Physique statistique et apprentissage machine : une méthode et trois exemples

Rémi MONASSON

Laboratoire de Physique de l'Ecole Normale Supérieure, CNRS UMR8023 & PSL Research, 24, rue Lhomond, 75005 Paris, France

monasson@lpt.ens.fr

Résumé – Depuis les années 80, les physiciens statisticiens ont développé des outils et concepts permettant de comprendre quantitativement les propriétés de problèmes computationnels pertinents pour l'apprentissage machine. Ces méthodes sont particulièrement puissantes pour détecter et caractériser des changements abrupts de comportements, appelés transitions de phase. Le but de cet article est double. D'abord, je présente une des méthodes (dite des répliques) dans le cas simple de la transition de condensation. Ensuite, deux exemples d'application de cette méthode à des problème issus de l'apprentissage non supervisé sont brièvement exposés.

Abstract – Since the 80's, statistical physicists have developed tools and concepts allowing them to quantitatively understand the properties of computational problems relevant to the field of machine learning. These methods are particularly powerful to unveil and characterize abrupt changes of behaviors, called phase transitions. The scope of this article is two-fold. First, I present one of the methods (called replica method) in the simple case of the condensation phase transition. Secondly, two applications of the method to problems relevant to unsupervised learning are briefly exposed.

Au début des années 80, les concepts et méthodes de la physique statistique furent appliqués à d'autres domaines. Un exemple fameux fut l'invention de l'algorithme du recuit simulé par S. Kirkpatrick et ses collaborateurs [1], où l'introduction d'une température fictive permettait, lorsqu'elle était baissée progressivement, de miminiser une fonction de coût sans rester piégé dans des minima locaux. Quelques années plus tard, la résolution analytique du modèle de mémoire auto-associative de J. Hopfield marqua un tournant dans les applications interdisciplinaires de la physique statistique [2]. Par la suite, de nombreux travaux portèrent sur les problèmes d'optimisation [3] et d'apprentissage machine [4], avec une attention particulière pour les transitions de phases qui y survenaient [5].

Ce court article ne peut rendre justice à un domaine qui a donné lieu à des centaines de publications et connaît actuellement un renouveau certain. J'expose ici la méthode des répliques, non rigoureuse mais puissante, introduite par Mark Kac dans les années 60 et qui est maintenant un outil standard de la physique statistique, dans un cas relativement simple, la transition de condensation [6]. Je présente ensuite succinctement deux de ses applications dans le cadre de l'apprentissage non supervisé. Le lecteur désireux d'en savoir plus pourra trouver bien d'autres d'exemples et références dans les livres [4, 7] et articles de revues récents [5, 8].

1 Transition de condensation

Soit σ un paramètre réel positif et $\mathbf{W}$ une matrice symétrique aléatoire, dont les éléments hors-diagonaux sont des variables gaussiennes i.i.d. centrées en zéro et de variances σ^2/N (les éléments diagonaux sont nuls). On appelle $\lambda_a, \mathbf{e}_a$ les valeurs et vecteurs propres de $\mathbf{W}$, avec $\lambda_1 \geq ... \geq \lambda_N$ et $a = 1...N$.

A $\mathbf{W}$ fixée, considérons une variable aléatoire vectorielle $\mathbf{x}$ de dimension N, dont la distribution est gaussienne multivariée, de moyenne nulle et de covariance inverse $\mathbf{C}^{-1} = \mu\, \mathrm{I}_N - \mathbf{W}$. Ici, I_N est la matrice identité en dimension N et μ est choisi pour que la trace de $\mathbf{C}$ soit égale à N. Informellement, cela revient, lorsque N est très grand, à fixer la variance de chacune des composantes de $\mathbf{x}$ à la valeur unité. Il est facile de se rendre compte que les projections $x_a = \mathbf{x} \cdot \mathbf{e}_a$ du vecteur $\mathbf{x}$ sur les modes propres de $\mathbf{C}$ sont des variables normales, centrées en źero et de variances

$$v_a(\mu) = \frac{1}{\mu - \lambda_a} \; . \tag{1}$$

La condition de normalisation de la trace de $\mathbf{C}$ peut donc s'écrire comme une équation implicite sur μ :

$$1 = \frac{1}{N} \sum_{a=1}^{N} v_a(\mu) \; . \tag{2}$$

Cette équation et la contrainte $\mu > \lambda_1$ (due au fait que $\mathbf{C}$ doit être définie positive) déterminent μ de manière unique.

