MARC MÉZARD *directeur de l'École normale supérieure, ancien professeur à l'École polytechnique*

PHYSIQUE STATISTIQUE ET INFORMATION : LE DÉFI DES DONNÉES MASSIVES

Lorsqu'on imagine les frontières de la physique, viennent à l'esprit immédiatement l'infiniment grand et l'infiniment petit. On oublie parfois cette autre frontière aux ramifications pourtant extraordinaires, à la fois par la richesse de sa construction intellectuelle et par la portée de ses applications : l'infiniment complexe. Le traitement des données massives nous en offre un bel exemple.

DÈS LE DÉBUT DU VINGTIÈME SIÈCLE, les physiciens curieux s'étaient interrogés devant les différentes phases de la matière. Un glaçon dans un verre d'eau ? Ce ne sont que des molécules d'H_2O. Les interactions entre deux molécules sont toujours les mêmes, alors qu'est-ce qui amène ces molécules à parfois s'aligner pour construire un cristal bien ordonné, ou bien au contraire à diffuser dans l'eau ? La réponse est justement qu'il s'agit d'un comportement collectif. Vous ne pouvez ni l'observer ni le comprendre en étudiant cinq ou dix molécules. En revanche, lorsqu'il y en a un grand nombre il se produit un phénomène coopératif, purement statistique, qui les amène à s'aligner dans un ordre cristallin, à une température très bien définie et très précise, le zéro degré de l'échelle Celsius. Il a fallu attendre les travaux du physicien norvégien Lars Onsager en 1944, pour obtenir une analyse détaillée de ce phénomène pourtant d'expérience quotidienne : la transition de phase entre un solide et un liquide.

© KATELEIGH / FOTOLIA.COM

Un glaçon dans un verre d'eau : des molécules d'H_2O organisées différemment.

« Lorsqu'il y a un grand nombre de molécules, il se produit un phénomène coopératif, purement statistique »

LES SINGULARITÉS DE LA TRANSITION DE PHASE

Depuis une quarantaine d'années, l'attention des physiciens s'est déplacée vers une autre frontière, celle des transitions de phases dans les systèmes très désordonnés, les verres. Un verre est en effet un état solide (au moins sur des échelles de temps pas trop longues), et pourtant les molécules du verre ne s'alignent pas dans un ordre cristallin, mais se figent dans des positions aléatoires, un peu comme si on restreignait les mouvements moléculaires d'un liquide, ne laissant plus à chaque molécule que la possibilité de petites vibrations dans la cage créée par ses voisines, en supprimant le mouvement collectif qui permet à l'eau de couler. En avançant dans la compréhension de ces phases vitreuses, les physiciens et mathématiciens ont développé un impressionnant corpus de concepts et de méthodes. Pour saisir la difficulté, on peut transposer

REPÈRES

Le domaine de la physique statistique d'aujourd'hui est celui qui explore les phénomènes « émergents », les comportements collectifs nouveaux qui apparaissent lorsqu'un système est composé d'un grand nombre de particules. Le prix Nobel Philip Anderson l'avait résumé dans le titre d'un de ses articles célèbres : *More is different*.

le problème en étudiant des comportements émergents d'agents qui décident d'acheter ou vendre sur un marché financier. Si tous les agents suivent la même stratégie, l'analyse est assez simple : on considère un « agent représentatif », et l'effet des autres agents sur le résultat de sa propre stratégie est pris en compte en se disant que, justement, les autres agents sont aussi dans le même cas que lui. Une relation d'autocohérence permet alors de comprendre le fonctionnement du marché. En physique, c'est la théorie du champ moyen, inventée par Pierre Weiss en 1907, qui rend compte ainsi des transitions de phases ordinaires. Dans les systèmes désordonnés c'est une tout autre affaire : chaque agent suit sa propre stratégie, on ne peut pas étudier un seul agent représentatif, on doit développer une étude statistique car le comportement moyen d'un individu n'apporte pas d'information suffisante. C'est ce qui a été fait pour l'étude des verres de spin (des systèmes magnétiques présentant des phases vitreuses), et les méthodes développées dans ce contexte ont trouvé des applications dans des domaines très variés, de la finance à la biologie en passant par l'informatique. La théorie des verres de spin, motivée à l'origine par des comportements anormaux comme le vieillissement de ces matériaux dont on n'a trouvé aucune application, développée par pure curiosité intellectuelle, est souvent comparée à une corne d'abondance.

