Inverse inference in the asymmetric Ising model

Jason Sakellariou

Université Paris-Sud 11
Faculté des sciences d'Orsay

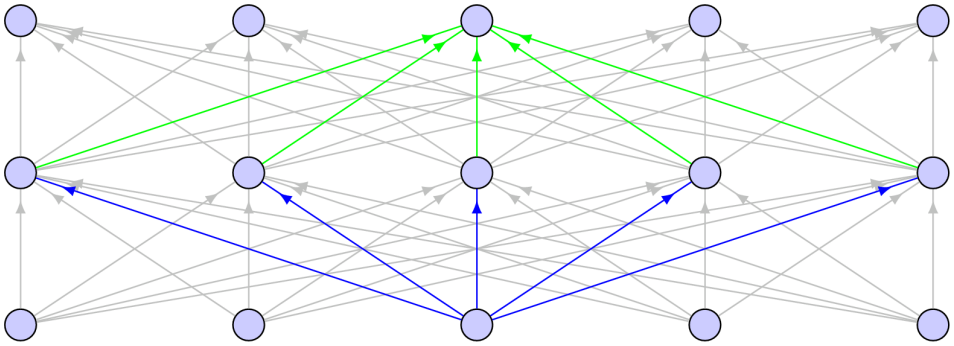
Thesis presented to obtain the degree of

Doctor of Sciences of the University Paris XI

Specialization : Theoretical Physics

by

Jason SAKELLARIOU



Inverse Inference in the Asymmetric Ising Model

Defended on February 22, 2013, in front of the Thesis Committee :

Matteo	MARSILI	referee
Federico	RICCI-TERSENGHI	referee
Silvio	FRANZ	
Martin	WEIGT	
Marc	MÉZARD	thesis advisor

Contents

Résumé Détaillé

Introduction

Récemment, des nouvelles techniques expérimentales, notamment en biologie, ont permis l'acquisition d'un très grand nombre de données concernant plusieurs systèmes complexes, comme les réseaux neuronaux, les réseaux de régulation de gènes etc. Ces techniques souvent consistent à enregistrer l'état des constituants de ces systèmes (neurones, taux d'expression des gènes etc.) à des instantanés différents. Un exemple typique est l'enregistrement, à l'aide d'électrodes, de l'activité de plusieurs neurones faisant partie d'un tissu neuronal, pendant une période de temps. Ces données nous permettent de calculer facilement certaines quantités statistiques, comme les valeurs moyennes ou les corrélations des variables qui décrivent les constituants du système en question. L'information contenue dans ces quantités ne reflète pas la structure du système d'une façon évidente et est donc de valeur scientifique limitée. Affin d'obtenir de l'information pertinente sur la structure de ces systèmes, comme la vraie connectivité d'un réseau de neurones, ces données doivent subir un traitement particulier. Ce problème est connu sous le nom *problème d'Ising inverse* affin de mettre en évidence le lien avec sa version "duale" qui est d'inférer les quantités observables quand les paramètres du modèle sont connues, appelée *problème d'Ising direct*.

Un nombre de méthodes pour résoudre ce problème d'Ising inverse existent déjà dans la littérature. Méthodes qui permettent l'exploitation de ce genre de données de façon utile. Cette thèse a comme but d'étudier ces méthodes et éventuellement de proposer des alternatives qui pourront surpasser les approches déjà connues en ce qui concerne leur précision et leur efficacité de calcul. La plupart des méthodes existantes sont basées sur une hypothèse de symétrie des interactions. Au cours de cette thèse nous nous sommes aperçus que si les interactions ne sont pas forcement symétriques, c'est-à-dire si la manière qu'un élément du réseau influence un de ses voisins n'est pas forcement identique à la manière que son voisin l'influence, la solution du problème de l'inférence peut s'écrire de façon simple ce qui conduit à une méthode qui est exacte et efficace à la fois. De plus, le interactions asymétriques sont peut-être rarement utilisées en physique statistique "traditionnelle" où les constituants des systèmes en question sont souvent des atomes ou des molécules, mais en biologie elles abondent. En effet, dans plusieurs des réseaux intéressants les interactions sont dirigées dans un sens seulement. Par exemple, dans les réseaux de neurones les neurones individuels communiquent grâce à des signaux électriques transmis le long de leur *axone* depuis le *soma* vers l'extérieur, voir fig. 1. Ceci veut dire que le paramètre qui sera utilisé pour modéliser l'influence qu'un neurone A exerce vers un neurone B aura une valeur différente pour le sens inverse.

D'autres exemples de réseaux biologiques à interactions asymétriques, comme les réseaux

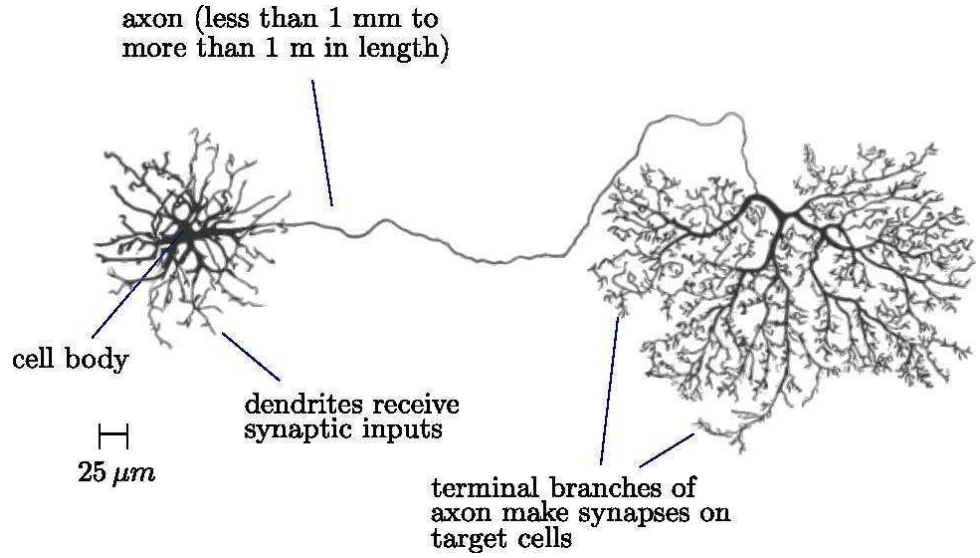


FIGURE 1 – Diagramme d'un neurone [AlbertsJL+ 02].

de régulation de gènes, sont exposés dans le manuscrit principal. Un autre aspect très important qu'on retrouve dans la plupart de ces réseaux biologiques est qu'ils correspondent à des graphes creux, c'est-à-dire chaque élément n'interagit qu'avec un petit sous-ensemble du reste des éléments. Cette propriété peut être exploitée affin de créer des méthodes moins exigeantes en nombre d'observations. Certaines des méthodes présentées dans cette thèse, ainsi que une variante de nôtre algorithme principale (voir section *Modèles creux*), tiennent compte de ce fait.

Avant de procéder avec la présentation des méthodes de résolution du problème d'Ising inverse nous allons introduire le modèle d'Ising. À la suite de ce résumé nous allons présenter brièvement l'état de l'art des méthodes utiliser pour résoudre le problème d'Ising inverse et puis, à la dernière partie, nous allons exposer une nouvelle méthode spécialement conçue pour traiter les systèmes à interactions asymétriques. Cette méthode est la contribution originale de cette thèse.

Le modèle d'Ising

Le modèle d'Ising a été proposé au départ comme un modèle de ferromagnétisme. Il consiste du Hamiltonien suivant

$$\mathcal{H} = -J \sum_{<i,j>} s_i s_j - H \sum_i s_i \quad , \tag{1}$$

où J est l'énergie d'interaction entre deux spins et H est un champ magnétique externe. Les spins sont des variables binaires prenant deux valeurs $s_i = \pm 1$. Ces variables vont nous permettre de modéliser des systèmes biologiques, comme un réseau de neurones, en faisant les correspondances $s_i = +1 \rightarrow$ neurone actif, et $s_i = -1 \rightarrow$ neurone inactif.

La probabilité d'une configuration est donnée par la distribution de Boltzmann

$$P(s_1, \ldots, s_N) = \frac{1}{Z} \, \mathrm{e}^{-\beta \mathcal{H}(s_1, \ldots, s_N)} \quad , \tag{2}$$

où $\beta = \frac{1}{k_B T}$ est la température inverse multipliée avec la constante de Boltzmann k_B. La fonction de partition Z est donnée par

$$Z = \sum_{\underline{\sigma}} \mathrm{e}^{-\beta \mathcal{H}(\sigma_1,\ldots,\sigma_N)} \quad . \tag{3}$$

Pour toute quantité qui dépend par les spins $A(\underline{s})$ on définit sa *moyenne thermique*

$$\langle A(\underline{s}) \rangle = \frac{1}{Z} \sum_{\underline{s}} A(\underline{s})\ \mathrm{e}^{\beta \mathcal{H}(s_1,\ldots,s_N)} \quad . \tag{4}$$

Un cas intéressant et facilement soluble est le cas des interactions à porté infinie, autrement le modèle *Curie-Weiss*. Dans ce cas la somme dans le Hamiltonien $\sum_{<i,j>}$ est prise sur toutes les paires de spins. Il peut être facilement démontré que l'aimantation dans ce cas est donnée par

$$m = \tanh\left(\beta J m + \beta H\right) \quad , \tag{5}$$

forme qui signale l'existence d'une transition de phase entre une phase *paramagnétique* est une phase *ferromagnétique*.

La version du modèle d'Ising la mieux adaptée aux systèmes complexes est celle qui correspond aux verres de spins. Dans ce cas chaque paire de spin participe dans une interaction portant une *constante de couplage* J_{ij} différente et de plus chaque spin peut être exposé à un *champ local* H_i diffèrent. Voici le Hamiltonien correspondant

$$\mathcal{H}(\underline{s}) = -\sum_{\langle i,j \rangle} J_{ij} s_i s_j - \sum_i H_i s_i \quad . \tag{6}$$

Les valeurs différentes des couplages servent à modéliser la variation qu'on retrouve dans le comportement de chaque synapse différente au sein d'un tissu neuronal.

Le problème inverse dans le modèle d'Ising symétrique

Formulation du problème

Affin de formuler le problème d'Ising inverse de façon précise nous devons d'abord introduire une notion clé de la théorie de l'information, la divergence de Kullback-Leibler (KL)

$$D_{\mathrm{KL}}(P \| Q) \equiv \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \quad , \tag{7}$$

qui est perçue comme une sorte de distance entre deux distributions P et Q. Plus précisément, la divergence KL quantifie la perte d'information quand on utilise la distribution Q pour modéliser un système qui obéit à la distribution P.

Désormais nous pouvons formuler le problème de la façon suivante : étant donné un ensemble de p échantillons de spins $\mathcal{S} = \{\underline{s}^{(1)}, \ldots, \underline{s}^{(p)}\}$ générés par le système en question ou étant donnés les aimantations et les corrélations des spins

$$m_i = \frac{1}{p} \sum_{\mu=1}^{p} s_i^{(\mu)} \text{ and } C_{ij} = \frac{1}{p} \sum_{\mu=1}^{p} s_i^{(\mu)} s_j^{(\mu)} - m_i m_j \ , \tag{8}$$

trouver les paramètres d'un modèle d'Ising J et H tells que la divergence KL entre la vraie distribution du système et du celle modèle soit minimale.

Par la suite nous allons présenter brièvement une série de méthodes qui ont été conçues pour résoudre ce problème.

La machine de Boltzmann

La première tentative historiquement a aboutit à un algorithme connu sous le nom *machine de Boltzmann*. La dérivation consiste à une simple réalisation de la minimisation décrite dans le paragraphe précédant. Ceci résulte à des règles d'apprentissage pour les couplages et les champs locaux

$$\delta J_{ij} = \epsilon \left(\langle s_i s_j \rangle_{\mathcal{S}} - \langle s_i s_j \rangle_{\mathcal{M}} \right) \quad \text{et} \tag{9}$$

$$\delta H_i = \epsilon \left(\langle s_i \rangle_{\mathcal{S}} - \langle s_i \rangle_{\mathcal{M}} \right) \quad , \tag{10}$$

où les indices $\mathcal{S}$ et $\mathcal{M}$ signifient que les moyennes thermiques sont effectuées par rapport aux données de spins et par rapport au modèle qu'on est en train d'inférer respectivement.

Si on itère la mise-à-jour de ces équations pour un nombre suffisant de pas on finira par obtenir les valeurs correctes des couplages et des champs locaux. Cependant, le calcul des moyennes thermiques par rapport au modèle qu'on retrouve à chaque pas de l'itération peut s'avérer très exigeant puisque, dans le cas général, ce problème appartient à la classe NP. Pour ça une série d'approximations ont vu le jour affin de pouvoir résoudre le problème d'Ising inverse dans un temps de calcul raisonnable.

Inférence exacte dans les modèles arborescents

Une façon de diminuer la complexité du calcul est de restreindre le recherche dans un sous-ensemble particulier de tous les modèles possibles. La classe de modèles la plus simple pour ce genre de calculs est la classe des modèles arborescents. Les modèles probabilistes à plusieurs variables peuvent être associés à un graphe dont les nœuds représentent les variables et les liens représentent les interactions, c'est-à-dire les facteurs de la distribution du modèle de la forme $f(s_i, s_j)$ regroupant une paire de variables. Les modèles arborescents sont donc des modèles dont le graphe sous-jacent est un arbre, c'est-à-dire ne contient pas de boucles.

La distribution de ces modèles peut se factorise selon la structure de l'arbre

$$P_t(\underline{s}) = \prod_{(ij)\in E_t} P_{ij}(s_i, s_j) \prod_{i \in V} P_i(s_i)^{1-|\partial i|} \quad , \tag{11}$$

où P_{ij} et P_i sont les marginaux à une et deux variables et où E_t et V sont les ensembles des liens et des nœuds respectivement. Cette propriété de factorisation signifie que les quantités extensives comme l'entropie peuvent s'écrire comme une somme de termes locaux.

En prenant la divergence KL entre la distribution initiale et la distribution d'un modèle arborescent on obtiens

$$D(P \| P_t) = -H(\underline{S}) + \sum_{i \in V} H(S_i) - \sum_{(ij)\in E_t} I_{ij}(S_i, S_j) \tag{12}$$

où le premier terme est l'entropie du système et le deuxième est la somme des entropies des variables individuelles, donc des quantités indépendantes du choix de l'arbre. Le troisième terme par contre, qui est la somme des informations mutuelles entre les paires qui interagissent, dépend de E_t. Ceci veut dire que affin de trouver la structure de la distribution qui minimise la divergence KL il suffit de trouver l'arbre qui maximise $\sum_{(ij)\in E_t} I_{ij}(S_i, S_j)$. L'algorithme bien connu du *maximum spanning tree* peut facilement résoudre ce problème en un temps polynomial. Une fois que la structure du graphe a été trouvée les couplages et les champs locaux peuvent aussi être calculés grâce à la méthodes des *paires indépendantes* (voir texte principal).

Si la distribution originale est arborescente, cette méthode est capable de retrouver le modèle correcte à un coût de calcul très bas ($\mathcal{O}(N^2)$). Par contre, si ce n'est pas le cas, comme dans la plupart des systèmes réels, les résultats ne sont qu'une approximation.

Méthodes champ moyen simples

Les deux méthodes précédentes montrent le jeu entre la précision de la méthode et sa complexité de calcul. Une classe de méthodes qui atteint un bon compromis entre ces deux aspects là est la classe des méthodes champ moyen. Guidés par l'intuition physique des systèmes à un grand nombre de particules, ces méthodes là offrent un cadre pour transformer un problème à N corps à un problème plus simple à 1 corps. Les détails de leur dérivation sont donnés dans le texte principal. Ici nous donnons juste leurs équations

$$m_i = \tanh\left(H_i + \sum_j J_{ij} m_j\right) \quad \text{et} \tag{13}$$

$$m_i = \tanh\left(H_i + \sum_{j\neq i} J_{ij} m_j - m_i \sum_{j\neq i} J_{ij}^2 (1 - m_j^2)\right) \quad . \tag{14}$$

Le premier système d'équations provient de la version la plus simple de la théorie du champ moyen et est donc appelé *théorie champ moyen naïve* (naïve mean-field theory). Les équations suivants portent le nom *TAP* (Thouless, Anderson et Palmer) et sont mieux adaptées pour le cas des verres de spins où elles prévoient un comportement correct dans la phase paramagnétique.

Des équations pour les corrélations peuvent être produits à partir de ces deux systèmes d'équations grâce au théorème *fluctuation-dissipation*, équations qui peuvent être inversées par la suite pour donner des méthodes permettant de calculer les couplages étant donné les corrélations et les aimantations. Les équations qui résultent de cette procédure ont été largement utilisées dans le cadre du problème d'Ising inverse.

Méthodes champ moyen avancées

En restant dans le cadre des théories champ moyen il y a moyen d'améliorer encore plus les résultats des algorithmes. Deux autres méthodes plus avancées ont été proposées les dernières années.

La première, un développement à petites valeurs des corrélations proposée par Vitor Sessak et Rémi Monasson, est capable de donner de résultats qui sont meilleurs dans la phase à haute température (phase paramagnétique).

La deuxième est une méthode basée sur la théorie des méthodes de passage de message, comme *Belief Propagation*. Elle porte le nom *Susceptibility Propagation* et a été proposée par Marc Mézard et Thierry Mora. La méthode consiste à mettre à jour itérativement un système d'équations qui sont interprétées comme des messages envoyés entre les paires de nœuds du graphe sous-jacent. La procédure atteint éventuellement un point fixe et ces valeurs finales des messages nous permettent de calculer les couplages. Ces deux méthodes là sont assez compliquées pour être décrits dans ce résumé, le lecteur peut donc se reporter dans le texte principal.