Que se passe-t-il lorsque N tend vers l'infini ? Si nous utilisons le fait que, lorsque N est grand, la densité des valeurs propres λ de $\mathbf{W}$ converge vers la loi du demi-cercle de Wigner sur l'intervalle $[-2\sigma; 2\sigma]$, l'equation (2) devient $1 = -s(\mu) = \frac{1}{2\sigma^2}\left(\mu - \sqrt{\mu - 4\sigma^2}\right)$, où s est la transformée de Stieljes de la loi du demi-cercle. Cette équation admet une unique solution $\mu = 1 + \sigma^2$. Bien sûr, ce résultat n'a de sens que si $\mu > \lambda_1$, c'est-à-dire si $\sigma < 1$.

Lorsque $\sigma \geq 1$, le calcul ci-dessus est faux. En effet, lorsque μ s'approche très près de $\lambda_1 = 2\sigma$, à une distance d'ordre $1/N$, la variance v_1 définie dans l'équation (1) devient d'ordre N et la contribution du terme $a = 1$ à la somme dans l'équation (2) est "macroscopique", c'est-à-dire qu'elle ne disparaît pas dans la limite $N \to \infty$. Géométriquement, cela veut dire que le vecteur aléatoire $\mathbf{x}$ est fortement aligné le long de la composante principale de $\mathbf{C}$: sa projection x_1 est d'ordre $\pm\sqrt{N}$, alors que les autres projections sont petites [1]. Nous en déduisons que, lorsque $\sigma \geq 1$, μ reste "bloqué" par la valeur propre λ_1 et l'équation (2) devient

$$1 = \frac{v_1}{N} - s\big(\mu = 2\sigma\big) \ , \quad \text{donc} \quad \frac{v_1}{N} = 1 - \frac{1}{\sigma} \ . \tag{3}$$

Le phénomène survenant à $\sigma = 1$ est un exemple de transition de phase, que l'on peut résumer informellement de la manière suivante. Dans la phase $\sigma < 1$ (appelée phase de haute température en physique), les vecteurs $\mathbf{x}$ sont libres d'occuper toutes les directions possibles de l'espace de dimension N, avec des projections suivant les vecteurs $\mathbf{e}_a$ certes d'autant plus grandes que a est petit mais qui restent d'ordre 1 lorsque N croît. Dans la phase $\sigma > 1$ (dite de basse température), les vecteurs $\mathbf{x}$ sont fortement alignés le long de $\mathbf{e}_1$, alors qu'ils peuvent fluctuer le long des directions transverses : on parle de transition de condensation.

2 Méthode des répliques

L'idée centrale de la méthode des répliques est la suivante. Dans le problème ci-dessus, nous avons deux types de variables aléatoires à considérer : la matrice $\mathbf{W}$ et les vecteurs $\mathbf{x}$. La première est fixée une fois pour toutes et définit une mesure sur les secondes. Du fait que plusieurs vecteurs $\mathbf{x}$ "voient" (c'est-à-dire sont soumis à) la même matrice $\mathbf{W}$, ils sont corrélés entre eux. Changer $\mathbf{W}$ en $\mathbf{W}'$ (issus du même ensemble statistique) ne détruira pas ces corrélations relatives. On peut donc moyenner sur $\mathbf{W}$ et les corrélations induites entre plusieurs $\mathbf{x}$ seront représentatives de la nature de la mesure induite par un $\mathbf{W}$ quelconque. Essayons maintenant de mettre cette idée en pratique.