CC

Pierre Weiss (1865-1940) invente en 1907 la théorie du champ moyen.

LA THÉORIE DE L'INFORMATION, NOUVEAU CHAMP DE RECHERCHE

Un des champs de recherche majeurs où l'on rencontre ces phénomènes d'émergence et de transitions de phases est la théorie de l'information. Fondée par Claude Shannon en 1948, avec en tête des questions très concrètes de télécommunications, elle s'est transformée désormais en un champ scientifique fertile et profond, aux ramifications actuelles essentielles dans de nombreux domaines où interviennent les notions de stockage, de transfert et de manipulation d'information. Que l'on songe par exemple au transfert d'information de la rétine au cerveau et à sa manipulation entre les différentes aires du cortex visuel pour extraire d'une image un contenu, ou bien, toujours en biologie, à la transmission d'information génétique d'une cellule à ses filles lors de la division cellulaire. Ce domaine scientifique, qui traverse les disciplines, devient d'importance capitale de nos jours avec l'arrivée de données massives et la quête de bons algorithmes pour en extraire l'information pertinente *via* les techniques d'apprentissage machine.

CC

Claude Shannon, né en 1916, est le père de la théorie de l'information.

BRUIT ET INTÉGRITÉ DE L'INFORMATION

Comme premier exemple penchons-nous sur un des sujets mis en avant par Shannon lui-même : l'utilisation des codes de correction d'erreurs pour transmettre de l'information. Dès que le canal de transmission est bruité, ce qui est toujours le cas, qu'il s'agisse de récupérer des données mesurées par un satellite, de téléphoner, de stocker un document sur son disque dur ou tout simplement de parler à quelqu'un, pour que la personne qui reçoit le message puisse comprendre le message qui lui est transmis, il faut utiliser ce qu'on appelle un code correcteur, c'est-à-dire en pratique envoyer un message redondant. C'est le cas lorsque je v**s p*rle ou l*rsqu* j'é*ris : le langage est redondant (on peut montrer qu'il contient à peine plus d'un

« Dès que le canal de transmission est bruité, il faut utiliser un code correcteur »

bit d'information par lettre envoyée, très en dessous des 4,7 bits par lettre correspondant aux possibilités offertes par les combinaisons des 26 lettres de l'alphabet si on n'utilisait pas les seuls mots du dictionnaire), et cette redondance vous permet de corriger des fautes de prononciation ou des fautes de frappe. Pour comprendre le mécanisme du codage par redondance, on peut considérer le code le plus simple, dit par répétition : pour transmettre le « mot » 1101, on envoie trois fois chaque bit, donc 111111000111. Avec 10 % de bruit sur la ligne on va altérer un bit sur dix, et recevoir par exemple 110111010111 ; le décodage (par majorité au sein de chaque triplet) est évident : dans ce cas il permet de retrouver l'information transmise, au prix d'une redondance d'un facteur 3. Mais un tel code, dit code par répétition, ne peut pas corriger toutes les erreurs : il reste toujours une certaine probabilité que le code ne restitue pas le mot d'origine, c'est le cas si le bruit a retourné deux bits d'un même triplet.

SUDOKU GÉANT

La grande découverte de Shannon est qu'il existe des codes qui savent corriger toutes les erreurs, lorsque le niveau de bruit est inférieur à un seuil critique. La construction des meilleurs codes de correction d'erreurs, permettant d'atteindre les performances idéales prévues par Shannon, a énormément progressé ces dernières années. Ces codes sont fondés sur une redondance collective impliquant un grand nombre de bits à la fois. En un sens, le « jeu » du décodage s'apparente à un sudoku géant : il faut retrouver le signal envoyé, à partir d'une version bruitée de ce signal qu'on a reçue, en utilisant la connaissance du code, qui se traduit en une série de contraintes reliant les bits envoyés. La grande difficulté était de trouver des algorithmes de décodage efficaces, alors que la dynamique à l'œuvre dans ces algorithmes, manipulant beaucoup de bits à la fois, est justement sujette à des phénomènes émergents et à des transitions de phases qui limitent leurs performances. C'est désormais chose faite, et certaines catégories de codes atteignent des performances proches du seuil idéal de Shannon, grâce à un ensemble d'idées qui sont intimement liées à la physique des verres de spin : le rôle des molécules est joué par les bits d'information, leurs interactions sont les contraintes du code, et on cherche à accélérer une dynamique qui est celle de l'algorithme de décodage.

