Notes spéculatives sur la concentration de la mesure

Yann Ollivier

11 mai 2002

1 Rappels

Soient $X_1, \ldots, X_n$ des espaces munis de mesures de probabilité $\mu_1, \ldots, \mu_n$. On pose $X = \prod X_i$ muni de la mesure produit μ. On définit sur X la distance de Hamming :

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \frac{1}{n} \left| \{i, x_i \neq y_i\} \right|$$

Avec cette distance, X est un espace de diamètre 1.

Le théorème de concentration de la mesure, dans sa forme la plus simple, s'énonce ainsi (Talagrand) :

Théorème 1 – *Soit $f : X \to \mathbb{R}$ une fonction k-lipschitzienne pour la distance de Hamming sur X. Alors pour tout $\varepsilon > 0$,*

$$\mu\left(\{x \in X, |f(x) - \mathbf{E}f| > \varepsilon\}\right) \leqslant 2 \exp -\frac{nt^2}{k^2}$$

où $\mathbf{E}f$ désigne la moyenne de f.

Démonstration cf Talagrand 96 Annals of Prob...

Ce théorème est une généralisation très forte de la loi des grands nombres, où $X = \{0,1\}^n$ et où f est le nombre de "1".

Évidemment, la condition d'être Lipschitz par rapport à la distance de Hamming est parfois restrictive. Quand les X_i sont finis ce n'est pas trop gênant, mais lorsque les X_i sont infinis, cela revient à demander que l'influence de chaque composante est bornée. On peut lever cette restriction en considérant pour X_i des espaces métriques, mais cela impose de demander que l'infini ne soit pas trop gros par rapport à la mesure.

Les conditions qu'on prend en général demandent que les moments de l'exponentielle de la fonction distance à un point soient contrôlés, i.e. par exemple

$\log \mathbf{E}(\exp \lambda \operatorname{dist}(x, x_0)) \leqslant C\lambda^2$. On dit d'un tel espace qu'il a une concentration gaussienne. De même que le théorème ci-dessus assure qu'un produit d'espaces finis a une concentration gaussienne (avec la bonne constante qui donne le diamètre observable), des théorèmes assurent qu'un produit d'espaces à concentration gaussienne est encore à concentration gaussienne avec la constante qu'on attend (diamètre observable environ égal à la racine de la somme des carrés des diamètres observables). Si on part d'espaces à concentration exponentielle et non gaussienne, leur produit est à concentration gaussienne à petite échelle, tandis qu'à grande échelle la même concentration exponentielle est conservée.

En résumé : la concentration gaussienne apparaît par produit d'espaces finis ; si on part d'espaces où le comportement à l'infini est à peu près contrôlé (exponentiel), leur produit fait apparaître de la concentration gaussienne à petite échelle (et conserve le même contrôle à grande échelle) ; enfin la concentration gaussienne est stable (et s'améliore) par produit.

On aimerait relier ces résultats à la distance "carrée" entre espaces métriques mesurés décrites par Gromov, un espace proche d'un espace concentré est encore concentré, mais on voudrait des théorèmes plus généraux de passage au produit (pas seulement pour de la concentration exponentielle aux grandes échelles).

La concentration sur le cube (et sur les produits) est clairement liée au fait que la contrainte d'être lipschitzien est forte, parce que beaucoup de chemins existent d'un point à un autre. De manière générale, sur les graphes où l'opérateur de la chaleur converge rapidement, on a de la concentration exponentielle (cela vaut en particulier pour les graphes aléatoires), sans doute aussi sur les grpahes de Cayley de groupes nilpotents (où on a des évaluations du trou spectral). On doit pouvoir démontrer de la sorte que les graphes aléatoires, dans certains modèles, sont concentrés.

2 Choix de la mesure produit, concentration de la concentration

Par exemple en physique statistique, si on a affaire à des particules indépendantes qui chacune peuvent se trouver parmi un ensemble fini d'états d'énergie avec certaines probabilités, la valeur de l'énergie totale se trouvera très concentrée autour de son espérance. Mais la valeur de l'espérance dépend, bien sûr, des probabilités qu'on prend. En physique statistique microcanonique, à défaut d'autre choix, on suppose en général que les particules se répartissent équiprobablement parmi les états disponibles. Ceci n'a évidemment de sens que si le nombre d'états accessible à chaque particule est fini.