Considérons la densité de probabilité d'un vecteur $\mathbf{x}_1$,

$$\rho(\mathbf{x}_1|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} \exp\left(-\frac{1}{2}\mathbf{x}_1^T(\mu\,\mathrm{I}_N - \mathbf{W})\mathbf{x}_1 \right) \ , \tag{4}$$

où

$$Z(\mathbf{W}) = \int d\mathbf{x} \, \exp\left(-\frac{1}{2}\mathbf{x}^T(\mu\,\mathrm{I}_N - \mathbf{W})\mathbf{x} \right) , \tag{5}$$

est un facteur de normalisation (qui pourrait s'exprimer simplement à l'aide des valeurs propres de $\mathbf{W}$). Cette densité de probabilité est elle-même aléatoire et nous voudrions calculer ses moments. Ceci est techniquement difficile car $\mathbf{W}$ apparaît, dans l'équation (4), à la fois au numérateur (dans l'exponentielle) et au dénominateur. Nous pouvons cependant écrire l'inverse de Z sous la forme Z^{n-1} lorsque $n \to 0$. Oublions un instant que n doit tendre vers 0, et traitons le comme un entier ≥ 1. Exprimons $Z(\mathbf{W})^{n-1}$ comme une intégrale sur $n-1$ vecteurs $\mathbf{x}_\alpha$, $\alpha = 2, ..., n$, appelés répliques,

$$Z(\mathbf{W})^{n-1} = \int \prod_{\alpha=2}^{n} d\mathbf{x}_\alpha \, \exp\left(-\frac{1}{2}\sum_{\alpha=2}^{n} \mathbf{x}_\alpha^T(\mu\,\mathrm{I}_N - \mathbf{W})\mathbf{x}_\alpha \right) . \tag{6}$$

Nous pouvons maintenant exprimer formellement les moments de ρ; un exemple trivial est que la moyenne de ρ sur $\mathbf{W}$, une fois intégrée sur $\mathbf{x}_1$, est égale à l'unité :

$$1 = \lim_{n\to 0} \int \prod_{\alpha=1}^{n} d\mathbf{x}_\alpha \, \exp\big(N\,\mathcal{M}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) \big) \ , \tag{7}$$

où

$$\begin{aligned} \mathcal{M} &= \frac{1}{N} \log \int D\mathbf{W} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{n} \mathbf{x}_\alpha^T(\mu\,\mathrm{I}_N - \mathbf{W})\mathbf{x}_\alpha \right) , \\ &= -\frac{\mu}{2}\sum_{\alpha=1}^{n} q_{\alpha,\alpha}(\mathbf{x}) + \frac{\sigma^2}{4}\sum_{\alpha,\beta=1}^{n} q_{\alpha,\beta}(\mathbf{x})^2 + O(1) \end{aligned} \tag{8}$$

lorsque $N \to \infty$, et où nous avons introduit les recouvrements (produit scalaires) entre paires de répliques

$$q_{\alpha,\beta}(\mathbf{x}) = \frac{1}{N}\,\mathbf{x}_\alpha^T\mathbf{x}_\beta \ . \tag{9}$$

Les formules (8) et (9) illustrent le principe général évoqué ci-dessus. Après avoir moyenné sur la matrice $\mathbf{W}$, les n répliques $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ interagissent entre elles : le logarithme de la densité de probabilité d'une configuration des n vecteurs, proportionnel à $\mathcal{M}$, n'est plus additif mais contient des termes de "couplages" (égaux à σ^2) entre toutes les paires de répliques α, β. Ces couplages produisent des corrélations entre répliques.