« L'acquisition comprimée est devenue un thème majeur en traitement de signal »

COMPRESSION DE L'INFORMATION

Un autre exemple récent de cette fertilisation croisée entre disciplines est celui de l'acquisition comprimée en traitement du signal. Chacun sait désormais que l'information peut être comprimée. Pour stocker une image, on peut décider d'utiliser un algorithme de compression qui certes fera perdre un peu de résolution, si l'on cherche un jour à zoomer sur un détail, mais qui permettra de garder l'essentiel, en économisant l'espace mémoire. Dans une base appropriée, l'image est « parcimonieuse » : un bon nombre de ses composantes sont presque nulles (par exemple, dans une transformée en ondelettes, on utilise la somme et la différence de deux pixels voisins, et cette dernière est souvent très petite, dès lors que les deux pixels sont dans une même zone presque uniforme).

© JULIASUDNITSKAYA / FOTOLIA.COM

Le « jeu » du décodage s'apparente à un sudoku géant.

© SFAM_PHOTO / SHUTTERSTOCK.COM

On peut espérer acquérir une image de RMN en accélérant le processus d'un facteur 3 ou 4.

COMPRESSION DU SIGNAL

Peut-on exploiter cette propriété pour acquérir le signal en faisant un nombre de mesures minimal ? Pour une photo, cela reviendrait à se dispenser de la double étape : mesure de chaque pixel, puis compression informatique, et la remplacer par une acquisition du signal directement sous forme comprimée. Si cela présente peu d'intérêt pour nos photos de famille, on peut en revanche espérer acquérir une image de RMN en accélérant le processus d'un facteur 3 ou 4, augmentant d'autant l'efficacité de nos dispositifs d'imagerie médicale, ou bien diminuer l'irradiation lors de l'examen par microscopie de molécules biologiques. Rien de surprenant à ce que l'acquisition comprimée soit devenue un thème majeur en traitement de signal dans la dernière décennie.

PLUS DE VARIABLES QUE D'ÉQUATIONS

À première vue, le problème de l'acquisition comprimée peut sembler insurmontable : il s'agit en effet de reconstituer un signal en faisant un nombre de mesures inférieur au nombre d'inconnues. Dans le cas très étudié où les mesures sont des combinaisons linéaires des composantes du signal, on a donc affaire à un système linéaire mal conditionné : il comporte moins d'équations (les mesures) que de variables (les composantes du signal). Comment faire dans ce cas ? En cherchant une solution où un certain nombre de variables sont nulles.

UN INTENSE TRAVAIL THÉORIQUE

La théorie de l'information nous enseigne que, lorsqu'on mesure des signaux parcimonieux, dès lors que le nombre de mesures est supérieur au nombre de variables non nulles, il existe en principe une solution permettant de retrouver le signal d'origine. Reste à trouver une méthode pour atteindre cet objectif. L'intense activité théorique sur ce sujet a en effet porté ses fruits. Pour réussir à reconstruire un signal à partir d'un petit nombre de données, il faut avant tout que chaque mesure porte sur un grand nombre de composantes du signal (au lieu de regarder une image pixel par pixel, on utilisera un verre dépoli), et on doit ensuite utiliser un puissant algorithme de reconstruction. Dans une des approches les plus récentes, on cherche le signal comme la combinaison la plus probable, prenant en compte les contraintes dues aux mesures et favorisant la parcimonie.

« Pour réussir à reconstruire un signal à partir d'un petit nombre de données, on doit utiliser un puissant algorithme de reconstruction »

ANALOGIE AVEC LES ÉTATS CRISTALLINS

Le signal d'origine, celui que l'on cherche à reconstruire, est bien le plus probable. Mais, dans un problème typique qui peut comporter des centaines de milliers voire des millions de variables, le retrouver n'a rien de simple. C'est en fait comme trouver un état « cristallin » dans un système physique : le rôle de la position des atomes est ici joué par les composantes du signal, les interactions entre atomes sont les contraintes dues aux mesures ; il existe un état cristallin optimum, et pour le trouver on peut utiliser des algorithmes de champ moyen, notre fameuse statistique des « agents », appliquée ici aux composantes du signal. L'analogie avec le système physique va très loin. En effet, l'application directe de ces principes se heurte à l'existence d'une phase vitreuse : l'algorithme n'arrive pas à reconstruire l'ensemble du signal, au lieu de trouver un cristal sa dynamique