Inférence dans le régime $p < N$

Une importance particulière a été accordé dans cette thèse dans le cas où l'inférence doit être faite à un nombre d'échantillons bas $p < N$. En effet, la plupart des méthodes présentées jusqu'à présent comportent l'inversion de la matrice des corrélations, ce qui est faisable que pour $p > N$. De plus, ces méthodes là produisent des erreurs de l'ordre de $1/\sqrt{p}$ à l'inférence des couplages ce qui veut dire qu'ils ont besoin d'un grand nombre d'échantillons pour produire des résultats acceptables.

Dans ce paragraphe nous allons présenter un méthode due à P. Ravikumar, M.J. Wainwright et J.D. Lafferty qui permet d'inférer des modèles creux (sparse) à l'aide d'un petit nombre d'échantillons (de l'ordre de $\log N$).

Le premier pas de la dérivation est de considérer les voisinages entrant dans chaque spin comme indépendants. Par la suite, en suivant l'approche Bayésienne pour l'inférence nous écrivons la log-vraisemblance négative pour le voisinage de chaque spin

$$\mathcal{L}^{(i)}(J_{\setminus i}, H_i) = \frac{1}{p}\sum_{\mu=1}^{p} f(J_{\setminus i}, \underline{s}_{\setminus i}^{(\mu)}) - H_i m_i - \sum_{j \in V\setminus i} J_{ij}\widetilde{C}_{ij} \quad , \tag{15}$$

avec

$$f(J_{\setminus i}, \underline{s}_{\setminus i}^{(\mu)}) \equiv \log 2\cosh\left(H_i + \sum_{j\in V\setminus i} J_{ij} x_j^{(\mu)}\right) \quad , \tag{16}$$

où $\widetilde{C}_{ij}$ sont les corrélations non-connexes $C_{ij} + m_i m_j$. La minimisation de ces fonctions peut nous donner les couplages du modèle sous certaines conditions. La clef pour traiter le cas des modèles creux à l'aide d'un petit nombre d'échantillons est d'inclure encore un terme qui est la norme ℓ_1 de la matrice des couplages multipliée par un paramètre de contrôle $\lambda\|J_{\setminus i}\|_1$. Ceci a deux effets importants : ça nous permet d'inférer des résultats creux et ça nous permet de faire de l'inférence dans le régime $p < N$ pour des raisons qui sont exposées dans le texte principal.

Quelques simulations

Une série de simulations de Monte Carlo ont été effectués affin de comparer toutes ces méthodes. Dans le texte principal le lecteur peut retrouver plusieurs graphes pour les courbes des erreurs dans divers situations. Ici nous donnons juste deux figures pour le cas le plus pertinent pour la biologie : le cas des modèles correspondant à des graphes creux.

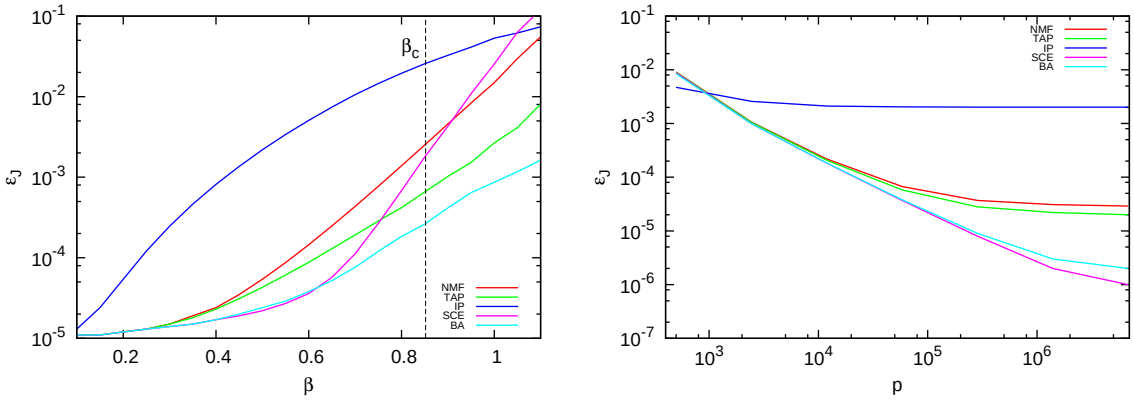


FIGURE 2 – Gauche : Erreur de l'inférence des couplages contre le température inverse β pour un système creux de $N = 100$ nœuds et de connectivité moyenne de $d = 10$ fait en utilisant $p = 10^5$ échantillons. Droite : La même chose mais cette fois contre le nombre d'échantillons p. Le système a été simulé à une température inverse de $\beta = 0.5$

L'erreur de l'inférence des couplages dans les deux figures a été calculé par la formule

$$\epsilon_J = \overline{(\beta J_{ij}^{\text{true}} - \beta J_{ij}^{\text{inferred}})^2} = \frac{2}{N(N-1)} \sum_{i<j} (\beta J_{ij}^{\text{true}} - \beta J_{ij}^{\text{inferred}})^2 \quad . \tag{17}$$

Théorie champ moyen exacte pour le modèle d'Ising asymétrique

Comme on a dit dans l'introduction, plusieurs systèmes qu'on retrouve en biologie n'ont pas des interactions qui sont forcement symétriques. Or, toutes les méthodes qu'on a présenter jusqu'à présent ont été conçues pour des systèmes à interactions symétriques. Toutes ces méthodes sont ou bien des approximations ou des algorithmes exacts mais très exigeants en calcul (machine de Boltzmann) et ces difficultés sont directement reliées à la symétrie des couplages.

Dans cette section, qui contient la grosse partie du travail original de cette thèse, nous proposons une nouvelle méthode qui, en tenant compte de l'asymétrie des interactions des éléments des systèmes en question, parvient à produire des solutions exactes à un temps de calcul très raisonnable. Nous commençons par introduire les équations concernant le problème direct et par la suite nous inversons ces équations pour obtenir une méthode de résolution du problème inverse.

Le problème direct

Tout d'abord, le modèle utilisé dans cette section est diffèrent des précédents puisque c'est forcement un modèle hors-équilibre à cause de l'asymétrie des couplages

$$P(s(t)|s(t-1)) = \prod_{i=1}^{N} \frac{1}{2\cosh(\beta h_i(t))} e^{\beta s_i(t) h_i(t)} \quad , \tag{18}$$

où

$$h_i(t) = H_i + \sum_j J_{ij} s_j(t-1) \quad . \tag{19}$$

Dans ce modèle chaque configuration de spins est conditionnée à la configuration précédente. Une notion donc de l'écoulement du temps est pertinente, c'est pour ça que nous introduiront des aimantations dépendantes du temps $m_i(t)$ ainsi que deux types de corrélations : les *corrélations à temps égaux* $C_{ij}(t)$ et les *corrélations à temps décalés* $D_{ij}(t)$, eux mêmes des fonctions du temps.

L'idée centrale du travail de cette thèse repose sur la remarque suivante : les termes de la somme dans l'équation 19 sont indépendants à cause de l'asymétrie des couplages. En effet, toute corrélation entre ces termes qui pourrait provenir de la contribution des spins à deux pas de temps en arrière est détruite à cause du fait que $J_{ij} \neq J_{ji}$. Autant que somme d'un grand nombre de termes indépendants, le terme $\sum_j J_{ij} s_j(t-1)$ a une distribution Gaussienne avec une moyenne et une variance donnés par

$$g_i \equiv \left\langle \sum_j J_{ij} s_j(t) \right\rangle = \sum_j J_{ij} m_j \quad \text{et} \tag{20}$$

$$\Delta_i \equiv \left\langle \left(\sum_j J_{ij} s_j(t) \right)^2 \right\rangle - \left\langle \sum_j J_{ij} s_j(t) \right\rangle^2 = \sum_j J_{ij}^2 (1 - m_j^2) \quad , \tag{21}$$

ce qui nous permet de remplacer les moyennes thermiques par des intégrales Gaussiennes. Le résultat pour les aimantations devient donc

$$m_i = \int Dx \tanh \left[\beta \left(H_i + g_i + x\sqrt{\Delta_i} \right) \right] \quad . \tag{22}$$

Un calcul similaire nous donne pour les corrélations la relation suivante sous forme matricielle

$$D = A\, J\, C \ , \tag{23}$$

où A est la matrice diagonale : $A_{ij} = a_i \delta_{ij}$, avec :

$$a_i = \beta \int Dx \ \left[1 - \tanh^2 \beta \left(H_i + g_i + x\sqrt{\Delta_i} \right) \right] \ . \tag{24}$$

Dans le texte principal la validité de tout ces équations a bien été vérifiée grâce à des simulations de Monte Carlo.

Le problème inverse

Une fois les relations pour le problème directe établies nous pouvons passer au problème inverse. La première remarque est que on ne peut pas simplement inverser l'équation 23 puisque

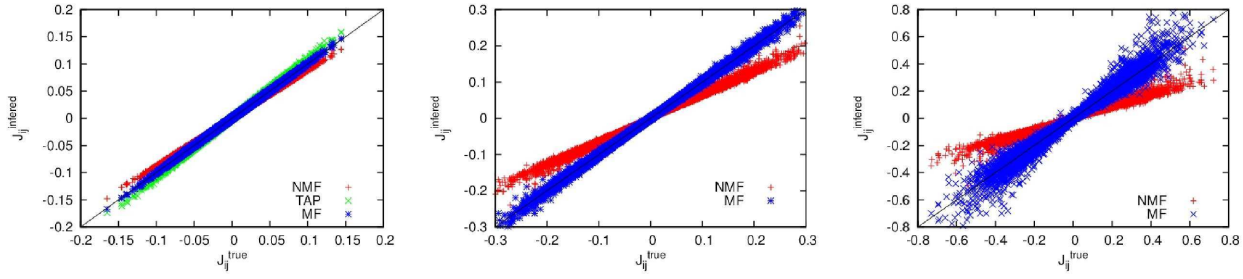


FIGURE 3 – Nuages de points. Couplages inférés contre les couplages réels de modèles de taille $N = 100$ en utilisant $p = 10^6$ échantillons. Nôtre algorithme est en bleu. Les versions hors-équilibre du champ moyen naïve (rouge) et du TAP (vert) ont aussi été incluses pour comparer. Les échantillons ont été produits à $\beta = 0.4$, 1 et 2 de gauche à droite.

la matrice A dépend de J. On proposera donc une procédure itérative qui convergera à la valeur correcte de J.

Un autre remarque importante est que puisque la matrice des couplages est asymétrique, nous pouvons inférer chaque voisinage de spin entrant indépendemment. Ceci veut dire qu'on peut laisser de côté les indices i. On réécris donc l'équation pour les aimantations comme

$$m = \int Dx \ \tanh\left[H + g + x\sqrt{\Delta}\right] \ . \tag{25}$$

Pour le calcul des vecteurs des couplages entrants J_j on introduit d'abord les vecteurs $b_j^{(i)} = \sum_k D_{ik} C_{kj}^{-1}$. Les couplages s'écrivent désormais comme

$$J_j = b_j / a \quad , \tag{26}$$

avec

$$a = \int Dx \left(1 - \tanh^2\left[H + g + x\sqrt{\Delta}\right]\right) \ . \tag{27}$$

Pour finir nous avons besoin d'une relation entre a et la variance des champs Gaussiens Δ

$$\Delta = \frac{1}{a^2} \sum_j b_j^2 (1 - m_j^2) \equiv \frac{\gamma}{a^2} \tag{28}$$

Nous sommes prêts à introduire nôtre méthode itérative pour inférer les couplages d'un système asymétrique. Il s'agit d'une itération pour trouver la valeur correcte de Δ qui nous permettra par la suite de calculer les couplages :

- Initialiser Δ
- En utilisant les valeurs empiriques de m trouver $H + g$ en inversant (25)
- En utilisant $H + g$ et Δ calculer a par l'équation (27)
- Calculer la nouvelle valeur de Δ par l'équation (28)

Il est garanti que, étant donné suffisamment d'échantillons, cette procédure convergera à la valeur correcte de Δ. Une analyse sur le nombre d'échantillons nécessaire peut être trouvé dans

le texte principal. Une fois Δ calculé nous pouvons utiliser les équations données plus haut pour trouver les couplages et les champs locaux facilement. Cet algorithme est asymptotiquement exact pour $p \to \infty$ et a une complexité de calcul de $\mathcal{O}(N^3)$.

Des résultat de la performance de cet algorithme sont exposés dans figure 3.

Modèles creux

En ce qui concerne les modèles définis sur des graphes creux nous avons adapter les idées présentées dans le paragraphe *Inférence dans le regime* $p < N$ dans le cas des systèmes à interactions asymétriques. Encore une fois nous pouvons exploiter la Gaussianité du champ effectif h_i.

On arrive à écrire une paire de règles d'apprentissage comme on l'a fait pour la machine de Boltzmann, adaptés cette fois à nôtre modèle hors-équilibre et qui utilise aussi les intégrales Gaussiennes à la place des moyennes thermiques

$$\delta H_i = \varepsilon \left(m_i - \int Dx \tanh \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) \tag{29}$$

et

$$\delta J_{ij} = \varepsilon \left(D_{ij} - [JC]_{ij} \int Dx \left(1 - \tanh^2 \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) \right) \quad . \tag{30}$$

Comme on a fait pour la méthode du paragraphe sur l'inférence à petit p on peut ici aussi introduire le terme avec la norme ℓ_1 pour obtenir la règle suivante pour les couplages

$$\delta J_{ij} = \varepsilon \left(D_{ij} - [JC]_{ij} \int Dx \left(1 - \tanh^2 \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) - \lambda \, \mathrm{sign} J_{ij} \right) \quad . \tag{31}$$

Ceci offre les mêmes avantages que ceux discutés dans ce paragraphe. Un série de simulations nous ont confirmé que ce deuxième algorithme est capable d'inférer des modèles d'Ising asymétriques avec une très bonne précision même dans des régimes où $p < N$.

Pour finir notons que nôtre algorithme a été testé sur des données réelles provenant de la rétine d'une salamandre. Les résultats sont clairement meilleurs que ceux obtenus par les autres méthodes champ moyen, mais ne sont tout de même pas en accord avec les résultats de l'algorithme exact (machine de Boltzmann). Plus de travail est nécessaire pour comprendre où est due ce désaccord.

Part I

Introduction

Chapter 1

Motivation

Statistical physics is primarily concerned with establishing a link between the microscopic and macroscopic scales of our world. In the microscopic level elementary particles are forming simple structures such as atoms and molecules through their interactions and the collective behavior of a huge number of them is creating the infinite variety of patterns that we observe in macroscopic scales. The machinery of statistical physics made possible the theoretical understanding and prediction of observable quantities that emerge in the macroscopic level starting from a description of the behavior of the elementary constituents of large systems. No realistic system being exactly solvable, statistical physicists usually made symmetry and homogeneity assumptions that enabled them to obtain solutions. In the last decades however the new field of *disordered systems* arose, where the interest shifted to systems that completely lack the simple kind of homogeneity that is present in the models of past works. In this new paradigm the elementary components, or *degrees of freedom* as we generically call them, are no longer identical but each one of them might "see" a completely different environment from its neighbors. Although initially the interest emerged from a particular type of materials, amorphous alloys of magnetic and non-magnetic metals called *spin glasses*, the theoretical ideas that where used to describe them proved to be much more promising than the materials themselves. Eventually, the mathematical difficulties that arose from this new approach were solved by the celebrated replica and cavity methods and a rich new set of behaviors emerged from these solutions, notably the striking *hierarchic* structure of their state space.

In a nutshell spin glasses are magnetic systems whose magnetic moments (spins) interact with each other in ways that tend to either align or oppose their direction in a random way. The wandering of the system as it tries to satisfy as much interactions as possible lies at the heart of its complex behavior. As it was immediately realized, a whole new class of systems that used to be considered too "complex" to be treated by physics could now be approached by the theory of spin glasses. Indeed, the fundamental assumption made in spin glasses is the randomness of the interactions between the degrees of freedom and systems of many interacting components with random interactions can be found everywhere in nature and in our technological world. In biology, examples can be found in networks of neural cells, gene regulatory networks and networks of amino acid interactions within protein-protein interactions. In one of the cornerstones of information theory, error-correcting codes, information for retrieving a corrupted message can be stored in the structure of a network involving the bits of information.

In theoretical computer science a problem of fundamental importance, the so called *satisfiability* problem which was the first to be demonstrated to belong to the NP complexity class, can be mapped to a spin glass and analyzed from a statistical physics point of view and so can a series of other constraint satisfaction problems. What is common to all the above systems and what makes them similar to a spin glass is that their components (spins, neurons, genes, amino-acids, bits, logical propositions) take part in a non-trivial network of competing constrains.

The kinds of questions that one might ask may differ however from case to case especially between biological systems on one hand (neural, gene or protein networks) and artificial ones (constrain satisfaction problems). In the latter case one usually knows a priori the network of interactions and is mainly interested in questions concerning the collective behavior of its components. Typical questions in such cases might be to find the configuration of the system which violates the least number of constrains or, in the case of probabilistic systems where different configurations can appear with different probabilities, to determine the statistics of the configurations such as the average values or correlations of the components of the system. On the other hand, in many situations appearing in biology, one might be able to measure such observable quantities, while the details of the network itself might be impossible to determine experimentally. The information contained in the measurable quantities can be however of limited scientific value without the proper processing since it doesn't necessarily reveal the actual structure of the system in an obvious way. Two variables might for example appear to be strongly correlated without being in direct interaction with each other. In such cases, a method that could predict the details of the network in question starting from the measured observables could provide valuable information about the structure of such biological systems. For various, mostly historical[i] reasons the epithet *inverse* is usualy applied to problems belonging to the second class, as opposed to *direct* for the ones belonging to the first.