Il reste à calculer l'intégrale $n \times N$–dimensionnelle sur les $\mathbf{x}$ dans l'équation (7). Comme la log-densité $\mathcal{M}$ ne dépend que de la matrice des recouvrements $q_{\alpha,\beta}(\mathbf{x})$, nous allons procéder en deux temps, en (1) estimant l'intégrale sur les n-uplets de vecteurs $\mathbf{x}_\alpha$ à recouvrements fixés, puis en (2) intégrant sur ces derniers. A l'étape (1), pour N grand, nous trouvons [2]

$$\begin{aligned} \mathcal{I}(q_{\alpha,\beta}) &\equiv \frac{1}{N} \log \int \prod_{\alpha=1}^{n} d\mathbf{x}_\alpha \prod_{\alpha\leq\beta} \delta\Big(q_{\alpha,\beta} - q_{\alpha,\beta}(\mathbf{x}) \Big) \\ &= \frac{1}{2} \log\det\big(q_{\alpha,\beta}\big) + O(1) \ . \end{aligned} \tag{10}$$

L'intégrale sur les $q_{\alpha,\beta}$ à l'étape (2) se fait, lorsque $N \to \infty$, par la méthode du col sur les matrices de recouvrements. Au

1. Ce raisonnement intuitif est qualitativement correct mais doit être précisé. Du fait de l'écart d'ordre $N^{-2/3}$ (dûe à la singularité en racine carrée de la densité des valeurs propres au bord du spectre) entre λ_1 et λ_2, la variance v_2 est d'ordre $N^{2/3}$, et la projection x_2 d'ordre $N^{1/3}$ n'est donc pas vraiment petite ... mais le terme $a = 2$ (et a fortiori les termes $a \geq 3$) n'apporte pas de contribution finie à la somme lorsque $N \to \infty$.

2. On peut se convaincre sans calcul de ce résultat, qui coïncide avec l'intégrale de la gaussienne sur les n-uplets de vecteurs ayant les $q_{\alpha,\beta}$ pour matrice de covariance. Aucune autre mesure ne peut donner une intégrale plus grande, car la mesure gaussienne est la mesure d'entropie maximale à covariance fixée.

col, les éléments diagonaux de $q_{\alpha,\beta}$ valent 1 (par choix de μ) et dont les éléments hors-diagonaux sont solutions de

$$\frac{\partial}{\partial q_{\alpha,\beta}}(\mathcal{M}+\mathcal{I}) = \frac{\sigma^2}{2}\, q_{\beta,\alpha} + \frac{1}{2}(q^{-1})_{\beta,\alpha} = 0\,, \qquad (11)$$

pour $\alpha \neq \beta$. Comme toutes les répliques sont statistiquement équivalentes, il est naturel de chercher une solution de cette équation où $q_{\alpha,\beta} = q$ pour toute paire $\alpha \neq \beta$. Cette hypothèse, appelée symétrie des répliques, donne, une fois plongée dans l'équation (11), $\sigma^2\, q - \frac{q}{(1-q)(1+(n-1)q)} = 0$. Cette équation admet toujours $q = 0$ comme solution. Pour $n < 1$ et $\sigma > 1$, il existe une autre solution $q \in]0;1[$ (et c'est elle qui optimise $\mathcal{M}+\mathcal{I}$), qui tend vers $q = 1 - \frac{1}{\sigma}$ lorsque $n \to 0$.

Nous retrouvons donc le résultat de la section 1. Lorsque $\sigma < 1$, deux vecteurs $\mathbf{x}_1$ et $\mathbf{x}_2$ tirés indépendamment avec la distribution (4) sont statistiquement orthogonaux, et leur recouvrement q est nul. Lorsque $\sigma > 1$, ces deux vecteurs ont typiquement un recouvrement fini, qui correspond au carré de leur composante le long de $\mathbf{e}_1$, cf. équation (3). Nous voyons donc comment cette méthode permet de retrouver le comportement de notre modéle lorsque σ varie, en s'intéressant uniquement aux interactions effectives créées entre répliques suite à la moyenne sur les matrices $\mathbf{W}$. Bien sûr, la démarche n'est pas rigoureuse et la limite $n \to 0$ a été faite de manière acrobatique. Cependant, la validité des résultats obtenus par cette méthode sur des modèles non triviaux, comme le modèle de verre de spin de Sherrington et Kirkpatrick, ont pu être démontrés.