SÉRENDIPITÉ

Lorsqu'ils réfléchissaient aux verres de spin il y a trente ans, jamais les physiciens n'auraient pensé que leur savante théorie des systèmes désordonnés s'appliquerait un jour à des problèmes majeurs de théorie de l'information, où la dynamique vitreuse n'est pas liée au mouvement des atomes mais à des algorithmes de reconstruction de signaux. Bel exemple de ce qu'on appelle la sérendipité.

se ralentit et il se fige dans un état vitreux éloigné du signal d'origine. Il faut alors appliquer un autre principe physique : en utilisant une matrice de mesure structurée, l'algorithme parvient à créer un germe de cristal à partir duquel se propage une onde de cristallisation, reconstruisant le signal désiré, jusqu'au seuil prédit par Shannon.

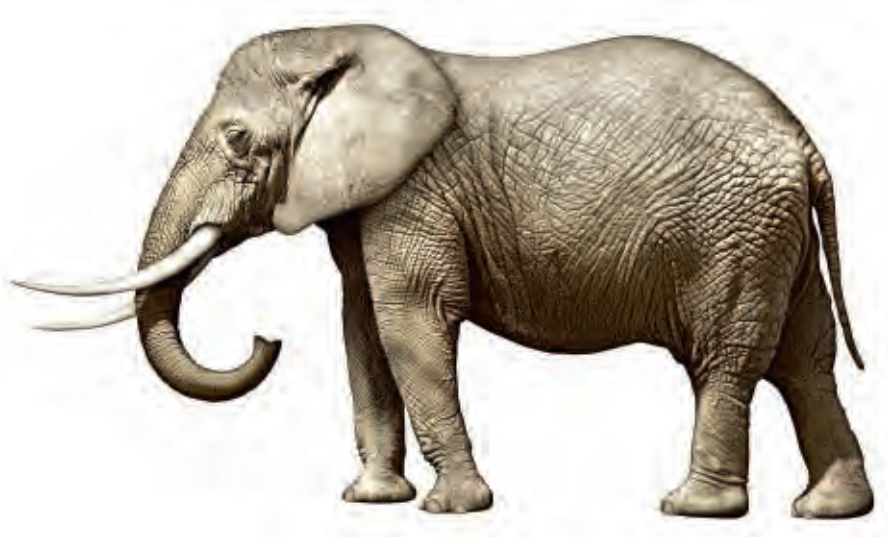
DE NOUVEAUX OUTILS STATISTIQUES

Les succès dans la reconstruction de « signaux parcimonieux » ouvrent des perspectives qui vont bien au-delà de l'acquisition comprimée, et proposent une nouvelle évolution de la démarche scientifique. Dans sa forme générique établie depuis, disons, Galilée, celle-ci procède par hypothèses, par construction de modèle, et par vérification expérimentale du modèle. Dans cette phase de vérification, on s'appliquera à déterminer les divers paramètres du modèle, ce qui n'est possible habituellement que lorsque le nombre de paramètres est assez petit, en tout cas bien plus petit que le nombre de mesures (on se souvient de la boutade de John von Neumann : pour expliquer un phénomène, avec un *fit* à quatre paramètres je peux *fitter* un éléphant, et avec cinq paramètres je peux lui faire bouger la trompe). La démarche que nous avons décrite est un peu différente : même en l'absence de modèle, on peut introduire des centaines ou des milliers de paramètres, tout en exigeant du *fit* qu'il soit parcimonieux : il devra donc mettre un certain nombre de paramètres à zéro, et ne garder que les paramètres les plus pertinents statistiquement. C'est ainsi l'analyse statistique elle-même qui peut apporter l'information sur les paramètres à considérer en priorité, amenant à réfléchir ensuite sur la construction d'un modèle. Contrairement à ce qui est parfois dit, l'irruption des données massives dans l'étude des systèmes complexes ne va pas se substituer à la théorie. Il est toujours aussi nécessaire, et de plus en plus difficile, de comprendre, analyser, construire un modèle, mais le théoricien peut s'appuyer sur de nouveaux outils statistiques extrêmement performants. ■

« Le théoricien peut s'appuyer sur de nouveaux outils statistiques extrêmement performants »

CC

Galilée est un des pères de la démarche scientifique.



© JIMLARKIN / FOTOLIA.COM

Avec 5 paramètres, on peut faire bouger la trompe de l'éléphant.