In recent years new experimental methods made possible the acquisition of an overwhelming amount of data of precisely that nature for a number of different biological systems. Such experiments, whether concerning assemblies of neurons, gene or protein networks, usually record a big number of configurations of the system from which the average values and the correlations of the variables representing the different components can be deduced easily. The aim of this thesis is to investigate ways of exploiting such data usefully and to develop a method that could predict the true network of interactions starting from the measured statistics. In the past decades a series of methods for solving such inverse problems has been proposed with different degrees of success. Most of those methods rely on the assumption that the interactions between the components are symmetric in nature meaning that if element A influences element B in a particular way then B also influences A in the same way.

This thesis is organised in the following way. In the introductive, first part we find a description of some biological systems where such methods of inverse inference would be useful on one hand and a short description of the main model used in such cases, the *Ising Model*, on the other. A short section concerning the biological applicability of the Ising model was also added. In the second part we review the state of the art algorithms for solving the *Inverse Ising Problem* (IIP). All the existing methods for solving the IIP are presented in a compact way with the emphasis placed in the core idea behind each method and their computational complexity. In the end we find a chapter presenting numerical results obtained by the most important of

i. Problems in the second class where studied much more recently than problems in the first one.

these methods, in order to compare them. All methods in this part are based on the Ising model with symmetric interactions. It turns out that the symmetry of the interactions poses serious computational limitations and as a result all methods are either not efficient computationally or not exact. Finaly, in the third and last part, we study the IIP for a kinetic Ising model with asymmetric interactions. A new method for solving this particular problem is presented and compared, both theoreticaly and numericaly, with two other methods derived for the same kind of systems. As we will explain the symmetry of the interactions is not a feature necessarily present in biological contexts and models with asymmetric interactions might be much more realistic from a biological point of view. As we will see, the asymmetry of the interactions can lead to an inverse method that is is both efficient and exact, a feature not found in past works. Besides the derivation of the algorithm a full analysis of its *time complexity* and *sample complexity*[ii] is included. The main idea of the new method is applied to four variations of the inverse asymmetric Ising problem yielding four different algorithms. One of these variations concerns the case of sparse systems, where only a small number of components pairs actually interact, which are of great importance in the study of biological systems. This last part includes all the original contributions of this thesis, most of which can be also found in the published articles [MezardS 11] and [SakellariouRMH 12] reprinted in part IV.

1.1 Neural Networks

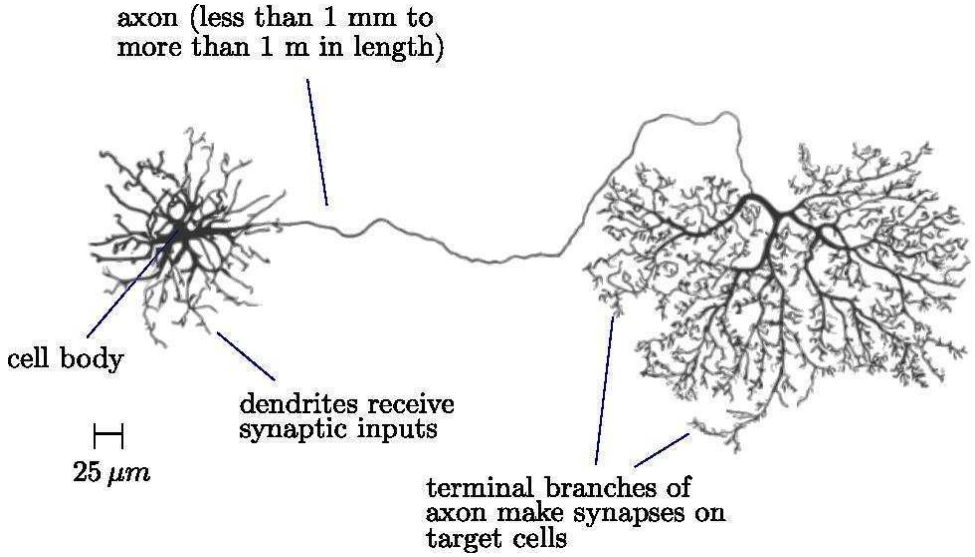


Figure 1.1: Diagram of a neuron [AlbertsJL+ 02]. The length of the axon can reach, in some cases, severals orders of magnitude higher than the size of other parts of the neuron.

One of the most complex objects in the known universe, the brain, remains a big mystery to our days. Its understanding is widely seen as one of the top scientific challenges of the 21st

ii. Basically, how the performance of the algorithm is affected by the number of samples or configurations used.

century. The human brain is composed of around 10^{11} individual cells, called *neurons*, that are connected with each other forming a total of around 10^{14} connections, called *synapses*. Its connectivity pattern is not at all random. Millions of years of evolution have shaped it, creating a very highly organized structure. It is accepted that complex functions of our organism like *memory*, *emotions*, *conscience*, *self-awareness* and *rational thought*, not to mention the hundreds of automated functions (control of breathing, heartbeat etc.) are carried by the brain and are due to the collective activity of the individual neurons which reflects the particular way they are organized.

One neuron is itself a complex object although its behavior can roughly be described in a simple way. It is composed of three distinct parts: the cell body or *soma*, the *dendrites* and the *axon*, see figure 1.1. Both the dendrites and the axon are connected to the cell body and are responsible for receiving and sending electrical signals to other neurons respectively. The axon is much longer than the dendrites as it can reach in some cases 1 m in length and is usually connected to the dendrites of some other neuron, establishing a synapse via which the neurons communicate by exchanging electrical signals.

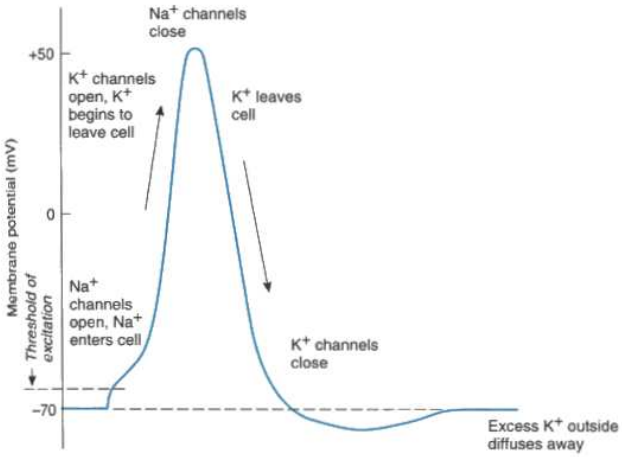


Figure 1.2: Schematic representation of the neuron's action potential together with the basic steps of how ion flow transmits the action potential down the length of the axon.

The cell membrane of the axon and soma contain ion channels that allow the neuron to generate and propagate an electrical signal. These signals are generated and propagated by charge-carrying ions including sodium (Na^+), potassium (K^+), chloride (Cl^-), and calcium (Ca^{2+}). The ion channels regulate the electrical potential difference between the cytoplasm and the extracellular medium. When the neuron is not receiving any signal from other neurons its potential difference is about -70 mV. If the voltage reaches a certain threshold (typically about -50 mV) a feedback mechanism makes ion channels to open thus increasing the voltage up to 100 mV in a very short amount of time (of the order of 1 ms) after which it quickly falls back to its initial levels, see figure 1.2. This process is called *firing* or *spiking* and the signal output generated is called *action potential*.

When a neuron fires, the action potential is transmitted through the axon who then releases *neurotransmitters* to the synapses. The neurotransmitters are causing the membrane potential

of neighboring neurons to change and thus they might trigger a spike. Depending on the type of the neurotransmitter the potential might increase, in which case we call the synapse *excitatory*, or decrease, in which case we call it *inhibitory*. Since the effect on the membrane potential might differ from case to case we usually assign a *synaptic weight* to each synapse, positive for excitatory synapses and negative for inhibitory ones.

Two important features of neural networks that will be particularly relevant in modeling them should be mentioned:

- They are *sparse*. As mentioned in the beginning of the section for $N = 10^{11}$ neurons contained in a human brain the total number of synapses is "only" around 10^{14} as opposed to $N(N-1)/2$ which would be the number for a fully connected network. This means that each neuron is connected to only a small subset of the remaining neurons and many of the possible connections are not present. These networks are called sparse and they often present a number of advantages, from a computational point of view, as we will see in later chapters.
- Neural networks are also *directed*. The action potential is transmitted along the axon in one direction, from the body cell outwards. Thus, if neuron A is excites neuron B with some synaptic weight the converse is not necessarily true: neuron B can excite neuron A with a completely different synaptic weight or, more probably, its axon might not even be connected to neuron A. In the physics jargon we say that their interaction is not symmetric. This remark will play a very important role later in this thesis where we will propose a new method for solving the inverse problem described in the beginning of this chapter. As was already mentioned, we will show that a model lacking symmetry in the interactions can be solved in a more efficient way.

1.1.1 Multi-neuron recording experiments

The complete understanding of the physiology and behavior of an individual neuron is already a challenging task. However, the mysteries of the brain's complex behavior are locked in the collective behavior of the neural network as a whole. In order to acquire information about this collective behavior experimentalists have recently developed techniques for recording simultaneously the electrical activity of multiple individual cells [MeisterPB 94]. In these experiments a number of electrodes, up to a couple of hundreds, are placed in contact with a piece of neural tissue and record the local changes in the membrane potential for a given time frame. Then, a procedure known as *spike sorting* is applied to the data in order to distinguish the true activity of each neuron from background electrical noise that can be caused from neighboring neurons. The result of this procedure is a collection of time sequences, called *spike trains*, one for each neuron, indicating the instantaneous state of the neuron in a binary way: firing or at rest. An example of a spike train can be seen in figure 1.3.

Spike trains like the one depicted above contain a lot of information about statistical dependencies between single neurons and can be used in principle to extract information about the true synapses and their weight. This problem has been actively studied in recent years [CoccoLM 09, SchneidmanBSB 06, ShlensFGG$^+$ 06]. This cannot be done in a naive way, however, since neurons that appear to be correlated might not be directly connected but may interact instead trough some other, intermediate neuron. These data must be approached in a

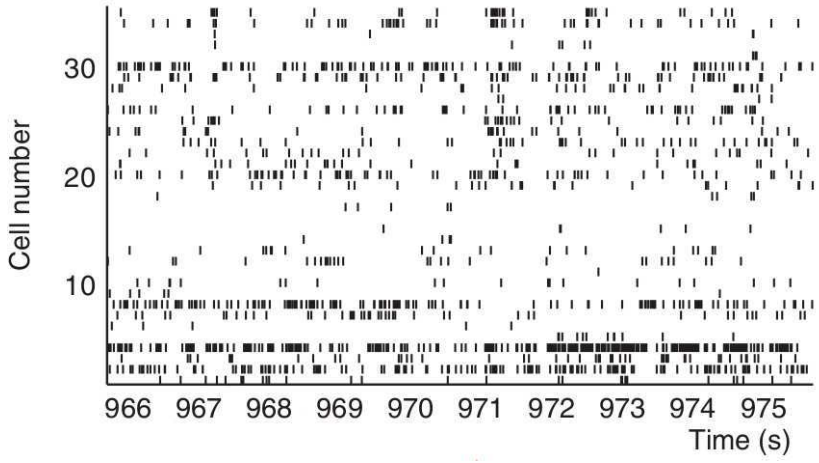


Figure 1.3: Example of spike trains [PeyracheBK$^+$ 09]. Each line corresponds to a single neuron and is divided into time bins. Each vertical bar indicates if the corresponding neuron was firing in the particular time bin.

global way, meaning that one must select the best network of synapses that can reproduce the statistics of the given data as a whole.

1.2 Other biological systems

Apart from neural networks a number of other systems found in biology present similar features and can be modeled in a similar way. We will make a short description of those systems and highlight their complex network nature.

1.2.1 Gene-regulatory networks

Transcriptional gene regulation is one of the cornerstones of developmental biology. It constitutes a feed-back mechanism in the transcription of genes to mRNA that allows the genome to be expressed in different ways, thus allowing different patterns to emerge starting from the same genetic information. This mechanism forms the basis for *cell differentiation* and *development*. A simple example of such a procedure is *transcriptional repression* and *activation* of a gene by one *transcription factor* (*i.e.* a regulatory protein). The molecular machine that transcribes genes to mRNA, called RNA polymerase, has specific binding sites on the DNA from where the transcription starts. It can happen that its binding site overlaps with the binding site of a transcription factor and thus a high concentration of the transcription factor can *repress* the transcription rate of the gene in question. On the other hand, other transcription factors can act as *activators*. In these cases the binding sites of the transcription factor and of the polymerase have close positions on the DNA and an attractive interaction between them can exist. The result is a cooperation between the transcription factor and the polymerase which means that high levels of the first can enhance the gene transcription.

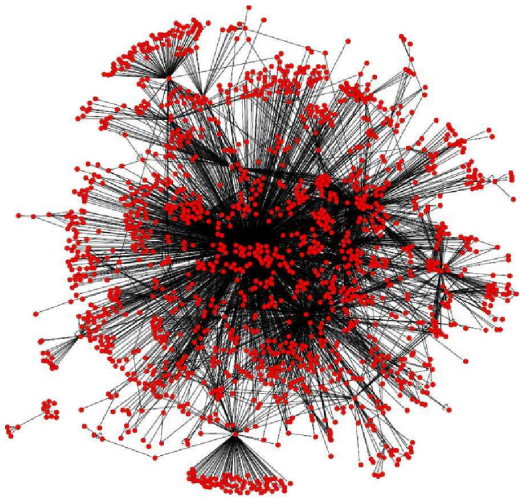


Figure 1.4: The transcriptional regulatory network of the *E. coli* [Freyre-GonzalezT 10]. Red nodes represent genes and links represent regulatory interactions. The figure highlights the extreme complexity of the GRN in such a simple organism.

The expression of transcription factors is regulated by other transcription factors. The set of all genetic interactions of this kind, between transcription factors and genes is called a *gene-regulatory network* (GRN), (see figure 1.4). Their understanding is of utmost importance in developmental biology since they describe the complex mechanisms leading from the genome to the formed organism. From a theoretical point of view they present similarities with the neural networks of the previous sections. Instead of a set of neurons that can be active or at rest we have a set of genes that can be expressed or not. The interactions between the elements of the networks are similar in nature also, since each gene can repress or activate the expression of others just like a neuron can inhibit or excite other neurons. Gene-regulatory networks also exhibit the two important features of sparsity and directionality that we discussed in section 1.1 since each gene is regulated or regulates only a small number of other genes compared to the total number of genes present in an organism and the regulation mechanism is such that interactions don't necessarily need to be symmetric.

Modern micro-array techniques enable the simultaneous measuring of the expression levels of order 10^4 RNAs. Statistical methods have been proposed recently for inferring the structure of gene-regulatory networks starting from those measurements [BraunsteinPWZ 08]. As in the case of neural networks, those methods rely on the Ising model which will be introduced in the next chapter.

1.2.2 Protein-protein interaction

Many of the most important molecular processes in a cell such as *DNA replication* or *signal transduction* (*i.e.* the propagation of chemical signals from the exterior to the interior of the cell) are carried out by large molecular complexes that are build from many interacting proteins. The understanding of the precise way proteins interact with each other is one of the outstanding

challenges in biology.

Proteins are large molecules formed of one or more chains of amino-acids folded in space and thus have a three-dimensional structure. When two proteins interact some parts of the first protein, thus some of its amino-acids, appear close in space with amino-acids of the second protein. Understanding the interaction involves identifying which amino-acids interact with which others and thus inferring the stereometry of the resulting protein complex. Again, the availability of large databases in recent years has made the use of statistical methods an attractive option as opposed to traditional methods such as crystallography. In recent works [WeigtWS$^+$ 09] it was proposed that the information on amino-acid interactions could be found in evolution.

Homologous proteins are proteins with a common evolutionary origin and thus they usually share a number of amino-acids in similar positions, which usually imply also a similar biological function. However, a number of amino-acids underwent mutations and are differentiated from one protein to another. When a mutation occurs in some part of a protein that interacts with some other protein, the interaction might be affected and the particular function might be lost. Hence either the mutation is not established in the population, either it is accompanied by a compensatory mutation in the second protein. Thus, by observing which pairs of amino-acids are correlated between two proteins one might deduce which pairs are interacting (the same line of though can be applied to pairs within a protein since those also interact and are causing the protein to fold in space).

As with the rest of the systems presented in this introduction, a statistical correlation doesn't necessarily imply a direct interaction. An element of a network (amino-acid, gene, neuron) A might interact with a number of others $C, D, \ldots$ who in turn can interact with an other one B so that A and B might appear correlated without being in direct interaction. The solution is found in treating the system in a global way by trying to find a system of interactions that reproduce the whole set of correlations. The authors of [WeigtWS$^+$ 09] have shown that by using a *Potts model* they where able to discern the direct from indirect interactions. The model used follows the same principle as the models used in neural and gene networks (it will be introduced in the next chapter) with the difference that instead of the variables taking binary values, they can take 21 different values, one for each kind of amino-acid plus one for an empty space.