3 Inférence d'une direction en haute dimension et apprentissage retardé

Nous nous intéressons maintenant à l'inférence d'une direction particulière dans un espace de haute dimension N à partir d'exemples. Considérons la distribution sur les vecteurs $\mathbf{x}$ de densité

$$\rho(\mathbf{x}|\mathbf{B}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x} - V\left(\mathbf{x}^T\mathbf{B}\right)\right)\,, \qquad (12)$$

à un facteur de normalisation près. Le premier terme dans l'exponentielle définit une distribution gaussienne isotrope, alors que le deuxième terme biaise cette distribution le long d'un vecteur $\mathbf{B}$ normalisé. Un choix naturel pour ce biais est le "potentiel" quadratique

$$V(u) = -\frac{s}{1+s}\, u^2\,, \qquad (13)$$

où $s > 0$. Ce choix définit toujours une distribution gaussienne, avec une matrice de covariance ayant une valeur propre égale à $1+s$ (associé au vecteur propre $\mathbf{B}$) et les $N-1$ autres valeurs propres égales à l'unité : il s'agit du modèle dit de spiked covariance, qui a été beaucoup étudié dans la littérature [9]. D'autres choix, ne préservant pas la nature gaussienne de ρ, sont possibles, comme $V(u) = a\,u^2 + b\,u + c\,|u| + d\,u^3 + ...$, qui permet de jouer avec les puissances ou la valeur absolue de la projection de $\mathbf{x}$ sur $\mathbf{B}$.

Supposons que nous tirons P exemples, $\mathbf{x}_\mu$, avec $\mu = 1...P$, indépendamment les uns des autres selon la distribution ρ en équation (12). La distribution postérieure de la direction $\mathbf{B}'$ à inférer est donc

$$\hat{\rho}(\mathbf{B}'|\{\mathbf{x}_\mu\}) \propto \exp\left(-\sum_{\mu=1}^{P} V\left(\mathbf{x}_\mu^T\mathbf{B}'\right)\right)\,\delta\left(|\mathbf{B}'|-1\right)\,. \qquad (14)$$

Cette distribution postérieure fut étudiée par Reimann et al. avec l'aide de la méthode des répliques [10], dans la double limite $N, P \to \infty$ à rapport $\alpha = P/N$ fixé (qui est nombre d'exemples par degré de liberté à inférer). Une quantité importante qui émerge de cette étude est le recouvrement R entre l'espérance de $\mathbf{B}'$ selon la distribution postérieure de l'équation (14) et la "vraie" direction $\mathbf{B}$, moyenné sur les P-uplets $\{\mathbf{x}_\mu\}$ tirés au hasard selon $\rho(\mathbf{x}|\mathbf{B})$. Son comportement dépend de

$$\bar{u} = \int du\, e^{-u^2/2 - V(u)}\, u\,. \qquad (15)$$

Si $\bar{u} \neq 0$, le recouvrement $R(\alpha)$ est une fonction croissante de α, allant de 0 à 1 lorsque α passe de 0 à l'infini. Pour α grand, R s'approche de 1 avec des corrections algébriques en $1/\alpha$. Pour α petit, R se comporte comme une puissance de α, avec un exposant dépendant du comportement de $V(u)$ pour u petit. Ceci est relativement simple à comprendre : comme $\bar{u} \neq 0$, chaque exemple $\mathbf{x}_\mu$ est en moyenne aligné selon $\mathbf{B}$. Leur somme pointe dans la direction de $\mathbf{B}$, même si α est petit.

Si $\bar{u} = 0$, la situation est radicalement différente. Le recouvrement R est nul tant que α est plus petit qu'un rapport critique α_c, qui dépend du potentiel V. Pour $\alpha > \alpha_c$, R est une fonction croissante de α et tend vers 1 lorsque $\alpha \to \infty$. Ce phénomène, appelé "apprentissage retardé", fut découvert en toute généralité par Nadal et Watkin [11]. Un cas particulier est bien sûr le "spiked-covariance model", cf. équations (12) et (13). Reimann et al. ont effectué dans leur travail de 1996 une étude complète de ce cas : toutes leurs prédictions (en particulier l'expression de $R(\alpha)$ et celle de α_c en fonction de s et) ont été par la suite démontrées et sont connues dans la communauté mathématique sous le nom de transition BBP [12].