On va montrer ici que ce choix arbitraire n'a pas grande conséquence, et qu'en fait, pour la plupart des choix des probabilités, on trouve la même valeur de l'espérance.

L'ensemble des mesures produits sur X est lui-même un espace produit : on se donne une mesure μ_1 sur X_1, etc. Notons $\mathcal{M}(Y)$ l'espace des mesures de masse 1 sur un espace Y.

L'ensemble des mesures produits sur X est simplement $\mathcal{M}(X_1) \times \ldots \times \mathcal{M}(X_n)$. Notons $\mathcal{M}_\times(X)$ cet ensemble.

Montrons que si $f : X \to \mathbb{R}$ est une fonction k-lipschitzienne pour la distance de Hamming sur X, alors la fonction

$$\begin{aligned} \mathbf{E} : \mathcal{M}_\times(X) &\to \mathbb{R} \\ \mu &\to \mathbf{E}_\mu f \end{aligned}$$

est une fonction k-lipschitzienne sur $\mathcal{M}_\times(X)$ pour la distance de Hamming.

Il suffit de montrer que si $\mu = (\mu_1, \mu_2, \ldots)$ et $\mu' = (\mu'_1, \mu_2, \ldots)$ sont deux mesures produits sur X, alors $|\mathbf{E}_\mu f - \mathbf{E}_{\mu'} f| \leqslant k$. Par linéarité (et densité ?) on peut se restreindre au cas où μ_1 et μ'_1 sont des Diracs δ_x, δ_y. Alors, notant $\nu = \mu_2 \times \mu_3 \times \ldots$:

$$\begin{aligned} \mathbf{E}_\mu f - \mathbf{E}_{\mu'} f &= \int_{z \in X_2 \times X_3 \times \ldots} (f(x,z) - f(y,z)) d\nu(z) \\ &\leqslant k \end{aligned}$$

Cela veut dire que la valeur autour de laquelle se concentre f sous une mesure produit est elle-même une fonction concentrée de la mesure produit.

Plus précisément, si sur chaque $\mathcal{M}(X_i)$ on s'est donné une mesure de probabilité, qu'on commence par choisir une mesure sur X_i selon cette loi, et qu'on tire ensuite un $x \in \prod X_i$ selon la loi obtenue, la valeur de $f(x)$ sera très concentrée autour d'une certaine valeur.

Si on tire la mesure produit μ suivant une certaine loi Λ, la valeur autour de laquelle cette concentration se produit est $\mathbf{E}_\Lambda \mathbf{E}_\mu f$ qui est par linéarité égale à $\mathbf{E}_{\mathbf{E}_\Lambda \mu} f$ (où $\mathbf{E}_\Lambda \mu$ est la loi moyenne sur X).

Ce résultat est plus fort que simplement la concentration de f autour de son espérance sous la loi $\mathbf{E}_\Lambda \mu$: on sait que pour beaucoup (sous Λ) de mesures μ, on aura concentration de f autour de la même valeur $\mathbf{E}_{\mathbf{E}_\Lambda \mu}$.

C'est évident lorsque tous les X_i sont égaux à $\{0, 1\}$ et que f est symétrique en tous les X_i : alors, on doit tirer des paramètres p_i au hasard dans $[0; 1]$, mettons uniformément, puis tirer des 0 ou des 1 ensuite, et il est intuitivement clair que cela revient directement à tirer des 0 et des 1 avec proba

1/2. C'est beaucoup moins clair lorsque les espaces ne sont pas égaux entre eux ou que la fonction f n'est pas symétrique.

On peut itérer le processus autant que l'on veut. À noter que le résultat final dépend du choix initial d'une mesure de proba sur l'espace des mesures de probas sur... Si on prend une mesure qui sort, au final, un Dirac sur un point de X, on n'obtiendra pas autre chose que la valeur de f en ce point.

Peut-on utiliser cette méthode, itérée, pour obtenir des mesures intéressantes sur l'espace de départ ? (Par exemple, des maxwelliennes ?)