Chapter 2

The Ising model

Systems with many interacting components, as the ones discussed in the previous chapter, are often modeled using the celebrated *Ising* model. This model completely disregards the details of the interacting elements and treats them as discrete, binary, variables. It is also a probabilistic model in the sense that one doesn't have to work with the coupled equations of motion of all elements, an immensely complex task, but can examine their statistical equilibrium properties instead. In this chapter we will introduce the Ising model and some of its variants and review some classical results.

2.1 The ferromagnetic Ising model

The Ising model was invented by physicist Wilhelm Lenz in 1920 who gave it as a problem to his student Ernst Ising. It was intended as a model for ferromagnetism that could account for the ferromagnetic/paramagnetic phase transition occurring in ferromagnetic materials. Ising was not able to solve the two-dimensional version of the model but gave the solution for the one-dimensional one in his thesis in 1924. He showed that in one dimension there is no phase transition in finite temperature and that long-range ferromagnetic order appears only in zero temperature. The two-dimensional case is much harder and its solution was only found some twenty years later by Lars Onsager, a solution that shows the existence of a phase transition in a non-zero temperature.

The model supposes that the magnetic moments or spins, one from each atom of a ferromagnet, are arranged in a lattice and they interact with their neighbors. The spins are described by binary variables taking the values $s_i = \pm 1, i = 1, \ldots, N$ where N is the total number of them. The energy of the system is given by the Hamiltonian

$$\mathcal{H} = -J \sum_{<i,j>} s_i s_j - H \sum_i s_i \quad , \tag{2.1}$$

where J is the energy of the interaction between two spins and H is an external field favoring one one of the two directions: up or down. The notation $\sum_{<i,j>} s_i s_j$ means that the summation is over the closest neighbors in the lattice.

The probability of a configuration is given by the Boltzmann distribution

$$P(s_1, \ldots, s_N) = \frac{1}{Z} \, \mathrm{e}^{-\beta \mathcal{H}(s_1, \ldots, s_N)} \quad , \tag{2.2}$$

where $\beta = \frac{1}{k_B T}$ is the inverse temperature rescaled by the Boltzmann constant k_B. The partition function Z is given by

$$Z = \sum_{\underline{\sigma}} \mathrm{e}^{-\beta \mathcal{H}(\sigma_1,\ldots,\sigma_N)} \tag{2.3}$$

where $\sum_{\underline{\sigma}} \equiv \sum_{\sigma_1=\pm 1} \sum_{\sigma_2=\pm 1} \ldots \sum_{\sigma_N=\pm 1}$. For the rest of this work we will absorb the Boltzmann constant in T. In fact, when treating the inverse Ising problem later in this thesis, we will completely omit the factor β since it just re-scales the magnitude of the interactions through J and H. For any quantity that depends on the spins $A(\underline{s})$ we define its *thermal average* as

$$\langle A(\underline{s}) \rangle = \frac{1}{Z} \sum_{\underline{s}} A(\underline{s}) \, \mathrm{e}^{\beta \mathcal{H}(s_1,\ldots,s_N)} \quad . \tag{2.4}$$

In one dimension there is no phase transition and the ordered phase can only occur in $T = 0$. It can be easily shown that the two-site *correlation* for a zero external field $H = 0$ is given by

$$\langle s_i s_j \rangle = (\tanh \beta J)^{|i-j|} \quad . \tag{2.5}$$

We see that is decays exponentially with the distance between the two spins signaling the absence of long-range order.

In the two-dimensional case Onsager's solution shows the existence of two phases, a ferromagnetic one where *magnetizations* $m = \langle s_i \rangle$ are non-zero and a paramagnetic one where they are zero. The phase transition occurs at the *critical temperature*

$$T_c = \frac{2J}{\ln(\sqrt{2}-1)} \quad . \tag{2.6}$$

Another interesting case is the so called *Curie-Weiss* model [MezardM 09]. It is an Ising model with infinite-range interactions (it can also be seen as describing an infinite dimensional system). In this case the Hamiltonian is slightly different

$$\mathcal{H} = -\frac{J}{N} \sum_{i<j} s_i s_j - H \sum_i s_i \quad . \tag{2.7}$$

Now the summation runs over all $N(N-1)/2$ possible pairs and the factor $1/N$ has been introduced in order to keep the energy extensive. Infinite-range interactions or infinite-dimensional systems are not physical but the advantage of this model is that it is exactly solvable. In a nutshell the infinite number of neighbors allows us to omit the fluctuations when taking a thermal average in eq.(2.4) and to replace the average of a function by the function of the average which in the case of the magnetization leads to

$$m = \tanh(\beta J m + \beta H) \quad . \tag{2.8}$$

The above equation can be solved numerically and the solution displays also two distinct phases, see figure 8.11. For a zero external field $H = 0$ we have that for $\beta \leq 1/J \equiv \beta_c$ the only solution of eq.(8.11) is $m(\beta) = 0$. For $\beta > 1/J$ however two new solutions appear at $m_\pm(\beta)$, with $m_+(\beta) = -m_-(\beta) > 0$. It turns out that they are also the only stable solutions in this

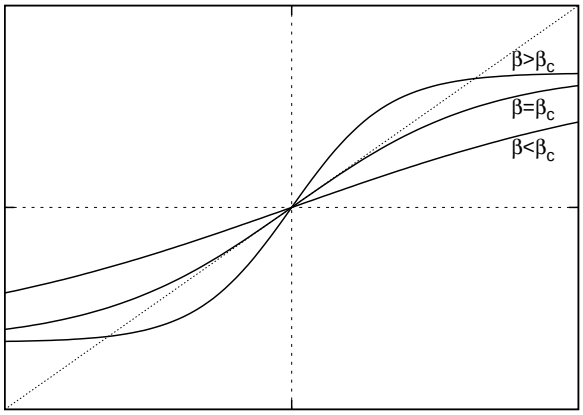


Figure 2.1: The function $\tanh(\beta Jm + \beta H)$ plotted against m, for $H = 0$ and for three values of β: above, equal and below the critical value $\beta_c = 1/J$. In the low temperature regime ($\beta > \beta_c$) the curve intersects the diagonal (dashed line) and solutions with non-zero magnetization appear.

phase. Their meaning is that below a given temperature the interactions between spins become important enough to keep them aligned in either the up or down direction (the two directions being equivalent since $H = 0$).

The Curie-Weiss model is an example of a very important family of models: the *mean-field models*. The mean-field character of this model comes from the fact that one can omit the fluctuations of the effective field felt by each spin $h = J/N \sum_j s_j + H$ and replace it with its mean value $h = Jm + H$ in the thermal average. This can be shown to yield exact results in the limiting cases mentioned earlier: infinite-dimensional systems or infinite-range interactions. However, this mean-field ansatz can be can also be used as an approximation in other cases. Many of the methods described in later chapters are based on mean-field arguments.

2.2 The Sherrington-Kirkpatrick model

When the magnetic properties of the newly created spin-glasses where investigated in the 70s theoreticians turned to the Ising model for answers. However, the simple model described in the previous section had to be modified to account for the randomness of the interactions between the spins. Spin-glasses consist basically of metallic materials hosting magnitic impurities in random positions. Around the impurity, the spins have a polarization that oscillates at large distances, rendering the sign of the interaction random since the distances between impurities is a random variable.

In order to model the situation described above a modified version of the Ising model was used where now each spin variable models an impurity and the interaction energy J is not constant any more but takes a random value for each interaction. The short-range version [EdwardsA 75] proved too difficult to be solved but results where obtained for the infinite-range case [SherringtonK 75], the so called *Sherrington-Kirkpatrick* (SK) model which we will

present here.

The Hamiltonian of the SK model is given by

$$\mathcal{H} = -\sum_{i<j} J_{ij} s_i s_j - H \sum_i s_i \quad . \tag{2.1}$$

The *coupling constants* J_{ij} (or simply the couplings), as they will be called from now on, are chosen at random from a Gaussian distribution

$$P(J_{ij}) = \frac{N}{\sqrt{2\pi J^2}}\, \mathrm{e}^{-\frac{N}{2J^2}\left(J_{ij}-\frac{J_0}{N}\right)^2} \quad , \tag{2.2}$$

where J_0 is a ferromagnetic bias of each interaction, J defines a typical scale for the couplings and the scaling with $\sqrt{N}$ is chosen so that the energy is extensive. In the infinite-range ferromagnet of the previous section the couplings had obviously to be rescaled by a factor N since each spin interacted with $\mathcal{O}(N)$ others through a fixed coupling. Now the J_{ij}'s are randomly distributed around 0 so that their random signs create cancellations between the different terms. Since the sum of a large number N of variables with random signs is of order $\mathcal{O}(\sqrt{N})$ they have to be rescaled by $\sqrt{N}$ in order for the energy to be of order $\mathcal{O}(N)$.

Unlike in the ferromagnetic case, in the spin-glass case the simple mean-field method described before doesn't yield correct results. A more elaborated mean-field approach, the *Thouless-Anderson-Palmer* (TAP) method, succeeds in providing correct results in the high temperature phase by taking into account some peculiarity of spin-glasses. We won't get into the details of the TAP approach here since we will discuss it in details in the second part of this thesis. Instead we will discuss very briefly the results yielded by the celebrated *replica method*. For a more detailed analysis see [MezardPV 87, Nishimori 01].

As usual in statistical mechanics one needs to evaluate the free-energy $F = -\ln Z$ which in the case of the SK model depends on the particular realization of the disorder, *i.e.* on the particular sampling of the J_{ij}'s. Since the free-energy is extensive we expect that it will coincide with its average value in respect with the distribution of the couplings $-\overline{\ln Z}$, where the overline is used to denote averages with respect to $P(J_{ij})$ (as opposed to the thermal average $\langle \cdot \rangle$). Now, because of the logarithm the above average is very difficult to evaluate but one can instead evaluate $\overline{Z^n}$ and then use the identity

$$\overline{\ln Z} = \lim_{n\to 0} \frac{\overline{Z^n} - 1}{n} \quad . \tag{2.3}$$

For n integer the average $\overline{Z^n}$ is much easier to compute since it is just the partition function of n replicated systems. Then however one must analytically continue to the reals in order to take the limit which is a very tricky matter from a rigorous point of view. The above approach, known as the *replica trick* is widely used for solving statistical mechanics problems in the field of disordered systems.

The partition function of the replicated system is

$$Z^n = \sum_{\underline{s}} \exp\left[\sum_{\alpha=1}^{n} \sum_{i<j} J_{ij} s_i^\alpha s_j^\alpha + H \sum_{\alpha=1}^{n} \sum_i s_i^\alpha\right] \quad . \tag{2.4}$$

Then, taking the average over the disorder

$$\begin{aligned}
\overline{Z^n} &\equiv \int \left(\prod_{i<j} dJ_{ij} P(J_{ij}) \right) Z^n \\
&= \mathrm{e}^{N^2 J^2 n/4} \sum_{\underline{s}} \exp \left[\frac{1}{N} \sum_{i<j} \left(J_0 \sum_{\alpha=1}^{n} s_i^\alpha s_j^\alpha + \frac{J^2}{2} \sum_{1 \le \alpha < \beta \le n} s_i^\alpha s_j^\alpha s_i^\beta s_j^\beta \right) + H \sum_i \sum_{\alpha=1}^{n} s_i^\alpha \right] \\
&\approx \mathrm{e}^{N^2 J^2 n/4} \sum_{\underline{s}} \exp \left[\frac{J_0}{2N} \sum_{\alpha=1}^{n} \left(\sum_i s_i^\alpha \right)^2 + \frac{J^2}{2N} \sum_{1 \le \alpha < \beta \le n} \left(\sum_i s_i^\alpha s_i^\beta \right)^2 \right. \\
&\qquad \left. + \; H \sum_i \sum_{\alpha=1}^{n} s_i^\alpha \right] \quad . \qquad (2.5)
\end{aligned}$$

In order to linearize the squared quantities $\left(\sum_i s_i^\alpha s_i^\beta\right)^2$ and $\left(\sum_i s_i^\alpha\right)^2$ we introduce Gaussian integrals, with integration variables $q_{\alpha\beta}$ and m_α respectively, such that their linear term corresponds to the above squared quantities.

$$\begin{aligned}
Z^n &= \mathrm{e}^{N^2 J^2 n/4} \int \prod_{\alpha<\beta} dq_{\alpha\beta} \int \prod_\alpha dm_\alpha \exp \left(-\frac{NJ^2}{2} \sum_{1 \le \alpha < \beta \le n} q_{\alpha\beta}^2 \right. \\
&\qquad \left. - \frac{NJ_0}{2} \sum_\alpha m_\alpha^2 + N \ln \sum_{\underline{s}} \mathrm{e}^L \right) \qquad (2.6)
\end{aligned}$$

with

$$L = J^2 \sum_{\alpha<\beta} q_{\alpha\beta} s^\alpha s^\beta + \sum_\alpha (J_0 m_\alpha + H) s^\alpha \quad . \qquad (2.7)$$

Notice how we have dropped the spin index in the last expression since only a single index i appears in the last line of eq.2.5. This is because the replica trick has the effect of decoupling the spins and coupling the replicas. This is not just a technical detail: the quantity resulting from this coupling between replicas, the *overlap* $q_{\alpha\beta}$, plays an important role in characterizing the *spin-glass phase* (together with the usual ferromagnetic order parameter m_α).

Since the exponent of the integrand in eq.(2.6) is proportional to N we can evaluate the integral by steepest descent. After some calculations it can be shown that the free-energy averaged over the disorder is

$$\overline{F} = N \lim_{n \to 0} \left\{ -\frac{J^2}{4n} \sum_{\alpha \neq \beta} q_{\alpha\beta}^2 - \frac{J_0}{2n} \sum_\alpha m_\alpha^2 + \frac{1}{4} J^2 + \frac{1}{n} \ln \sum_{\underline{s}} \mathrm{e}^L \right\} \quad , \qquad (2.8)$$

where the variables $q_{\alpha\beta}$ and m_α must be chosen such that they extremize the quantity in braces.

Since the replicas where introduced artificially in order to compute Z^n one naively expects that they are completely equivalent, *i.e.* $q_{\alpha\beta} = q$ for $\alpha \neq \beta$ and $m_\alpha = m$. It turns out that this *replica symmetric* hypothesis yields correct results as long as the temperature is high

enough. It is interesting to note that in the high temperature regime the equation of state of the magnetization takes a Gaussian form [Nishimori 01]

$$m = \int Dx \tanh\left[J_0 m + H + J\sqrt{q}x\right] \quad , \tag{2.9}$$

where $Dx \equiv \frac{dx}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is a Gaussian measure. The comparison of the above equation with eq.(8.11) suggest that the effective field felt by each spin has a Gaussian distribution due to the randomness.

The replica symmetric hypothesis can be used to derive the phase diagram of the model. Without going into details we will just describe the three phases: the simplest case $m = q = 0$ corresponds to a paramagnetic phase where the magnetization is zero because each spin spends and equal amount of time pointing upwards and downwards. If $m \neq 0$ then we are in the ferromagnetic phase where the spins have a tendency to be aligned. If $H = 0$ this can happen only when the ferromagnetic component of the interactions J_0 is important compared to the typical scale of the random component of the couplings J. Finally, as opposed to the simple ferromagnetic case of the previous section, there is a third scenario where $m = 0$ but $q \neq 0$, the *spin-glass phase*. In this phase, the free-energy develops a complex landscape and the system is "locked", due to the low temperature, in some valley of this landscape. For a given sampling of the couplings J_{ij} the individual magnetizations $m_i \equiv \langle s_i \rangle$ are non-zero but when the average over the realization of the J_{ij}'s is taken the overall magnetization vanishes. Interestingly however, q is not zero since it reduces to $\overline{\langle s_i \rangle}^2$ in the replica symmetric situation which is an average of a positive quantity since the spins are frozen in some configuration for each set of J_{ij}. The boundaries of the three phases can be seen in figure 2.2

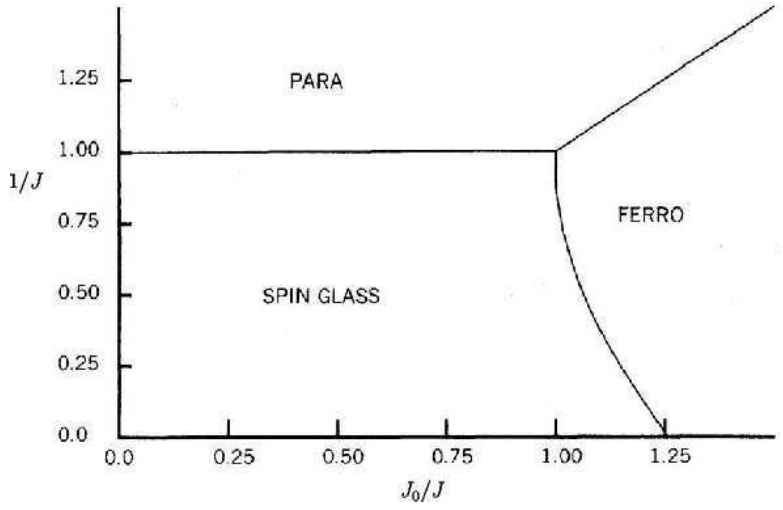


Figure 2.2: Phase diagram of the Sherrington-Kirkpatrick model [SherringtonK 75]

The replica symmetric hypothesis, however, turns out to be wrong for low temperatures since it predicts a negative entropy. The solution to this problem was given in a series of papers in the late 70s by Giorgio Parisi with the famous *replica symmetry breaking* ansatz. The key

idea is to abandon the equivalence of the replicas and to suppose that actually the quantity $q_{\alpha\beta}$ depends on the replica indices α and β. His solution involves a hierarchic construction where the $q_{\alpha\beta}$ matrix is builded by iteratively dividing diagonal and off-diagonal blocks and assigning them different values. As we said before the spin-glass phase is characterized by the existence of an infinite number of free-energy valleys separated by infinitely high barriers (in the thermodynamic limit) that forces the system to stay frozen in some particular state. The different values of the overlap matrix $q_{\alpha\beta}$ somehow reflect the varying similarities between systems locked into different valleys. A striking property of the Parisi solution is the structure of the overlaps between states: in the three possible overlaps between any triplet of states two have to be equal and one strictly larger than the other two. One can use the overlap to define a distance in state space. Because of the above property of the overlap the space acquires an *ultrametric* structure where, as opposed to the usual metric space, the triangle inequality no longer holds and is replaced by a stronger one stating that for any three points x, y and z the distances must obey $d(x, y) \leq \max\{d(x, z), d(z, y)\}$.