4 Nature des représentations apprises par les machines de Boltzmann restreintes

Une machine de Boltzmann restreinte (RBM) est un modèle génératif non-supervisé permettant d'apprendre une distribution sur un ensemble de données [13, 14]. Une RBM est définie sur un graphe bipartite connectant une couche visible comportant les unités v_i, $i = 1...N$, où sont présentées les données, et une couche cachée comportant les unités h_μ, $\mu = 1...M$, portant les représentations de ces données. La distribution jointe des configurations $\mathbf{v}$ et $\mathbf{h}$ des unités visibles et cachées est

$$\rho(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N} g_i(v_i) - \sum_{\mu=1}^{M}\mathcal{U}_\mu(h_\mu) + \sum_{i,\mu} h_\mu\, w_{i\mu}(v_i)\right) \qquad (16)$$

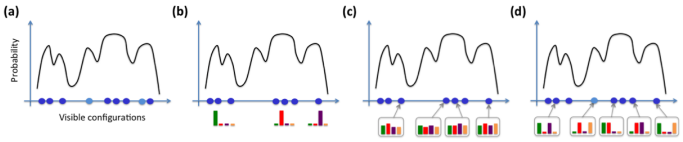

FIGURE 1 – **Nature des représentations avec une RBM. (a).** Les données (points en bleu foncé) forment un échantillonnage très sparse de la distribution (courbe noire) sur l'ensemble des configurations visibles ; beaucoup des configurations ayant une haute probabilité (points en bleu clair) ne sont pas dans les données. Régimes de représentations : **(b)** prototypes, **(c)** intriquées, **(d)** compositionnelles. Les barres de couleur correspondent aux activations h_μ des différentes unités cachées.

où les g_i et les $\mathcal{U}_\mu$ sont des potentiels locaux agissant sur les unités v_i (a priori catégorielles, par exemple, binaires : $v = 0, 1$) et h_μ (à valeurs réelles), et les poids $w_{i\mu}(v)$ couplent ces unités ; Z est telle que ρ est normalisée.

Dans la pratique, les paramètres de la RBM sont fixés pour maximiser la log-vraisemblance d'un ensemble de données $\mathbf{v}_k$, à savoir $L = \sum_k \log P(\mathbf{v}_k)$, où $P(\mathbf{v}) = \int d\mathbf{h}\, \rho(\mathbf{v}, \mathbf{h})$ est la distribution marginale des configurations visibles (Fig. 1(a)). Les potentiels sur les unités cachées sont paramétrisés pour définir des familles variationnelles relativement étendues. Un famille intéressante est $\mathcal{U}(h) = \frac{1}{2}\gamma^+(h^+)^2 + \frac{1}{2}\gamma^-(h^-)^2 + \theta^+ h^+ + \theta^- h^-$, où $h^+ = \max(h, 0)$, $h^- = \min(h, 0)$, qui englobe les unités de types Bernoulli, gaussiennes et rectifiées linéaires, qui sont toutes des choix populaires. Les paramètres $\gamma^\pm$, $\theta^\pm$ ainsi que les $g_i(v)$ et $w_{i\mu}(v)$ sont ensuite obtenus par montée de gradient de L [13].

Une question fondamentale est l'interprétation des paramètres inférés et, plus généralement, la manière dont les données sont représentées par la RBM. Pour ce, considérons un ensemble statistique de RBM dont les paramètres sont [14] : le rapport d'aspect $\alpha = \frac{M}{N}$, les seuils $\theta^\pm$ et courbures $\gamma^\pm$ des potentiels cachés $\mathcal{U}$, les potentiels $g_i(v) = g(v)$ agissant sur les unités visibles binaires. Les connexions $w_{i\mu}(v)$ sont tirées au sort, et valent 0, $+\frac{W}{\sqrt{N}}$ ou $-\frac{W}{\sqrt{N}}$ avec probabilités égales respectivement à $1 - p$, $\frac{p}{2}$ et $\frac{p}{2}$. Le paramètre p détermine la sparsité des poids : plus p est petit, moins il y a de poids (non nuls).