3 Concentration sous la mesure de Wiener

Si on part de l'espace $\{0,1\}^n$, l'espace des mesures produits sur cet espace est $[0;1]^n$, et on peut prendre, par exemple, comme mesure produit sur cet espace la mesure uniforme ; mais on peut aussi itérer la construction une fois de plus et considérer l'espace $\mathcal{M}([0;1])^n$, se donner une mesure produit dessus, et voir que pour beaucoup de tirages d'une loi de proba sur $[0;1]^n$ selon cette mesure, puis pour beaucoup de tirages d'une loi de proba sur $\{0,1\}^n$ selon la loi qu'on vient de tirer, une une fonction f sur $\{0,1\}^n$ est concentrée autour d'une *même* valeur.

Se donner une mesure produit sur $\mathcal{M}([0;1])^n$, c'est-à-dire une famille de mesures sur $\mathcal{M}([0;1])$, ne va pas de soi... on peut par exemple imaginer de se restreindre aux mesures ayant une densité continue par rapport à la mesure uniforme sur $[0;1]$, auquel cas on doit se donner une mesure sur les fonctions continues sur $[0;1]$, par exemple une mesure de Wiener avec une certaine constante de diffusion (conditionnée à être positive et d'intégrale 1).

Moralement, la mesure de Wiener est aussi une mesure produit, la probabilité d'une fonction s'interprétant comme le produit des exponentielles des carrés de ses variations en chaque point.

Peut-on rendre rigoureux le fait que la mesure de Wiener est une mesure produit et donner un résultat de concentration de $\mathcal{M}([0;1])$ sous cette mesure ?

Qu'est-ce qu'une fonction(nelle) lipschitzienne dans ce cadre ? On peut déjà donner un exemple : la valeur d'une fonction en un point doit être une fonctionnelle sur $\mathcal{M}([0;1])$. Or il se trouve qu'elle est concentrée : prenant un mouvement brownien partant de 0, sa valeur en un point est précisément une gaussienne. Cela marche aussi pour des combinaisons de valeurs de fonctions, etc.

4 Conditionnement

On observe un phénomène général : quand on prend un sous-espace d'un espace concentré défini par des équations pas trop compliquées, alors on obtient encore un espace concentré.

Un exemple de ce phénomène est la concentration de la mesure sur la sphère. On commence par considérer $\mathbb{R}^n$ gaussien qui en tant qu'espace produit est concentré, et on considère la sphère de rayon $\sqrt{n}$ dedans. La masse de $\mathbb{R}^n$ gaussien est proche de cette sphère, la distance "carrée" de Gromov entre les deux est faible, et la sphère est concentrée.

Un autre exemple est le principe de conditionnement de Gibbs : on prend des variables indépendantes identiquement distribuées $X_1, \ldots, X_n$ à valeur dans un espace X. On note $P(X_i)$ la mesure empirique de la distribution des X_i dans X. Soit f une fonction (régulière) sur X, et soit $Y \subset \mathcal{M}(X)$ l'ensemble des mesures sur X sous lesquelles l'espérance de f est comprise dans un certain intervalle I. On peut considérer l'événement $P(X_i) \in Y$ et conditionner les X_i par cet événement (par exemple, imposer à un système de particules indépendantes d'avoir une certaine énergie), c'est-à-dire qu'on restreint la mesure au sous-espace X' de X^n où $P(X_i) \in Y$. Alors, on peut montrer que la loi des X_i conditionnée par X' est proche (au sens de la théorie de l'information) d'une loi produit sur X^n, cette loi produit étant celle dont la marginale est la loi de Y la plus proche (au même sens) de la loi de X_1. Étant proche d'un espace produit, le sous-espace X' est concentré.

Ces deux exemples ont en commun qu'un sous-espace d'un espace produit est proche (au sens distance carrée) d'un espace produit, lequel peut être simplement décrit à partir des équations donnant le sous-espace. Ainsi le conditionnement de $\{0,1\}^n$ par le fait que $\sum f(x_i)/n = A$ est (proche de) l'espace produit $\{0,1\}^n$ où on met sur chaque copie de $\{0,1\}$ l'unique mesure ν telle que $\mathbf{E}_\nu f = A$. De même, si on prend $\mathbb{R}^n$ avec la mesure uniforme et le sous-espace défini par $\sum x_i^2 = 1$, on a ainsi un espace proche d'un espace produit $\mathbb{R}^n$ où sur chaque copie de $\mathbb{R}$ on met une mesure ν telle que $\mathbf{E}_\nu x^2 = 1$: autrement dit, la mesure produit approximant le sous-espace se lit directement sur les équations définissant ce sous-espace, dans ces cas (cela est sans doute lié à la symétrie de l'équation en les composantes).