2.3 On the biological applicability of the Ising model

The Ising model and its variations presented in the previous section has proved to yield a very good description of magnetic systems such as ferromagnets and spin-glasses. Why should we use it to model systems of much greater complexity such as the brain or a regulation network of genes? The answer is essentially the same as to the question why it works when applied to ferromagnets. In 1957 E.T. Jaynes wrote two papers [Jaynes 57a, Jaynes 57b] on the link between statistical mechanics and Shannon's newly formulated *information theory*. In these works he offered a new way of understanding why the Gibbsian formulation of statistical mechanics works. He realized that the *thermodynamic entropy* and the *information entropy* are essentially the same thing. We will discuss the subject in greater depth in chapter 3 but for the moment let us state, without explaining why, that the quantity

$$H \equiv -\sum_x p(x) \log p(x) \quad , \tag{2.1}$$

called entropy, is related to the concept of information. Precisely it quantifies one's lack of certainty on the outcome of the random variable x. With that remark Jaynes formulated what is known as the *maximum entropy principle*: given some prior measurable information on a probabilistic system (usually the average value of some function of its microstates, such as the magnetization) of all the distributions that agrees with that data, the one that best represents our state of knowledge is the one that maximizes the entropy. In other words, if we can measure only a number of observable quantities of a system, the scientifically most "honest" choice for a probability distribution for modeling that system is the one that, while agreeing with the measurements, codifies our complete lack of further knowledge. For example, let's say x is some variable describing the microscopic state of our system and $\langle \mathcal{H}(x) \rangle$, the average of some function $\mathcal{H}$, is an observable quantity. Then it can easily be shown with the help of Lagrangian multipliers [Jaynes 57a, Jaynes 57b] that the maximum entropy distribution is

$$P(x) = \frac{1}{Z} \, \mathrm{e}^{-\mathcal{H}(x)} \quad \text{with} \quad Z = \sum_x \mathrm{e}^{-\mathcal{H}(x)} \quad . \tag{2.2}$$

The above distribution is nothing but the Boltzmann distribution found everywhere in statistical mechanics. Under this light the choice of the Boltzmann distribution becomes much more understandable, as the distribution that contains no further information beyond what was assumed we had.

So, in the case of neural networks, the quantities we can measure in experimental settings as the one described in section 1.1.1 are the average values and pairwise correlations of the states of the neurons, $\langle s_i \rangle$ and $\langle s_i s_j \rangle$. It turns out that, using the above line of arguments, the distribution we have to choose is the Ising distribution. The only difference with the magnetic systems described in the previous sections is that in ferromagnets and spin-glasses we are able to embed the system in a uniform magnetic field, hence the term H in the Hamiltonians of eqns.(2.1,2.7,2.1), whereas in neural systems we choose to introduce different *local fields* to account for different biases in the state of each neuron. Hence the Hamiltonian used in this case is

$$\mathcal{H}(\underline{s}) = -\sum_{\langle i,j \rangle} J_{ij} s_i s_j - \sum_i H_i s_i \quad . \tag{2.3}$$

This Hamiltonian sufficiently describes neural systems since the nature of the physical interactions in such cases is pairwise. In GRN's the situation is more complicated since different combinations of genes or transcription factors can affect in different ways the expression of some gene, and thus the above model is just approximative and modifications must be made, such as the inclusion of three-body interactions $K_{ijk} s_i s_j s_k$.

Part II

The Symmetric Inverse Ising Problem

Since the dawn of scientific thought, science is concerned primarily with two concepts, *modeling* and *prediction*. These are really complementary and can be thought as two opposing motions between the "world" of scientific models and the real world. *Modeling* is finding the most adequate model that explains a given set of observable quantities and *Prediction* is computing the values of the observables given some model. Traditionally statistical physicists have concentrated most of their efforts in the second procedure. Starting from simplistic, yet powerful, models the main effort was to develop a computational inventory suitable for computing and understanding the behavior of observable quantities that quantify the collective behavior of the system. This is why we will use the epithet *Direct* for this kind of problems and reserve the word *Inverse* for the opposite procedure. In more recent decades, however, with the advent of the *disordered systems paradigm* and the applicability of its ideas to biological, socio-economic and other fundamentally complex systems, the modeling procedure has become itself complex and computationally difficult. The older concepts of symmetry and homogeneity no longer apply to disordered systems as they are described by a huge amount of different parameters the particular values of which might play an important role in the collective behavior of the system. Starting from measurements of observable quantities, inverse procedures, able to infer models fitting the observables, could provide extremely valuable information about the structure of many systems such as brain connectivity, causal dependencies between gene expressions, protein interaction patterns, DNA folding and even the details of financial networks. However, because of the underlying complexity, such inverse methods are not trivial.

In this chapter we will outline and compare the most important methods used in the literature for solving the *inverse Ising problem*. Our main aim here is not to go into every detail but to outline the derivation of the various methods and, most importantly, make a comparative study of their various traits, *i.e.* their time complexity, sample complexity and the limiting case where they are exact, if any exist. These methods vary a lot, ranging from the exact but computationally infeasible Boltzmann machine, to physics inspired mean-field and high temperature expansion methods and beyond to more sophisticated methods for treating particular varieties of models *e.g.* sparse networks. All these methods have their advantages but they all experience some regime where they become either infeasible or they fail to provide correct results. In part III we will propose a new method who is instead exact and efficient in any regime given only one condition which is usually found in biological applications, the asymmetry of the interactions. A direct comparison between our method and the ones found in this chapter will therefore not be always possible since most of the methods presented here were originally conceived for symmetric systems.

The process of inferring a model by analyzing a set of observed quantities is reminiscent of the biological process of knowledge acquisition, where one observes the outcomes of real world situations and figures out what kind of mechanisms are responsible for such outcomes. Hence inverse problem methods are often found in the literature under the name *learning*. Notions like *knowledge* and *learning* can be formalized with the help of a powerful tool, *Information Theory*, born some sixty years ago in the works of Claude Shannon. In the next section we will introduce some of the central ideas of information theory that will be used as the main framework of all the methods discussed in this thesis.

Chapter 3

Some information theory background

In order to quantify the notion of *information*, scientists have introduced a set of concepts which have a close parallel with many statistical physics concepts. The central role is played by the *entropy* of a random variable $H(X)$. In information theoretic contexts it is usually called *Shannon entropy*. When the random variable obeys a Boltzmann distribution its definition coincides with the statistical mechanics definition of the *thermodynamic entropy* up to a multiplicative factor, the Boltzmann constant k_{B}, who guarantees that the units match those of other thermodynamic quantities. Since we will treat only models obeying the Boltzmann distribution we will just use the name entropy throughout the rest of this thesis. The entropy is defined with respect to the distribution of some random variable in the following way

$$H_P(X) \equiv -\sum_{x\in\mathcal{X}} P(x)\log P(x) = \mathbb{E}_P[\log\frac{1}{P(x)}] \quad , \tag{3.1}$$

where $\log\frac{1}{P(x)}$ is called the *self-information* of the variable as it quantifies how much information is represented in the outcome of the random variable. Although the above definition has the same form as the one found in the works of Ludwig Boltzmann and J. Willard Gibbs in the 1870s, it was not until Claude Shannon's celebrated paper [Shannon 48] in 1948 that the connexion between entropy and information was made clear. Shannon showed that, given a source that assigns values to some variable at random, its entropy bounds the smallest average message length that we can use in order to communicate the outcomes of the variable without losing information. This means, informally, that the entropy quantifies our lack of certainty. The basic properties of the entropy that justify its use as an measure of information are the following:

(a) $H(X) \geq 0$.

(b) $H(X) = 0$ if and only if X is certain.

(c) For a given set of events $\mathcal{X}$, $H(X)$ is maximal when all events are equiprobable and takes the value $\log|\mathcal{X}|$.

(d) For any pair of random variables $H(X,Y) \leq H(X) + H(Y)$.

(e) $H(X,Y) = H(X) + H(Y)$ if and only if the two variables are independent, *i.e.* if $P(x,y) = P(x)P(y)$.

(f) For some partition of the events set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ and $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$, the entropy has two contributions $H(X) = H(\mathcal{X}_i) + H(X|x \in \mathcal{X}_i)$, where

$$H(\mathcal{X}_i) = -\left(\sum_{x\in\mathcal{X}_1} P(x)\right)\log\left(\sum_{x\in\mathcal{X}_1} P(x)\right) - \left(\sum_{x\in\mathcal{X}_2} P(x)\right)\log\left(\sum_{x\in\mathcal{X}_2} P(x)\right)$$

is the entropy associated with the choice of subset and

$$\begin{aligned} H(X|x\in\mathcal{X}_i) &= -\left(\sum_{x\in\mathcal{X}_1} P(x)\right)\sum_{x\in\mathcal{X}_1} P(x|x\in\mathcal{X}_1)\log P(x|x\in\mathcal{X}_1) \\ &- \left(\sum_{x\in\mathcal{X}_2} P(x)\right)\sum_{x\in\mathcal{X}_2} P(x|x\in\mathcal{X}_2)\log P(x|x\in\mathcal{X}_2) \end{aligned}$$

is a weighted sum of the entropies associated with the choice of the event inside each subset.

For a more detailed description of the entropy as well as proof of the above properties see for instance [CoverT 91].

3.1 The Kullback-Leibler divergence...

The above definition of the entropy gives rise to a number of derived concepts. Here we will focus on one of them of particular importance for inverse problems. As we have described in the beginning of this chapter, inverse problems are about *learning* the model that generated a particular set of measured data. Since we are interested in probabilistic models what we need to learn is the distribution of the original model. That is to find a distribution which is as "close" as possible to the distribution of the data. For this we need a notion of distance between distributions. The interpretation of entropy as a measure of information leads to a way to evaluate such a "distance" by means of the *Kullback-Leibler divergence* [KullbackL 51]

$$D_{\mathrm{KL}}(P\|Q) \equiv \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = -\left(H_P(X) - \mathbb{E}_P[\log\frac{1}{Q(x)}]\right) \quad . \tag{3.1}$$

It is the negative difference between the entropy of the first distribution on one hand and the average, with respect to the first distribution, of the self-information of the second distribution on the other hand. Thus, it measures how well distribution Q captures the probabilistic structure of distribution P. It is not a true distance for a number of reasons, *e.g.* it is not symmetric nor does it satisfy the triangle inequality, but can nonetheless be thought in a similar way since $D_{\mathrm{KL}}(P\|Q) \geq 0$, the equality holding only when $P = Q$. As for the entropy defined earlier, the log is usually taken to base 2 in information theoretic contexts while the natural logarithm is used in statistical physics contexts. Here we will use the natural logarithm.

The Kullback-Leibler divergence plays a central role in the context of inverse problems as it is used to measure how well the inferred distribution Q models the original distribution P[i]. In fact we will see that *all the inverse Ising methods presented in this thesis can be viewed as minimizing the Kullback-Leibler divergence between the original distribution* $P(\underline{s})$ *and some*

i. In this context we will use the words *model* or *trial* distribution for Q and *true* distribution for P.

model distribution $Q(\underline{s})$. The methods may differ in the order in which P and Q are placed in $D_{\mathrm{KL}}(\,\cdot\,\|\,\cdot\,)$. The distribution occupying the first position is the one used to take the average of $\log\frac{P}{Q}$. It is very often impossible to take averages with respect to an Ising distribution theoretically. On the other hand, in the context of the inverse Ising problem, this is done empirically by averaging over the measured data, which is easy. So, if one wants to build directly an inverse method, one must use $D_{\mathrm{KL}}(P\|Q)$ and minimize it over the models Q. As an alternative, one can build a method for the direct problem, using a model distribution such that averages can be carried easily [ii] and then inverse the equations to get an inverse method. In that case, one must use $D_{\mathrm{KL}}(Q\|P)$. This second approach is used in the mean field methods described later in this chapter.

An alternative way of looking at the Kullback-Leibler divergence is by means of the thermodynamic potentials. For instance $D_{\mathrm{KL}}(Q\|P)$ can be rewritten as

$$D_{\mathrm{KL}}(Q\|P) = U[Q] - S[Q] + \log Z \equiv \mathbb{F}[Q] - F \quad . \tag{3.2}$$

The first functional is the *internal energy* defined as $U[Q] \equiv \langle \mathcal{H}_P \rangle_Q$. $\mathcal{H}_P$ is the Hamiltonian of the true distribution. The second functional is the *entropy* [iii] defined as usual as $S[Q] \equiv \langle -\log Q \rangle_Q$. Their difference $\mathbb{F}[Q] \equiv U[Q] - S[Q]$ will be called the *free energy functional* to distinguish it from the usual free energy $F \equiv -\log Z$. The KL divergence can be expressed as a difference of these two potentials which means that $\mathbb{F}$ is greater than F, in general, and achieves its minimum value $\mathbb{F} = F$ when $Q = P$.

3.2 ...and its relation to the log-likelihood

Another widely used approach for inverse problems is the maximization of the *log-likelihood*. We will briefly outline this method and demonstrate its equivalence to the Kullback-Leibler divergence minimization. The main idea is to use Bayes theorem in order to write an expression for the likelihood of a model given the fact that we have observed a number of outcomes. Writing θ for the model parameters and x for *one* observed outcome we have

$$Q(\theta|x) = \frac{Q(x|\theta)Q(\theta)}{Q(x)} \quad . \tag{3.1}$$

$Q(x|\theta)$ is just the distribution we want to infer. $Q(\theta)$ is a prior distribution over all models. It can be just a uniform distribution over all sets of model parameters, but we will later see that one can use it to restrict the search to a particular, relevant class of models and thus achieve more efficient algorithms. $Q(x)$ is of no consequence in the present case since it doesn't depend on the model parameters. The next step is to maximize the above function, with respect to θ, in order to find the most likely model capable of generating the data we observed. If one has a set of independent measurements the joint likelihood is just the product of the individual one-measurement likelihoods. This is where the convenience of taking the logarithm of the

ii. Some particular classes of factorized distributions are usually used, where efficient inference can be done either exactly or approximately.

iii. We respect the convention of noting H the entropy in information theory contexts and S in statistical mechanics contexts.

likelihood becomes apparent. The logarithm is a monotonically increasing function and thus the logarithm of a function achieves its maximum in the same place as the function itself, hence it does not affect maximum likelihood estimation. On the other hand the logarithm transforms the product of the individual likelihoods in a sum of log-likelihoods which is easier to manipulate. The log-likelihood for p independent measurements reads

$$\mathcal{L} = \frac{1}{p} \sum_{\mu=1}^{p} \log Q(\theta|x^{(\mu)}) \quad . \tag{3.2}$$

It is easy to show that maximizing the above quantity is equivalent to minimize the Kullback-Leibler divergence between two appropriate distributions: ideally, when $p \to \infty$, the average over the measurements in eq.(3.2) can be rewritten as an average over the unknown distribution

$$\mathcal{L} = \sum_{x \in \mathcal{X}} P(x) \log Q(\theta|x) = \mathbb{E}_P[\log Q(\theta|x)] \quad . \tag{3.3}$$

Comparing the above equation with eq.(3.1) we conclude that maximizing the likelihood is equivalent to minimize the Kullback-Leibler divergence between the unknown distribution and the model distribution weighted by a prior distribution of models.

Because of their ultimate connexion with entropy, both the Kullback-Leibler divergence and the likelihood function are measures of our knowledge about the probabilistic structure of the unknown system, hence they provide a natural framework for inverse problems. All methods presented in this thesis have either one of the two approaches as a starting point. They differ only in the procedure used to find the extremum and in the restrictions about the target models. However, we will see that these differences can be crucial to the performance of the respective algorithms, both in terms of computational complexity and in terms of their precision. But before reviewing all those methods, let us take a few more paragraphs to state the problem and introduce some notation that will be used for the rest of this work.

Chapter 4

Formulation of the problem

The *symmetric Ising model* is defined by the distribution

$$P(\underline{s}) = \frac{1}{Z_\beta(J,H)} \mathrm{e}^{\beta\left(\sum_i H_i s_i + \sum_{i<j} J_{ij} s_i s_j\right)} \tag{4.1}$$

$$Z_\beta(J,H) = \sum_{\underline{\sigma}} \mathrm{e}^{\beta\left(\sum_i H_i \sigma_i + \sum_{i<j} J_{ij} \sigma_i \sigma_j\right)} \tag{4.2}$$

where $\underline{s} = (s_1, \ldots, s_N)$ is the spin vector, J is the *couplings matrix* and H is the *local fields vector*.

The inverse problem can be stated as follows. Given a set of p spin configurations $\mathcal{S} = \{\underline{s}^{(1)}, \ldots, \underline{s}^{(p)}\}$ generated from a model $\mathcal{M} = (J,H)$, find parameters J_{ij} and H_i.

