac{\psi_\theta(x_i)}{p(x_i)}\right); \qquad (2)$$

et, pour calculer le gradient de cette expression, il faudrait prendre en compte la dépendance en θ des échantillons $x_i \sim \Psi_\theta$. Une approche possible serait d'appliquer une reparamétrisation [6], c'est à dire réécrire les échantillons à l'aide d'une fonction inversible différentiable qui permet d'obtenir une expression où l'aléatoire et les paramètres sont découplés; ce qui revient à écrire $x_i = f^{-1}(y_i; \theta)$ où y_i est une réalisation d'une variable aléatoire qui ne dépend pas de θ, et f un C1-difféomorphisme. Dans le cas du FSD, trouver un tel changement de variable f n'est pas possible car l'échantillonnage de Ψ_θ fait intervenir une variable catégorielle discrète dont la fonction de répartition n'est pas différentiable. Finalement, il n'est pas possible de reparamétriser les échantillons d'un FSD de manière à en calculer les gradients. Pour répondre à ce problème nous proposons de construire une autre estimation MC basée sur le principe de la RB [3]. Considérons donc :

$$D_{\mathrm{KL}}\big(\Psi||P\big) \approx \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} w_k(z_i)\log\left(\frac{\psi\Big(\mu_k + \Sigma_k^{1/2} z_i\Big)}{p\Big(\mu_k + \Sigma_k^{1/2} z_i\Big)}\right), \tag{3}$$

où $z_1, ..., z_M \sim \mathcal{N}(0, \mathrm{I_d})$. Cet estimateur est différentiable par rapport aux paramètres θ, et peut être interprété comme le résultat d'une procédure de RB. En effet, soit J la variable aléatoire $\log\Big(\frac{\psi(X)}{p(X)}\Big)$ tel que $X \sim \Psi_\theta$ que l'on peut réécrire, par définition du FSD, comme :

$$J = \log\left(\frac{\psi\Big(\Sigma_U^{1/2} Z + \mu_U\Big)}{p\Big(\Sigma_U^{1/2} Z + \mu_U\Big)}\right),$$

où $z \sim \mathcal{N}(0, \mathrm{I}_d)$ et $u \sim \mathrm{Cat.}\big(w_1(z), ..., w_K(z)\big)$. Il est clair que $D_{\mathrm{KL}}(\Psi_\theta||P) = \mathbb{E}[J]$, où l'espérance est calculée par rapport à la loi jointe de (z, u). D'une part, échantilloner des couples $\big\{(z_i, u_i)\big\}_{i=1,...,M}$ aboutit à l'estimateur MC naïf (2). D'autre part, la RB est basée sur le fait que que $\mathbb{E}(J) = \mathbb{E}\Big(\mathbb{E}\big(J|Z\big)\Big)$ [1]. Mais, étant donné que l'espérance intérieure peut-être calculée explicitement :

$$\mathbb{E}\big(J|Z\big) = \sum_{k=1}^{K} w_k(Z)\log\left(\frac{\psi\Big(\Sigma_k^{1/2} Z + \mu_k\Big)}{p\Big(\Sigma_k^{1/2} Z + \mu_k\Big)}\right),$$

il est seulement nécessaire d'approximer par MC l'espérance extérieure en utilisant des échantillons $z_1, ..., z_M \sim \mathcal{N}(0, \mathrm{I}_d)$. Cette approche aboutit à l'approximation $D_{\mathrm{KL}}(\Psi||P) \approx \frac{1}{M}\mathbb{E}\big(J|Z = z_i\big)$, qui n'est autre que (3). Ainsi, utiliser une approximation basée sur la RB nous permet d'éviter le problème lié à la reparamétrisation de la variable catégorielle discrète U, et l'expression (3) est bien différentiable par rapport aux paramètres θ.

Conclusion

Dans ce papier, nous avons présenté le modèle du FSD qui est une famille paramétrée de distributions de probabilité. Ce modèle étend et inclut les MG et bénéficie d'une plus grande flexibilité ; il peut être paramétré simplement grâce à l'emploi d'une fonction de réseau de neurones définissant les poids du mélange. Nous montrons comment utiliser les FSD pour résoudre deux types de problèmes variationnels, et nous illustrons l'intérêt de ce modèle sur un exemple bidimensionel d'estimation de densité, dans lequel le FSD capture beaucoup plus de détails que le MG.

References

[1] David Blackwell. Conditional Expectation and Unbiased Sequential Estimation. *The Annals of Mathematical Statistics*, 18(1):105 – 110, 1947.

[2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[3] G. Casella and C.Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 03 1996.

[4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[5] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

[6] Diederik P Kingma and Max Welling. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, page 121, 2014.

[7] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[8] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[9] George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine learning research*, 22:1–64, 2019.

[10] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.