Autre exemple : la sphère est un espace concentré, de même que le plan projectif qui en est un quotient (la concentration est stable par un quotient dont toutes les fibres ont la même mesure), et Gromov a donné une méthode pour estimer la concentration sur une sous-variété de la sphère ou du projectif en fonction de la codimension, du volume riemannien et du degré (nombre maximum de points d'un équateur que la sous-variété peut rencontrer – cela impose de la régularité).

Les conditions générales sous lesquelles un sous-espace d'un espace produit (resp. concentré) est encore proche d'un espace produit (resp. est concentré) sont à ma connaissance fort peu claires.

5 Lien avec la théorie de l'information

Il existe des démonstrations du théorème central limite faisant appel au fait que les gaussiennes sont les lois qui, pour une variance fixée, maximisent l'entropie ; cela est lié au fait que sur un espace produit, parmi les mesures de marginales fixées, celle qui maximise l'entropie est la mesure produit.

Peut-on démontrer les inégalités de concentration en utilisant le même genre de principe : le fait de prendre une mesure produit ferait que la mesure d'une fonction lipschitzienne apporterait peu d'information ?

Dans le cas du conditionnement de $\mathbb{R}^n$ par $\sum x_i^2 = 1$, la mesure produit approximant la sphère est sur chaque composante la gaussienne, i.e. celle qui maximise l'entropie parmi les lois respectant la contrainte imposée.

Définitions sur des espaces métriques mesurés

Yann Ollivier

9 avril 2003

[Probablement bcp de fautes]

On se place sur l'espace $\mathcal{X}$ des espaces (réguliers) métriques probabilisés, avec la topologie de Hausdorff-Gromov, provenant d'une métrique de Hausdorff-Gromov à valeurs dans $\mathbb{R} \times \mathbb{R}$ qui dit que $\operatorname{dist}(X_1, X_2) \leqslant (\varepsilon, \kappa)$ si, en négligeant une part κ de la mesure de X_1 et X_2, il y a une bijection entre X_1 et X_2 préservant la mesure et modifiant les distances d'au plus ε.

Soit $f_{\kappa,\varepsilon} : \mathcal{X} \to \mathbb{R}$ une fonction dépendant de deux paramètres : une petite distance ε et une petite mesure κ (moralement, ce qui se passe sous cette échelle n'intervient pas dans f). On dit qu'une telle fonction est *régulière* s'il existe une constante C telle que pour tous $\eta > 0$, il existe (ε, κ) tels que si $\operatorname{dist}(X_1, X_2) \leqslant (\varepsilon, \kappa)$, pour tous $(\varepsilon_0, \kappa_0)$ on a

$$f_{\varepsilon_0,\kappa_0}(X_1) \leqslant \eta + \sup_{\substack{\varepsilon_0 - C\varepsilon \leqslant \varepsilon_1 \leqslant \varepsilon_0 + C\varepsilon \\ \kappa_0 - C\kappa \leqslant \kappa_1 \leqslant \kappa_0 + C\kappa}} f_{\varepsilon_1,\kappa_1}(X_2)$$

et

$$f_{\varepsilon_0,\kappa_0}(X_1) \geqslant -\eta + \inf_{\substack{\varepsilon_0 - C\varepsilon \leqslant \varepsilon_1 \leqslant \varepsilon_0 + C\varepsilon \\ \kappa_0 - C\kappa \leqslant \kappa_1 \leqslant \kappa_0 + C\kappa}} f_{\varepsilon_1,\kappa_1}(X_2)$$

[peut-on se passer de η ?]

Par exemple, le diamètre observable est une fonction régulière. La composée d'une fonction continue et d'une fonction régulière est régulière.