In the present work we will focus only in Ising models with one and two body interactions (local fields and couplings), thus the distribution can be fully characterized by the first and second moments

$$m_i \equiv \langle s_i \rangle \text{ and } C_{ij} \equiv \langle s_i s_j \rangle - m_i m_j \ . \tag{4.3}$$

referred to hereafter as the *magnetizations* and *correlations* respectively. It is customary to consider that the inverse Ising problem algorithms accept these quantities as inputs instead of the raw data $\mathcal{S}$. If the exact values of m_i and C_{ij} are known an inverse Ising algorithm could potentialy yield exactly the values of H_i and J_{ij}. When this is the case we say that those algorithms are *exact*. In practice however, one cannot know *a priori* the exact magnetizations and correlations so they must be estimated from the configurations $\mathcal{S}$

$$m_i = \frac{1}{p}\sum_{\mu=1}^{p} s_i^{(\mu)} \text{ and } C_{ij} = \frac{1}{p}\sum_{\mu=1}^{p} s_i^{(\mu)} s_j^{(\mu)} - m_i m_j \ . \tag{4.4}$$

In this case, the noise due to the finite number of configurations used in the above estimation will yield errors in the estimation of the model parameters, even if an exact algorithm is used. We will still use the term "exact algorithm" however, understanding that it would yield exact results if feeded with the exact m_i's and C_{ij}'s.

The computation of the empirical correlation matrix takes $\mathcal{O}(N^2 p)$ time so any algorithm using the correlations as a starting point will be at least that slow, but this is a harmless

constrain since any algorithm that estimates N^2 quantities (the couplings) using p measurements will take *at least* that time. The inverse Ising problem can be written schematically as $(m, C) \to (H, J)$.

4.1 Graphical models

In many systems of interest, where a large number of random variables is involved, the pattern of their mutual dependencies is often non-trivial. In mathematical language this means that their joint distribution (the Boltzmann distribution of eq. (4.2) in our case) can be decomposed into a product of different factors each containing a subset of variables. As we will encounter in later sections of this chapter this decomposition might play an important role in the solvability of both the direct and inverse problems. The physical origin of this feature is the local nature of physical interactions. In real spin systems spins usually interact only within some neighborhood of finite radius, neurons in the brain create synapses with a limited number of other neurons, genes might influence the expression of a limited number of other genes and so on. To highlight this property and treat it more easily it is convenient to represent graphically the structure of the dependencies or interactions. This is done usually with the help of a *graph*.

A graph G is an set of *nodes* or *vertices* V together with a set of *edges* or *links* E which themselves are sets of pairs of nodes[i]. The graph, written $G = (V, E)$, is then associated with a probability distribution that can be put in a factorized form. For pairwise systems we will use the following notation

$$P(\underline{x}) = \prod_i \psi_i(x_i) \prod_{(ij)} \psi_{ij}(x_i, x_j) \quad , \tag{4.1}$$

where we also allow the possibility for one-body interactions (local fields). The nodes in V are in one to one correspondence with the variables in $\underline{x} = (x_1, \ldots, x_N)$ and each edge, $(ij) \in E$ with $i, j \in V$, represent the factor $\psi_{ij}(x_i, x_j)$.

The structure of the interaction graph in real systems is, as we said before, non-trivial this is why in theoretical works they are often treated as random objects themselves, besides the randomness of the couplings and local fields. From the point of view of the inverse Ising problem it might be an important question to infer the graph structure, *i.e.* the edge-set E, as a first step before inferring the couplings and local fields J and H. We will see for instance in section 6 that inferring the graph structure first can make the task of finding the couplings particularly easy.

i. In general, a graph can contain more than one kind of nodes such as the case of factor graphs where factor nodes together with variable nodes are used to represent multiple variables factors or, in the physics jargon, many-body interactions see e.g. [MezardM 09]. In the present work only pairwise interactions are considered so simple graphs with only variable nodes are sufficient.

Chapter 5

The Boltzmann machine and its training

A Boltzmann machine is a type of stochastic network, invented by Geoffrey Hinton and Terry Sejnowski [AckleyHS 85, Hinton 89], and named after the Boltzmann distribution. From the perspective that we are interested in it is just an Ising model together with a Monte Carlo dynamics. The introduction of the Hopfield model in 1982 [Hopfield 82] sparked interest in networks capable of storing knowledge in the structure of their connexions. Both these models use a Boltzmann distribution together with an Ising Hamiltonian for their purposes, the difference being that, whereas in a Hopfield network a deterministic gradient descent is performed in order to retrieve a memorized "pattern", in a Boltzmann machine the system is left to "wander" stochastically in configuration space and thus generate sets of plausible configurations. This "wandering" is done by the Metropolis algorithm [MetropolisRRT$^+$ 53] so that the generated configurations are distributed according to the desired Boltzmann distribution. The thermal noise of the Metropolis algorithm enables the system to escape from local minima and hopefully to find the global minimum of the energy function. In the Hopfield model this local minima trapping is not a problem since the energy landscape has been designed so that the local minima correspond to memorized patterns. In fact getting stuck in some local minimum amounts in retrieving a memorized pattern in a Hopfield model so it is a desired behavior. However, in constrain satisfaction problems, where one wants to minimize the total number of violated constrains the search for the global minimum is essential, hence the stochastic nature of the Boltzmann machine.

An other potential application of such systems would be to train them to generate data similar to some particular dataset. For instance, once the system has been trained using a set of pictures, it could be used to complete an other picture which is partially missing. This leads us to the central problem of the Boltzmann machine applicability: its *training*. In this context training means inferring the couplings and local fields (weights and biases in the Boltzmann machine jargon) of the system that generated the particular dataset of interest. In other words training means solving the inverse Ising problem. Moreover, if the original system was not an Ising model, we could wish to find couplings and fields such that an Ising model would generate similar examples as the ones we used in the training. It turns out that there is a simple set of learning rules, as shown in [AckleyHS 85], that leads the system gradually to adopt the correct

values of the couplings and fields. We will now outline the derivation of this procedure.

Let's say that we have a set of data $\mathcal{S} = \{\underline{s}^{(1)}, \ldots, \underline{s}^{(p)}\}$ generated from an unknown distribution $P_0(\underline{s})$. We want to find Ising model parameters $\mathcal{M} = (J, H)$ such that the distribution of the inferred model $P_{J,H}$ minimizes the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(P_0, P_{J,H}) = \sum_{\underline{s}} P_0(\underline{s}) \log \frac{P_0(\underline{s})}{P_{J,H}(\underline{s})} \quad . \tag{5.1}$$

We differentiate with respect to J_{ij}

$$\begin{aligned} \frac{\partial D_{\mathrm{KL}}(P_0, P_{J,H})}{\partial J_{ij}} &= -\sum_{\underline{s}} P_0(\underline{s}) \left(s_i s_j - \frac{1}{Z} \sum_{\underline{\sigma}} \sigma_i \sigma_j \mathrm{e}^{\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i H_i \sigma_i} \right) & (5.2) \\ &= -\left(\langle s_i s_j \rangle_{\mathcal{S}} - \langle s_i s_j \rangle_{\mathcal{M}} \right) \quad , & (5.3) \end{aligned}$$

where $\langle \cdot \rangle_{\mathcal{S}}$ means average over the data and $\langle \cdot \rangle_{\mathcal{M}}$ average with respect to inferred model. This leads to the following learning rule by gradient descent

$$\delta J_{ij} = \epsilon \left(\langle s_i s_j \rangle_{\mathcal{S}} - \langle s_i s_j \rangle_{\mathcal{M}} \right) \quad . \tag{5.4}$$

ϵ defines the rate of the learning process. This rule is very simple, it means that the learning process adjust the couplings of the inferred model gradually until its correlations match the empirical ones we have from the data. Similarly for the local fields we get

$$\delta H_i = \epsilon \left(\langle s_i \rangle_{\mathcal{S}} - \langle s_i \rangle_{\mathcal{M}} \right) \quad . \tag{5.5}$$

After each update of the couplings and fields one must compute the theoretical predictions of the model's correlations and magnetizations in order to compute the next set of corrections. It can be shown that, if there are no hidden spins[i], D_{KL} is a convex function of the model parameters [AckleyHS 85, Hinton 89]. This guarantees that the simple gradient descent described above will eventually reach the global minimum. If the data where generated by an Ising model the global minimum has $D_{\mathrm{KL}} = 0$ and therefore the learning algorithm will recover the model parameters exactly. However, computing the model correlations and magnetizations is not easy. Since exact inference is NP-hard in the general setting [Cooper 90] one must turn to Monte Carlo simulations. Therefore, after each update, the system is simulated for a number of steps until it reaches thermal equilibrium. At low temperatures, the simulated system can get stuck in some local minimum, and spend a lot of time until it escapes. Since there is an exponential number of local minima [MezardPV 87] in the low temperature phase, we need an exponential number of Monte Carlo steps in order to explore the phase space sufficiently. We see that we have an exact algorithm at the cost of an unfeasibly high computational complexity. This is a first example of the interplay between the precision of the method and its complexity. Shortly we will encounter algorithms that have lower complexities but infer approximately the model parameters.

i. Originally the Boltzmann machine was conceived with the possibility of having hidden "units" to account for constrains in the data that cannot be explained solely with pairwise interactions.

Chapter 6

Exact learning on trees

The ambitious Boltzmann machine learning fails to provide a practical algorithm because of the inefficiency of inference in general. The source of this hardness are the numerous loops of the underlying factor graph [i], who force us to treat the system globally. Precisely for that reason, the exemplar class of "easy" models are *tree models*. Trees are the only structures that allow exact inference through local computations [Pearl 88]. This is why they have been widely used as approximations in inference problems. In this section we will show how restricting the target models to trees decreases drastically computation time. Since in most potential applications of inverse Ising methods the underlying networks are not trees, the method described here is just meant to demonstrate how prior knowledge on the system can alter the complexity of a learning algorithm. If the original system is not a tree, then the method described here guarantees to find the optimal product approximation [ii] having the structure of a tree, although it is a crude approximation since many interdependencies between variables will be ignored.

We can divide the problem into two parts. First, inferring the correct graph and, second, inferring the couplings and local fields. For the first part we will present a particularly elegant method invented by C. K. Chow and C. N. Liu in 1968 [ChowL 68], which makes use of a well known graph theory algorithm, the *Maximum Spanning Tree* (MST) algorithm. We will present their method in general, without a reference to a particular type of distribution and then we will show a possible variation, specially adapted to Ising models.

6.1 The Chow-Liu Method

First we need to introduce the notion of *mutual information* of two variables X_i and X_j

$$I_{ij}(X_i; X_j) \equiv \sum_{x_i, x_j} P_{ij}(x_i, x_j) \log \frac{P_{ij}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \quad , \tag{6.1}$$

where P_{ij} and P_i, P_j are the two and one variable marginals of the total distribution. This quantity is the Kullback-Leibler divergence between the joint distribution of X_i and X_j and the product of their marginals, and thus quantifies their lack of independence.

i. see section 4.1

ii. A product approximation of a distribution is a product of several of its marginals

First we need to introduce some facts about tree-graphical models. It is well known [MezardM 09] that the joint probability distribution of a tree model can be factorized in terms of local marginals. Precisely, for any tree

$$P_t(\underline{x}) = \prod_{(ij)\in E_t} P_{ij}(x_i, x_j) \prod_{i\in V} P_i(x_i)^{1-|\partial i|} \quad , \tag{6.2}$$

where the ensemble of edges E_t is chosen such that $|E_t| = |V| - 1$ and the graph is simply connected. These two requirements guarantee that the graph contains no loops, *i.e.* that it is a tree. As a consequence of this decomposition in local terms the entropy can be decomposed as well

$$H[P_t] = -\sum_{(ij)\in E_t} \sum_{x_i,x_j} P_{ij}(x_i, x_j) \log P_{ij}(x_i, x_j) - \sum_{i\in V} (1 - |\partial i|) \sum_{x_i} P_i(x_i) \log P_i(x_i) \quad . \tag{6.3}$$

Other extensive quantities can be decomposed in a similar way.

As we did for the Boltzmann machine in the previous section, we take the Kullback-Leibler divergence between the true distribution $P(\underline{x})$ and our tree model distribution $P_t(\underline{x})$

$$\begin{aligned} D(P\|P_t) &= \sum_{\underline{x}} P(\underline{x}) \log \frac{P(\underline{x})}{\prod_{(ij)\in E_t} P_{ij}(x_i, x_j) \prod_{i\in V} P_i(x_i)^{1-|\partial i|}} \\ &= \sum_{\underline{x}} P(\underline{x}) \log P(\underline{x}) - \sum_{\underline{x}} P(\underline{x}) \left[\sum_{(ij)\in E_t} \log P_{ij}(x_i, x_j) \right] + \sum_{\underline{x}} P(\underline{x}) \left[\sum_{i\in V} (|\partial i| - 1) \log P_i(x_i) \right] \\ &= -H(\underline{X}) - \sum_{x_i,x_j} P_{ij}(x_i, x_j) \left[\sum_{(ij)\in E_t} \log \frac{P_{ij}(x_i, x_j)}{P(x_i)P(x_j)} \right] + \sum_{x_i} P_i(x_i) \left[\sum_{i\in V} (|\partial i| - 1) \log P(x_i) \right] \\ &\quad - \sum_{(ij)\in E_t} \left[\sum_{x_i} P_i(x_i) \log P_i(x_i) + \sum_{x_j} P_j(x_j) \log P_j(x_j) \right] \end{aligned} \tag{6.4}$$

Now, it is clear that, in the last sum of the above equation, each variable will yield an entropic term $|\partial i|$ times, so that some of the terms will cancel out with terms of the preceding sum. Hence, the result is

$$D(P\|P_t) = -H(\underline{X}) + \sum_{i\in V} H(X_i) - \sum_{(ij)\in E_t} I_{ij}(X_i, X_j) \tag{6.5}$$

The first two terms are independent of any particular tree structure and, since $I(x_i, x_j)$ is non negative, minimizing the Kullback-Leibler divergence is equivalent to maximizing the quantity $\sum_{(i,j)\in E_t} I(x_i, x_j)$. Thus, the edge-set of the optimal tree is given by

$$E_t^* = \underset{E_t=\text{Tree}}{\arg\max} \left\{ \sum_{(i,j)\in E_t} I(x_i, x_j) \right\} \tag{6.6}$$

Although the space of trees is much smaller than the space of all possible graphs, it is still huge. We know from the Cayley formula that there are N^{N-2} distinct trees. Fortunately, we don't need to consider exhaustively all these possibilities to solve the above maximization problem.

The important result of Chow and Liu is that the quantity to be maximized is a sum of local terms, so that a simple greedy algorithm can solve the problem.

Given a weighted graph, the problem of finding a tree spanning all nodes and having maximum total weight is well known in graph theory. A popular algorithm is Prim's one [Prim 57]. For a full description of the algorithm and the proof of its correctness we refer the reader to the literature. Here we give only its description:

Algorithm 6.1.1: MAXIMUM SPANNING TREE(W)

$V \leftarrow \{1\}$, Any node can be used in the initialization without affecting the result
$E \leftarrow \emptyset$
while $|V| \neq N$
do $\begin{cases} (ij) \leftarrow \arg\max_{i \in V,\, j \notin V} W_{ij} \\ V \leftarrow \{V, j\} \\ E \leftarrow \{E, (ij)\} \end{cases}$
return (V, E)

Here W is the weights matrix. Returning to our problem, if we use the weights $W_{ij} = I(x_i, x_j)$ the maximum weight tree will be the one minimizing the Kullback-Leibler divergence, and hence the solution of our restricted inverse problem. In practice, one must compute the empirical estimates of the mutual information from samples and then proceed with the above algorithm, but we refer the reader to [ChowL 68] for the details.

In section 4 we stated that the usual starting point for the inverse Ising problem are the correlations and magnetizations. These can be used also in the context of the Chow and Liu method instead of the mutual information. Actually, in order for the MST algorithm to recover the correct tree, one can use as the inputing weights any quantity, defined on a pair of nodes, that is strictly decreasing with the distance between those nodes. The mutual information is an example but one can show that also the correlations decrease in absolute value with the distance. In general, although both quantities are measures of the dependence of two variables, there are not always related as the correlation captures only linear dependence whereas mutual information measures general dependence. In some special cases, however, there can be a relation between the two. It has been show in [Wentian 90] that one such case is when the variables are binary, as the spins in the Ising model. But it is quite easy to show that if correlations are used as weights, the MST algorithm will recover the correct tree, without using the above result. First we need to prove a lemma about correlation decay in Ising trees.