L'étude avec la méthode des répliques des RBM définies par cet ensemble statistique montre que trois types de représentations de $\mathbf{v}$ sont possibles, selon les paramètres ci-dessus :
1. Prototypes (Fig. 1(b)). L'espace des configurations visibles est partitionné en "régions", qui correspondent chacune à une unité cachée. Dans une région, cette unité est fortement activée ($h \sim \sqrt{N}$) alors que les autres sont faiblement activées ou éteintes. Les connexions w_i correspondantes sont un prototype des configurations visibles v_i correspondant à la région.
2. Intriquées (Fig. 1(c)). Presque toutes les unités cachées sont activées. Deux configurations visibles $\mathbf{v}$ et $\mathbf{v}'$ différentes correspondent à des activations $\mathbf{h}$ et $\mathbf{h}'$ légèrement différentes. Les connexions w ne peuvent pas être reliées simplement aux configurations visibles $\mathbf{v}$ ayant une grande probabilité marginale P.
3. Compositionnelles (Fig. 1(d)). Si p est suffisamment petit (beaucoup de poids nuls) et W suffisamment grand (les poids restants sont forts) un nombre substantiels d'unités cachées ($\sim \frac{1}{p}$, grand devant 1 et petit devant M) sont activées. La même unité cachée est active dans différentes régions de l'espace visible. Ses connexions définissent un motif sparse, et la composition combinatoire de multiples motifs permet de générer une très grande diversité de configurations $\mathbf{v}$. Ce régime compositionnel a beaucoup d'avantages : il permet de capturer des invariances de la distribution présentes dans des données éloignées les unes des autres ; les représentations sont sparses et plus facilement interprétables. En outre, il est présent dans les analyses de données réelles, comme les digits de MNIST [14] ou les séquences de protéines homologues [15].

Références

[1] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220, 671-680 (1983).

[2] J.J. Hopfield, Proc. Nat. Acad. Sci. (USA) 79, 2554 (1982).

[3] M. Mézard, G. Parisi and M. A. Virasoro, Spin Glass Theory and Beyond (World Scientific, Singapore, 1987)

[4] A. Engel and C. van den Broeck, Statistical Mechanics of Learning (Cambridge University Press, Cambridge, England, 2001)

[5] R. Monasson, *Introduction to Phase Transitions in Random Optimization Problems*, in Complex Systems, Les Houches, edited by J.-P. Bouchaud, M. Mézard, and J. Dalibard (Elsevier, New York, 2007), Vol. 85, pp. 1-65

[6] J.M. Kosterlitz, D.J. Thouless and R.C. Jones, Physical Review Letters 36, 1217 (1976)

[7] M. Mézard and A. Montanari, Information, Physics, and Computation (Oxford University Press, New York, 2009)

[8] L. Zdeborova and F. Krzakala, Advances in Physics 65, 453–552 (2016)

[9] I.M. Johnstone, in Proceedings of the ICM, Madrid, Spain, pp. 307-333 (2006)

[10] P. Reimann and C. van den Broeck, Physical Review E 53, 3989-3998 (1996)

[11] T.L.H. Watkin and J.-P. Nadal, J. Phys. A 27, 1899 (1994)

[12] J. Baik, G. Ben Arous and S. Péché, Ann. Probab. 33, 1643-1697 (2005)

[13] A. Fischer and C. Igel, in Iberoamerican Congress on Pattern Recognition (Springer, 2012) pp. 14-36

[14] J. Tubiana and R. Monasson, Physical Review Letters 118, 138301 (2017)

[15] J. Tubiana, S. Cocco and R. Monasson, eLife 2019 ;8 :e39397 (2019)