1 Entropies

On peut définir

$$H_{\varepsilon,\kappa}(X) = \inf_{(B_i)} \sum_i -\mu(B_i) \log \mu(B_i)$$

l'inf étant pris sur toutes les familles $(B_i) \subset X$ telles que les B_i sont deux à deux disjoints, avec $\operatorname{diam}(B_i) \leqslant \varepsilon$ et $\mu(\bigcup B_i) \geqslant 1 - \kappa$.

C'est une fonction régulière.

Motivation : sur un espace discret, l'entropie usuelle s'écrit $\sum_x -\mu(x)\log\mu(x)$; sur $\mathbb{R}$ avec une mesure μ absolument continue par rapport à la mesure de Lebesgue, l'entropie s'écrit $\int_x \mathrm{d}\mu(x)\log(d\mu(x)/dx) = \lim_{\varepsilon\to 0}\sum_i -\mu(I_i^\varepsilon)\log(\mu(I_i^\varepsilon)/\varepsilon)$ où les $(I_i^\varepsilon)_i$ forment une partition de $\mathbb{R}$ par des intervalles de taille ε ; sur $\mathbb{R}^2$ avec une mesure μ absolument continue par rapport à Lebesgue, cela donne $\lim_{\varepsilon\to 0}\sum_i -\mu(I_i^\varepsilon)\log(\mu(I_i^\varepsilon)/\varepsilon^2)$.

L'entropie est toujours une entropie d'une mesure par rapport à une autre (par rapport à la mesure de comptage dans un espace discret ou à la mesure de Lebesgue sur $\mathbb{R}$ ou $\mathbb{R}^2$...) laquelle n'est pas une mesure de probabilité mais une mesure "uniforme". Dans notre cas, l'entropie définie ci-dessus, sur un espace métrique mesuré, est prise par rapport à la "mesure de comptage" sur des partitions de diamètre ε : c'est la quantité d'information qu'il faut pour localiser un point de l'espace à ε près.

Le comportement quand ε tend vers 0 dépend de la dimension : par exemple, si X est une variété riemannienne de dimension n dont l'entropie ordinaire est h, on a

$$H_{\varepsilon,0}(X) \approx h + n\log 1/\varepsilon$$

On aimerait supprimer cette dépendance, en tenant compte de la dimension locale qu'on peut par exemple définir comme

$$d_x = \lim_{\varepsilon\to 0}\frac{\log\mu(B(x,\varepsilon))}{\log\varepsilon}$$

puis poser

$$\frac{d\mu(x)}{dx} = \lim_{\varepsilon\to 0}\frac{\mu(B(x,\varepsilon))}{\varepsilon^{d_x}}$$

et enfin

$$H = \int -d\mu(x)\log\frac{d\mu(x)}{dx}$$

mais la dimension d_x locale n'est pas du tout une fonction régulière sur $\mathcal{X}$ car elle dépend de la structure à très petite échelle. Cela marchera pour des espaces suffisamment réguliers.

Pour obtenir une fonction régulière, on peut remplacer d_x par une quantité dépendant de l'échelle ε comme

$$d'_{x,\varepsilon} = \frac{1}{\log 100}\log\frac{\mu(B(x,\varepsilon)}{\mu(B(x,\varepsilon/100))}$$

ce qui donne une fonction régulière (si on néglige aussi les ensembles de mesure $\leqslant \kappa$), mais le 100 est très arbitraire. On préfère donc garder le $H_{\varepsilon,\kappa}$ ci-dessus, en se souvenant que l'asymptotique en ε dépend de la "dimension".

2 Dimension statistique

La dimension n'étant pas une fonction régulière, par quoi peut-on la remplacer ? Le d'_x ci-dessus est un candidat possible, en voici d'autres.

Le diamète observable d'un produit de n espaces mesurés, ou d'une variété de dimension n de courbure $\geqslant 1$, est environ $1/\sqrt{n}$. Ceci donne envie de retrouver la dimension à partir du diamètre observable.

Soit $X \in \mathcal{X}$, on définit son rayon essentiel $EssR$ comme suit : c'est la médiane de la fonction distance de $X \times X$ dans $\mathbb{R}$. Si X est concentré, et qu'on tire deux points au hasard dans X, ils sont à distance $EssR(X)$. Le rayon observable de la sphère est $\pi/2$ en toute dimension. On peut aussi définir le diamètre essentiel comme $EssDiam_\kappa(X) = \inf\{\operatorname{diam} Y, Y \subset X, \mu(Y) \geqslant 1-\kappa\}$. Ces deux fonctions sont régulières.