Lemma 6.1.1 *Let $G = (V, E)$, with $V = \{1, 2, 3\}$ and $E = \{(12), (23)\}$, be a three nodes graph and let $\theta = \{J_{12}, J_{23}, H_1, H_2, H_3\}$ be a set of Ising parameters. Then, in the Ising model defined by G and θ, $|C_{12}| \geq |C_{13}|$, where $C_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$. Moreover, in any tree, $|C_{ij}|$ is a decreasing function of d_{ij}, the distance between the nodes on the tree.*

Proof The proof is a straightforward calculation of magnetizations and correlations. We start

from their definitions

$$\langle s_i \rangle = \frac{\sum_{\underline{s}} s_i \exp\left(\sum_{(kl)\in E} J_{kl} s_k s_l + \sum_{k\in V} H_k s_k\right)}{\sum_{\underline{s}} \exp\left(\sum_{(kl)\in E} J_{kl} s_k s_l + \sum_{k\in V} H_k s_k\right)} \quad \text{and}$$

$$\langle s_i s_j \rangle = \frac{\sum_{\underline{s}} s_i s_j \exp\left(\sum_{(kl)\in E} J_{kl} s_k s_l + \sum_{k\in V} H_k s_k\right)}{\sum_{\underline{s}} \exp\left(\sum_{(kl)\in E} J_{kl} s_k s_l + \sum_{k\in V} H_k s_k\right)} \quad .$$

We then expand the sums and use the variables $\zeta_{ij} \equiv \tanh J_{ij}$ and $\eta_i \equiv \tanh H_i$

$$\begin{aligned}
\langle s_1 \rangle &= \frac{1}{Z}\left(\eta_1 + \zeta_{23}\eta_1\eta_2\eta_3 + \zeta_{12}\eta_2 + \zeta_{12}\zeta_{23}\eta_3\right) \\
\langle s_2 \rangle &= \frac{1}{Z}\left(\eta_2 + \zeta_{12}\eta_1 + \zeta_{23}\eta_3 + \zeta_{12}\zeta_{23}\eta_1\eta_2\eta_3\right) \\
\langle s_3 \rangle &= \frac{1}{Z}\left(\eta_3 + \zeta_{12}\eta_1\eta_2\eta_3 + \zeta_{23}\eta_2 + \zeta_{12}\zeta_{23}\eta_1\right) \\
\langle s_1 s_2 \rangle &= \frac{1}{Z}\left(\zeta_{12} + \eta_1\eta_2 + \zeta_{23}\eta_1\eta_3 + \zeta_{12}\zeta_{23}\eta_2\eta_3\right) \\
\langle s_1 s_3 \rangle &= \frac{1}{Z}\left(\zeta_{12}\zeta_{23} + \eta_1\eta_3 + \zeta_{12}\eta_2\eta_3 + \zeta_{23}\eta_1\eta_2\right) \quad .
\end{aligned}$$

Where Z is the partition function. Putting all that together we have

$$\left|\frac{C_{12}}{C_{13}}\right| = \frac{1-\eta_3^2\zeta_{23}^2}{(1-\eta_3^3)\zeta_{23}^2} \geq 1 \quad . \tag{6.7}$$

The equality holds when $\zeta_{23} = \pm 1$ which means that $J_{23} = \pm\infty$, so for all practical purposes it is a strict inequality.

To prove the second statement we first remark that a similar result holds for a chain of any length. Indeed, by applying recursively the above formula one can show that, for any adjacent pair (ij), $|C_{ij}|$ is greater than any $|C_{ik}|$ with k being on the same side as j. Moreover, the magnetizations and correlations of any subgraph of a tree model can always be reproduced by a model having the same structure as the subgraph, the same couplings and the appropriate local fields [iii]. Therefore, since any path on a tree can be mapped to a chain with the same couplings and different local fields, by the above result the correlations along the path will decrease with the distance. ∎

This leads us to the following proposition.

Proposition 6.1.2 *Let $T = (V, E_t)$ be a tree and let $\theta = \{\{J_{ij} : (ij) \in E_t\}, \{H_i : i \in V\}\}$ be Ising parameters defined on that tree. Suppose $K_W = (V, E_k = V \times V, W)$ is a weighted complete graph with the same node set as T and weights given by $W_{ij} = |C_{ij}|$. Then the maximum weight spanning tree T_W^* of K_W is T.*

iii. When performing averages in a tree, we can always start by summing the variables of the leaves and gradually proceed towards the subgraph in question. It turns out that in the end we incorporate the influence of the rest of the tree by just modifying the local fields of those variables that are connected with the rest of the tree. This is a direct consequence of the fact that there are no loops in a tree.

Proof For any partition of the complete graph $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$ we define a cut-set as $E_{12} = \{(ij) : i \in V_1, j \in V_2\}$. First note that $(ij)^* \equiv \arg\max_{(ij)\in E_{12}} W_{ij}$ will be necessarily in T_W^*. Indeed, if another edge $(kl) \in E_{12}$ was in T_W^* then removing that edge and adding $(ij)^*$ would produce a tree of greater total weight. Now, from lemma 6.1.1 we have that the only edge $(ij) \in E_{12} \cap E_t$ is $(ij)^*$ since all other edges in E_{12} link nodes with greater distance in E_t. ∎

6.2 The Independent Pair Approximation

The above proposition enables us to use the simpler $|C|$ as weights in the MST algorithm instead of the mutual information, confirming that, as with the Boltzmann machine learning, all the information about Ising models can be found in the correlations and magnetizations. Of course, we still need to find the precise values of the couplings and local fields but these also can be found exactly from the correlations and magnetizations once the structure of the tree has be found. It was mentioned earlier that, if *effective* local fields are chosen appropriately, any subgraph of a tree will reproduce the same moments as the whole of the tree. We can thus break down the tree to an ensemble of connected pairs and solve with respect to the couplings each pair independently. We will now outline this method.

The distribution of a pair of spins is written

$$P(s_i, s_j) = \frac{1}{Z_{ij}} \mathrm{e}^{J_{ij}s_is_j + h_i^{(j)}s_i + h_j^{(i)}s_j} \quad . \tag{6.1}$$

In the context of the whole tree, the fields present in the above equation are the effective local fields acting on each spin. They contain contributions form the actual local field and from the remaining spins of the tree, not counting the second one of the pair in question. For instance, $h_i^{(j)}$ is the sum of the local field acting on i and the field felt by i from all other spins except j. It can be interpreted thus as the total field i would feel if we remove spin j from the graph. We can solve the above equation with respect to J_{ij}

$$J_{ij} = \frac{1}{4} \ln \left(\frac{P_{++}P_{--}}{P_{+-}P_{-+}} \right) \quad , \tag{6.2}$$

where $P_{++} \equiv P(+1, +1)$, $P_{+-} \equiv P(+1, -1)$ etc. Then we can express those probabilities in terms of magnetizations and correlations

$$J_{ij} = \frac{1}{4} \ln \left[\frac{((1+m_i)(1+m_j) + C_{ij})\,((1-m_i)(1-m_j) + C_{ij})}{((1+m_i)(1-m_j) - C_{ij})\,((1-m_i)(1+m_j) - C_{ij})} \right] \quad . \tag{6.3}$$

Once the couplings have been found we can compute the effective local fields for each pair by a similar formula

$$h_i^{(j)} = \frac{1}{2} \ln \left[\frac{(1+m_i)(1-m_j) - C_{ij}}{(1-m_i)(1-m_j) + C_{ij}} \right] \quad . \tag{6.4}$$

Now it is easy to find the actual local fields. We first note that $h_i^{(j)} = \tilde{h}_i^{(j)} + H_i$ where $\tilde{h}_i^{(j)}$ is the contribution on i's total field from all spins except j without counting H_i. If we sum $\tilde{h}_i^{(j)}$

over j we will over-count each spin's contribution $|\partial i| - 1$ times, where $|\partial i|$ is the number of neighbors of spin i. Thus we can write

$$H_i + \frac{1}{|\partial i| - 1} \sum_j \tilde{h}_i^{(j)} = \text{atanh}(m_i) \quad . \tag{6.5}$$

On the other hand if we sum $h_i^{(j)}$ we have

$$\sum_j h_i^{(j)} = \sum_j \tilde{h}_i^{(j)} + |\partial i| H_i \quad . \tag{6.6}$$

Combining the above equations we finally have

$$H_i = \sum_j h_i^{(j)} - (|\partial i| - 1)\text{atanh}(m_i) \quad . \tag{6.7}$$

Equations (6.3,6.4,6.7) can be used to find the couplings and fields of the tree model once the graph has been found. They form the so called *Independent Pair approximation* [RoudiTH 09, RoudiAH 09].

Although exact only on trees, the method described in this chapter can be applied on any model to find approximate solutions. The original paper proposes the MST method as a way to find the best tree model approximating any kind of distribution. This can have some advantages as trees allow efficient and exact inference. They can reproduce, however, a very limited spectrum of behaviors. In many well known systems, such as the Hopfield model or the Boltzmann machine, the richness of their behavior comes from the existence of metastable states. Such states require the existence of frustrated loops and cannot be realized in tree graphs. Moreover, since we are mostly interested in the inverse Ising problem as a method for inferring the actual structure of biological networks, this method is not well suited since such systems are typically not trees. As for the Chow-Liu method, the independent pair approximation can also be used in cases where the underlying graph is not a tree but with limited results. The assumption that every pair can be treated independently is valid only in weak couplings (high temperature) settings, where the correlations decay importantly beyond adjacent pairs.

In the next sections we will examine more realistic methods, able to capture the structure of more complex networks. None of them is a panacea as they all have limited regimes of applicability. Yet they are much more powerful as they can be be applied to more realistic systems with satisfactory results. The first family of methods comes from the class of *mean field* methods, well known in statistical physics.

Chapter 7

Mean field methods

Mean Field Theory (MFT) is a general term to describe a whole family of methods, primarily in statistical mechanics, whose aim is to solve a many-body problem by replacing it with an effective one-body one. In general, many-body problems are very difficult to solve exactly because the combinatorial "explosion", due to the great number of degrees of freedom, forbids the full enumeration of states. In many cases, however, the lack of important fluctuations allows a powerful simplification: since the "environment" seen by each degree of freedom doesn't vary a lot, on can replace the full system with a unique field acting on one degree of freedom. This field, called *effective* or *molecular* field, is the average field created by the rest of the system, hence the name. In some cases, despite the simplification, the main features of the system's behavior are reproduced by the mean field equations, thus MFT can provide important insights at a low cost since one-body problems are usually much easier. For instance, in a ferromagnetic Ising model, one can ignore the full system and study the behavior of one spin in the presence of the combined field of all other spins, and still predict the existence of a paramagnetic/ferromagnetic phase transition, although the predicted behavior around the critical point will be wrong. By considering the average field created by the other spins, MFT completely ignores their fluctuations and thus can be seen as a zeroth-order expansion of the Hamiltonian in fluctuations. Moreover, dimensionality plays an important role in determining the applicability of MFT. In high dimensional systems (or in the equivalent long range limit) spins "feel" the presence of a great number of neighbors whose fluctuations thus become negligible. However, unlike the ferromagnetic case, in the spin-glass Ising model fluctuations are important [i] and the naive mean field theory fails. A much better result can be obtained by taking into account the first-order term in fluctuations. The resulting equations are known as Thouless-Anderson-Palmer (TAP) equations. Since many of the methods presented in this thesis are mean-field methods we will reserve the name naive mean field (NMF) method for the simplest one, described in the next section.

In this chapter we will outline the derivation of mean-field methods. We will then present the way they have been used to solve the inverse Ising problem as was done in [KappenR 97, KappenR 98, Tanaka 98]. The same rationale, as the one we find in NMF and TAP methods, has been used in a number of similar contexts such as kinetic models [KappenS 00,

i. Because of the cancellations due to the random couplings, the average field is much weaker and thus its fluctuations are crucial in determining the behavior of the system.

HertzRTT$^+$ 10, RoudiH 11a, RoudiH 11b, ZengAAM 10] but we will introduce these results in part III where we will discuss kinetic Ising models. For the sake of coherence with the previous sections we will derive the mean field equations starting from a Kullback-Leibler divergence minimization problem, although they were originally derived using physical arguments. We will first derive and discuss the equations for the direct problem and then inverse them and present some results of their application to the inverse problem.

7.1 Naive Mean Field Approximation

In a mean field model we consider that each spin is independent of all others and feels a local effective field composed of two terms

$$h_i = \tilde{h}_i + H_i \quad . \tag{7.1}$$

As in the previous section, H_i is the actual local field of the full model and $\tilde{h}_i$ is the contribution of the other spins. The mean field measure thus reads

$$P_{\rm mf}(\underline{s}) = \prod_i \frac{\exp(h_i s_i)}{2\cosh(h_i)} \quad . \tag{7.2}$$

The main advantage of having decoupled spins is that magnetizations can be computed easily. The mean field magnetizations are just $m_i \equiv \langle s_i \rangle_{\rm mf} = \tanh(h_i)$. We need to fix the values of the $\tilde{h}_i$'s so that they agree as much as possible with the full model. This is done by minimizing the Kullback-Leibler divergence between measure (7.2) and the one of the Ising model (4.2). In this case we will derive a set of equations for solving the direct problem and then invert them. Hence we will change the order of the trial and true distribution appearing in the KL divergence $D_{\rm KL}(p\|q) = \left\langle \log \frac{p}{q} \right\rangle_p$, as it was explained in section 3.1.

We have

$$\begin{aligned} D_{\rm KL}(P_{\rm mf}\|P_{\rm Ising}) &= \sum_i h_i m_i - \sum_i \log 2\cosh(h_i) - \sum_{i<j} J_{ij} m_i m_j - \sum_i H_i m_i + \log Z \\ &= \sum_i \tilde{h}_i m_i - \sum_i \log 2\cosh(h_i) - \sum_{i<j} J_{ij} m_i m_j + \log Z \end{aligned} \tag{7.3}$$

The extremization then gives

$$\frac{\partial D_{\rm KL}}{\partial \tilde{h}_i} = \left(1 - m_i^2\right)\left(\tilde{h}_i - \sum_{j\neq i} J_{ij} m_j\right) = 0 \quad . \tag{7.4}$$

Hence the effective field is given by $\tilde{h}_i = \sum_{j\neq i} J_{ij} m_j$ and so the mean field equations read

$$m_i = \tanh\left(H_i + \sum_j J_{ij} m_j\right) \quad , \tag{7.5}$$

which is nothing but an approximation to the true magnetization $\left\langle \tanh\left(H_i + \sum_j J_{ij} s_j\right)\right\rangle$ when the fluctuations have been ignored. The naive mean field model eq.(7.2) predicts zero connected

correlations since $\langle s_i s_j \rangle_{\rm mf} = m_i m_j$. However, there is a simple way of computing a non-vanishing approximation for the connected correlations $C_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$, based on the MF approach, by means of the *fluctuation-response theorem*, as was done in [KappenR 97, KappenR 98, Tanaka 98]

$$\begin{aligned} C_{ij} &= \frac{\partial m_i}{\partial H_j} = \frac{\partial}{\partial H_j} \tanh \left(H_i + \sum_j J_{ik} m_k \right) \\ &= (1 - m_i^2) \left(\delta_{ij} + \sum_k J_{ik} C_{kj} \right) \quad , \end{aligned} \tag{7.6}$$

which in matrix notation reads

$$J = L^{-1} - C^{-1} \quad , \tag{7.7}$$

where $L_{ij} \equiv (1 - m_i^2)\delta_{ij}$.

Equation (7.7) can be used to infer the couplings, in the inverse problem context, once m and C have been computed from the data. Then, one can invert eq.(7.5) to find the local fields. This procedure has been used in the contexts of Boltzmann machine learning [KappenR 97, KappenR 98, Tanaka 98], extracting the connectivity from spike trains in cortical models [RoudiTH 09], the Hopfield model [Huang 10b]. We have mentioned earlier that one necessary condition for the correctness of the naive mean field method is high dimensionality, where the average of a great number of contributions to the effective field, from the neighbors of each spin, is a good approximation for the actual fluctuating value of the later. It is not always a sufficient one, however. In the case of ferromagnets, the homogeneity of the couplings makes that all spins tend to align with each other and individual fluctuations have negligible influences in the effective field. One way to see that is to notice that, since $J_{ij} \sim \mathcal{O}(1/N)$, the variance of the individual contributions is $\mathrm{Var}(J_{ij} s_j) \sim 1/N^2(1 - m_j^2)$. Considering the approximation that all spins are independent we have for the total variance of the effective field $\mathrm{Var}(h_i) \sim \mathcal{O}(1/N) \to 0$, for $N \to \infty$. On the other hand, for spin-glasses and in order to keep to effective field of order 1, we have that $J_{ij} \sim \mathcal{O}(1/\sqrt{N})$ so that $\mathrm{Var}(J_{ij} s_j) \sim 1/N(1 - m_j^2)$. This leads to a non negligible variance for the effective field $\mathrm{Var}(h_i) \sim \mathcal{O}(1)$. We conclude that naive mean field is far from correct in general for random couplings.

There is, nevertheless, a case where it is asymptotically correct: in the high temperature limit. One way to see why this is true is to notice that $\lim_{\beta \to 0} D_{\rm KL}(P_{\rm mf} \| P_{\rm Ising}) = 0$ once the effective fields have been fixed. Indeed, in all the aforementioned applications, naive mean field provide correct results only in the high-temperature/weak-couplings limit. However, this is not a particularly relevant case since it is the limit where a network is... not a network anymore but a collection if independent spins.

7.2 The TAP equations

The failure of naive mean field in the case of spin glasses can be cured by the addition of a suitable correction. This will lead eventually to another closed set of equations for the magnetizations m_i from where we can derive a relation for the correlations by applying once

more the fluctuation-response theorem. These equations are called TAP in the literature, an acronym for Thouless, Anderson and Palmer [ThoulessAP 77] who first derived them as a mean field theory for an infinite range model with Gaussian couplings (SK model).