Inspirés par l'exemple de la sphère, des espaces produits, des variétés à courbure positive, définissons la dimension statistique

$$StatDim(X) = \left(\frac{EssDiam(X)}{ObsDiam(X)}\right)^2$$

(on aurait pu remplacer EssDiam par EssR).

On a bien sûr introduit le diamètre essentiel, plutôt que de prendre directement le carré de l'inverse du diamètre observable, pour obtenir une quantité homogène en la métrique.

Cette quantité répond à la question : "si le phénomène que j'observe provient d'un grand nombre de contributions indépendantes, combien ?". Bien sûr, ceci est défini à des constantes près (puisque cela dépend de κ), seule l'asymptotique quand un paramètre tend vers l'infini est importante.

3 Théorèmes ?

On attend une inégalité entre dimension statistique et entropie : la dimension statistique est en gros le nombre d'informations élémentaires qu'il faut pour déterminer le résultat, l'entropie est la quantité d'information pour localiser un point...

Les variétés à courbure positive sont concentrées, et en regardant un dessin du cube $\{0,1\}^n$ pour n grand on perçoit une sorte de courbure positive. Peut-on définir une fonction régulière sur $\mathcal{X}$ qui rappelle la courbure, et, si oui, peut-on minorer la courbure qd l'espace est concentré ?

(Le lien concentration/espaces produits inspire plutôt à considérer de la commutativité que de la courbure positive. [Qu'est-ce qu'un mm-espace de courbure positive ?])

4 Espaces produits

Les espaces produits sont des espaces concentrés. La fonction de concentration est une fonction régulière sur $\mathcal{X}$ (un espace proche d'un espace concentré est concentré). Un quotient (image par une application 1-Lipschitz préservant la mesure) d'un espace concentré est un espace concentré.

Les espaces concentrés usuels proviennent, en probabilités, des espaces produits (plus des raffinements), et, en géométrie, de quotients de la sphère, qui est en fait proche d'un espace produit ($\mathbb{R}^n$ gaussien). Exception (?) : les variétés à courbure positive (cf. groupes de Lie ?). Mais dans certains cas au moins la courbure positive implique une structure pdrouit : pour les convexes, on peut (Dvoretzky) récupérer une section presque sphérique de grande dimension, i.e. une grande fraction de la dimension se comporte comme un produit.

Si on a un espace métrique concentré, peut-on l'écrire (à (ε, κ) près) comme un quotient d'un espace produit ? Suggestion de construction : prendre un point au hasard suivant la mesure, et considérer la fonction d_1 "distance à ce point" (qui est concentrée) comme une première coordonnée ; prendre un deuxième point, un troisième... ils sont tous à distance mutuelle $EssR$. Est-ce que les fonctions d_i distances à ces points sont à peu près indépendantes ? (quitte à les composer par une fonction 1-lipschitzienne), autrement dit est-ce que $\mu(d_1 \geqslant a, d_2 \geqslant b) \approx \mu(d_1 \geqslant a)\mu(d_2 \geqslant b)$?

Appliquée au cube $\{0,1\}^n$, cette construction donne une structure produit *différente de la structure standard* qui consiste à voir le cube en le tenant par deux points opposés (on voit qqch qui ressemble plus à une sphère) : (quitte à recentrer d_i autour de 0), la fonction d_1 peut être vue comme la distance (de Hamming) au point 0^n, ie le nombre de 1 ; la fonction d_2, comme la distance à $0^{n/2}1^{n/2}$ (on sait alors combien il y a de 1 à gauche et à droite) ; les fonctions d_3 et d_4, comme les distances à $0^{n/4}1^{n/4}0^{n/4}1^{n/4}$ et $0^{n/4}1^{n/4}1^{n/4}0^{n/4}$ respectivement, etc. Ces quantités sont à peu près indépendantes. Jusqu'où peut-on descendre en dimension ? Pour une raison entropique, devrait pouvoir descendre jusqu'à $n/\log n$ environ mais pas en-dessous, car chaque fonction d_1 fournit (à une constante près) une quantité d'information $\log n$ égale à l'entropie d'une gaussienne de variance $n/2$.