As we have mentioned in section 3.1 (eq.3.2), the KL divergence between a model distribution Q and the true distribution P can be written as

$$D_{\mathrm{KL}}(Q\|P) = \mathbb{F}[Q] - F \quad , \tag{7.1}$$

where $\mathbb{F}[Q] \equiv E[Q] - S[Q]$ is the free energy functional and F is the free energy of the true distribution. The divergence is minimized when $\mathbb{F}[Q]$ achieves its minimum. We minimize $\mathbb{F}[Q]$ (and thus minimize $D_{\mathrm{KL}}(Q\|P)$) by a two-stage minimization process:

The first step is to minimize $\mathbb{F}[Q]$, where the trial distributions Q are constrained by fixing the magnetizations to a set of values $\underline{m} = \langle \underline{s} \rangle_Q$. For this we define the following function

$$\mathbb{F}^*(m) = \min_Q \left\{ \mathbb{F}[Q] \mid \langle \underline{s} \rangle_Q = \underline{m} \right\} \quad . \tag{7.2}$$

The above constrained optimization problem can be transformed to an unconstrained one by introducing a set of Lagrange multipliers h_i. Now we need to minimize the following quantity

$$\mathbb{F}[Q] - \sum_i h_i(\langle s_i \rangle_Q - m_i) = E[Q] - S[Q] - \sum_i h_i(\langle s_i \rangle_Q - m_i) \quad . \tag{7.3}$$

It can be easily shown that the minimizing distribution is written

$$Q(\underline{s}) = \frac{1}{Z(h)} \mathrm{e}^{-\mathcal{H}(\underline{s}) + \sum_i h_i s_i} \quad , \tag{7.4}$$

where $\mathcal{H}(\underline{s}) = -\sum_{i<j} J_{ij} s_i s_j - \sum_i H_i s_i$ is the standard Ising Hamiltonian.

The second step is to minimize $\mathbb{F}^*(m)$ with respect to m. By introducing the minimizing distribution of eq.(7.4) back in eq.(7.2) we get the *dual optimization problem*

$$\mathbb{F}^*(m) = \max_h \left\{ \sum_i h_i m_i - \log Z(h) \right\} \quad , \tag{7.5}$$

where $Z(h) = \sum_{\underline{s}} \exp(-\mathcal{H}(\underline{s}) + \sum_i h_i s_i)$. It appears that $\mathbb{F}^*(m)$ is nothing but the *Legendre transform* of the free energy $F(h) = -\log Z(h)$ [ii]. The maximization in eq.(7.5) guarantees that

$$m_i = \frac{\partial \log Z(h)}{\partial h_i} \quad . \tag{7.6}$$

The exact computation of $\mathbb{F}^*(m)$ is as hard as the computation of the free energy $F = -\log \sum_{\underline{s}} \exp(\mathcal{H}(\underline{s}))$. It turns out, however, that we can make a perturbation expansion of $\mathbb{F}^*(m)$ around a null Hamiltonian (essentially a high temperature expansion). We replace the

ii. Usually $\mathbb{F}^*(m)$ bears the name *Gibbs free energy* although some authors use that name for the functional $\mathbb{F}[Q]$ in general.

Hamiltonian in eq.(7.5) with $\lambda\mathcal{H}(\underline{s})$ and expand. In the end we simply have to set $\lambda = 1$. The expansion gives

$$\mathbb{F}^*(m) = \mathbb{F}_0^*(m) + \mathbb{F}_1^*(m)\lambda + \mathbb{F}_2^*(m)\lambda^2 + \mathcal{O}(\lambda^3) \quad , \tag{7.7}$$

with $\mathbb{F}_n^*(m) = \frac{\partial^n}{\partial\lambda^n}\mathbb{F}(m)\Big|_{\lambda=0}$. The first two terms can be easily computed and yield

$$\mathbb{F}_0^*(m) = \sum_i \left\{ \frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2} \right\} \tag{7.8}$$

$$\mathbb{F}_1^*(m) = -\sum_{i<j} J_{ij}m_i m_j \tag{7.9}$$

They are the negative entropy and the internal energy of the naive mean field model. This is easily shown by minimizing $\mathbb{F}_0^*(m) + \mathbb{F}_1^*(m)$ with respect to the m_i's, which yields the set of naive mean field equations (7.5). The second order term is

$$\mathbb{F}_2^*(m) = -\frac{1}{2}\sum_{ij} J_{ij}^2(1-m_i^2)(1-m_j^2) \quad . \tag{7.10}$$

When this term is also taken into account the resulting extremization condition yields

$$m_i = \tanh\left(H_i + \sum_{j\neq i} J_{ij}m_j - m_i\sum_{j\neq i} J_{ij}^2(1-m_j^2) \right) \quad . \tag{7.11}$$

Equations (7.11) are the TAP equations, well known in the spin-glass literature [ThoulessAP 77]. Their original derivation was based on a, much more intuitive, cavity type argument which we will discuss soon. However, the above method recovers both naive mean field and TAP equations from a systematic expansion of the free energy, which in principle allows for improvements by adding higher order terms. This method was first proposed by Plefka [Plefka 82] how also showed the important result that, for the SK model, all terms beyond second order can be neglected, as long as the system is not in the spin glass phase [iii].

The TAP equations provide an important improvement compared to naive mean field as they take into account the effect of the fluctuations which is non-negligible in spin glasses. Indeed, in the paramagnetic phase the variance of the effective field felt by the i^{th} spin is $\text{Var}(\sum_j J_{ij}s_j) = \sum_{k,j} J_{ij}J_{ik}C_{jk} \approx \sum_j J_{ij}^2(1-m_j^2)$, which is the extra term appearing in the TAP equations. If these fluctuations are neglected one simply recovers the naive mean field equations.

The more intuitive way of looking at the TAP equations is the following: The naive mean field equations (7.5) are not correct in the spin glass case since $H_i + \sum_{j\neq i} J_{ij}m_j$ is not the true average field felt by the i^{th} spin. This field would be correct if we remove spin i from the system so that it doesn't influence the remaining ones. In spin glasses, where the effective field is weak because of the random cancellations, the influence that one spin can have on his neighbors cannot be neglected. Thus a shift in the remaining magnetizations m_j would occur caused by

iii. For Gaussian couplings with zero mean and variance $1/N$ the paramagnetic/spin-glass phase transition in the SK model occurs at $\beta_c = 1$.

the presence of spin i. The shift is determined by the *magnetic susceptibility*, *i.e.* the answer of each spin in the variation of his effective field

$$\chi_{jj} \equiv \left.\frac{\partial m_j}{\partial h_j}\right|_{h_j=0} = 1 - m_j^2 \quad . \tag{7.12}$$

Thus, each magnetization in the naive mean field equations (7.5) should be replaced by $m_j - \chi_{jj} J_{ij} m_i$ which leads to the TAP equations (7.11).

Concerning the inverse problem, a relation involving the correlations based on the TAP equations can be obtained by applying the fluctuation-response theorem as before. Differentiating eq.(7.11) with respect to m_j ($j \neq i$) gives

$$[C^{-1}]_{ij} \equiv \frac{\partial H_i}{\partial m_j} = -J_{ij} - 2 m_i m_j J_{ij}^2 \quad . \tag{7.13}$$

Solving the above equations yields the couplings who can then be used in eq.(7.11) to find the local fields.

Chapter 8

Small Correlations expansion

As long as we are in the high temperature phase and the model is fully connected (long range) all terms in the high temperature expansion, described in the previous chapter can be neglected [Plefka 82] and the TAP equations become asymptotically exact. As we depart, however, from these conditions higher order terms become relevant. The application of a high temperature expansion in disordered spin systems, first proposed by Plefka [Plefka 82], was extended up to order $\mathcal{O}(\beta^4)$ in [GeorgesY 91] and could be used to improve the accuracy of the TAP result. Inspired by this work, V. Sessak and R. Monasson noticed that, when a similar expansion is performed, a fraction of the resulting terms can be put in closed form [SessakM 09]. The central object of their construction is not the Gibbs free energy of eq.(7.5) as in the derivation of the TAP equations, but the entropy of the Ising model at fixed magnetizations and correlations. It turns out that their result can be viewed as a corrected version of the independent pair approximation of chapter 6. In this chapter we will outline the derivation of their result without going into the details of the calculations. We refer the reader to the original paper [SessakM 09] for the complete presentation.

The starting point is, as usual, the minimization of the KL divergence of eq.(7.1) through the minimization of the free energy functional $\mathbb{F}[Q]$. Except that now, instead of fixing just the magnetizations and performing a high temperature expansion, we fix also the correlations by an additional Legendre transform and then we perform a small correlations expansion. The resulting potential is the following entropy

$$\begin{aligned} S(m,C) &= \min_{J,H}\left\{\log Z(J,H) - \sum_{i<j} J_{ij}(C_{ij}+m_i m_j) - \sum_i H_i m_i\right\} \\ &= \min_{J,H}\left\{\log \sum_{\underline{s}} \exp\left(\sum_{i<j} J_{ij}[(s_i-m_i)(s_j-m_j)-C_{ij}] + \sum_i h_i(s_i-m_i)\right)\right\} \end{aligned} \quad (8.1)$$

where $h_i = H_i + \sum_j J_{ij} m_j$ is the usual mean effective field. The minimization guarantees that

$$m_i = \frac{\partial \log Z}{\partial H_i} \quad \text{and} \quad (8.2)$$

$$C_{ij} = \frac{\partial \log Z}{\partial J_{ij}} \quad . \quad (8.3)$$

Of course, by the well known duality of the Legendre transform, we can infer the couplings and fields using

$$J_{ij} = -\frac{\partial S}{\partial C_{ij}} \quad \text{and} \tag{8.4}$$

$$h_i = -\frac{\partial S}{\partial m_i} \quad , \tag{8.5}$$

once the entropy of eq.(8.1) has been computed. The minimization of S is, however, a tricky mater. By analogy of the high temperature expansion of the previous chapter, Sessak and Monasson have proposed to introduce a fictitious inverse temperature β [i] as a scale factor of the correlations and to expand the resulting entropy $S(m, \beta C)$ around $\beta = 0$. Neither this expansion is trivial, but they succeed by introducing a modified entropy who is, by construction, constant for any β while being related in a simple way with their original entropy. This allows them to compute up to order $\mathcal{O}(\beta^4)$ the entropy and then, using eq.(8.4,8.5), the couplings and fields. After defining the auxiliary quantities

$$L_i \equiv 1 - m_i^2 \quad \text{and} \quad K_{ij} \equiv (1 - \delta_{ij}) \frac{C_{ij}}{L_i L_j} \tag{8.6}$$

their result for the entropy is given by

$$\begin{aligned} S &= -\sum_i \left[\frac{1+m_i}{2} \log \frac{1+m_i}{2} + \frac{1-m_i}{2} \log \frac{1-m_i}{2} \right] \\ &- \frac{\beta^2}{2} \sum_{i<j} K_{ij}^2 L_i L_j + \frac{2\beta^3}{3} \sum_{i<j} K_{ij}^3 m_i m_j L_i L_j + \beta^3 \sum_{i<j<k} K_{ij} K_{jk} K_{ki} L_i L_j L_k \\ &- \frac{\beta^4}{12} \sum_{i<j} K_{ij}^4 [1 + 3m_i^2 + 3m_j^2 + 9m_i^2 m_j^2] L_i L_j - \frac{\beta^4}{2} \sum_{i<j} \sum_k K_{ik}^2 K_{kj}^2 L_i L_j L_k^2 \\ &- \beta^4 \sum_{i<j<k<l} [K_{ij} K_{jk} K_{kl} K_{li} + K_{ik} K_{kj} K_{jl} K_{li} + K_{ij} K_{jl} K_{lk} K_{ki}] L_i L_j L_k L_l \\ &+ \mathcal{O}(\beta^5) \end{aligned} \tag{8.7}$$

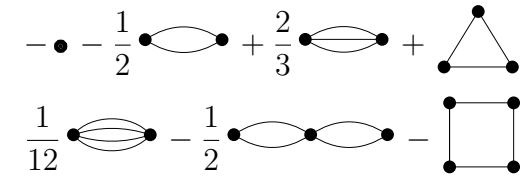
which leads to the following relation for the couplings

$$\begin{aligned} J_{ij} &= \beta K_{ij} - 2\beta^2 m_i m_j K_{ij}^2 - \beta^2 \sum_k K_{jk} K_{ki} L_k \\ &+ \frac{\beta^3}{3} K_{ij}^3 [1 + 3m_i^2 + 3m_j^2 + 9m_i^2 m_j^2] + \beta^3 \sum_{k \neq i,j} K_{ij} [K_{jk}^2 L_j + K_{ki}^2 L_i] L_k \\ &+ \beta^3 \sum_{\substack{k \neq i \\ l \neq j}} K_{jk} K_{kl} K_{li} L_k L_l \\ &+ \mathcal{O}(\beta^4) \quad . \end{aligned} \tag{8.8}$$

As we can see in the simulations run by the authors, the additional terms of the expansion indeed improve the accuracy of the inferred couplings up to a particular value of the inverse

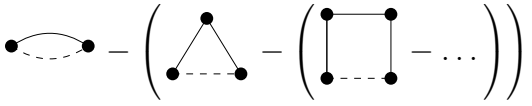
i. We will use the same symbol, although it doesn't correspond to the standard inverse temperature of the Ising model.

temperature at which the data were generated. However, beyond that point the inference error becomes worse as we take into account higher order terms. The reason is that the right hand side of eq.(8.8) contains a series of alternating signs which diverges beyond some radius of convergence. To make this point clear we must first identify the corresponding terms in the expansion of the entropy eq.(8.7). If we write the expression of the entropy in terms of Feynman diagrams, using vertices for spins and edges for the correlation terms K_{ij}, we have

$S(m, \beta C) \;=\;$ 

(8.9)

where the sums are implicit. If we isolate the terms containing simple loops we have the alternating series we were looking for. Their contribution in the expression for the couplings is the one that creates the divergence beyond some value of the inverse temperature. We can rewrite this contribution for the couplings in terms of Feynman diagrams as

$J_{ij}^{(\text{loop})} =$  , (8.10)

where dashed lines corresponds to the correlation term that is missing because of the differentiation (see eq.(8.4)). One possible interpretation of eq.(8.10) is the following: to a first-order approximation the coupling is given by the correlation of the pair in question. But, that way, we overestimate it because of the stronger correlation induced by neighboring spins, coupled with that pair. So we must subtract the correlations of 3-spins paths. But again, we are overestimating them because of the correlations of 4-spins paths and so on.

The authors noticed that this series can be re-summed. To clarify this point they give a simple example from the Curie-Weiss model, where the mean field magnetization is given by

$$m = \tanh\left(\beta J_0 m + \beta h\right) \quad , \tag{8.11}$$

where the couplings are all equal to $J_{ij} = \frac{J_0}{N}$. Differentiating with respect to h we get an expression for the correlation which we can then solve for J_0 and get

$$J_0 = \frac{C}{1+C} = C - C^2 + C^3 - C^4 + \ldots \tag{8.12}$$

where β was absorbed in the coupling J_0. The above equation is just the simpler version of eq.(8.10). By this analogy the authors were able to give the closed form of $J_{ij}^{(\text{loop})}$

$$J_{ij}^{(\text{loop})} = (L_i L_j)^{-1/2} \left[M \cdot (I + M)^{-1}\right]_{ij} \quad , \tag{8.13}$$

where M is defined by $M_{ij} = \beta K_{ij}\sqrt{L_i L_j}$ and $M_{ii} = 0$.

Interestingly, the authors also noticed that the 2-spins diagrams (*i.e.* the diagrams containing only powers of K_{ij}) can also be re-summed and the result is identical to the independent pair approximation of chapter 6

$$J_{ij}^{(\text{2-spins})} = \frac{1}{4}\ln\left[\frac{\left((1+m_i)(1+m_j)+C_{ij}\right)\left((1-m_i)(1-m_j)+C_{ij}\right)}{\left((1+m_i)(1-m_j)-C_{ij}\right)\left((1-m_i)(1+m_j)-C_{ij}\right)}\right] \quad . \tag{8.14}$$

The authors propose to use a combination of eq.(8.13,8.14) as a way to estimate the couplings from the observed magnetizations and correlations, ignoring the remaining terms in eq.(8.4). Their final formula for the couplings is

$$J_{ij} = J_{ij}^{(\text{2-spins})} + J_{ij}^{(\text{loop})} - \frac{K_{ij}}{1 - K_{ij}^2 L_i L_j} \quad . \tag{8.15}$$

The last term prevents double-counting of the terms present in both series for $J_{ij}^{(\text{2-spins})}$ and $J_{ij}^{(\text{loop})}$.

If the graph of the model in question was a tree, we know from chapter 6 that the correct result for the couplings is just the independent pair approximation $J_{ij}^{(\text{2-spins})}$ for edges present in the graph and zero for the others. This means that all other terms, corresponding to diagrams with loops, in eq.(8.4) should cancel out. Under this light, the result of eq.(8.15) is a corrected version of the independent pair approximation as we depart from trees and begin to have loops in our model.

One can also compute the local fields by applying eq.(8.5) to the entropy expansion of eq.(8.7) but no closed form has been found by the authors.

Chapter 9

Susceptibility Propagation

In chapter 6 we have exposed the advantages of working with tree models. Indeed, their decomposition in terms of purely local quantities allowed an exact and very efficient ($\mathcal{O}(N^2)$) algorithm to be found. It would be very useful to find a way to generalize these concepts in cases where the underlying graph is not a tree. The general form of a joint distribution of a tree model is, as we have seen in chapter 6,

$$P_t(\underline{x}) = \prod_{(ij)\in E_t} P_{ij}(x_i, x_j) \prod_{i\in V} P_i(x_i)^{1-|\partial i|} \quad . \tag{9.1}$$

The restriction of the graph being a tree is hidden in the edge set E_t. It is tempting to use the above factorized form on graphs that contain loops. It is known, nonetheless, that this form can be exact only for trees, so using it in loopy graphs is only approximative. It turns out that, under some conditions, it can be a good approximation for a certain class of graphical models. However, the simple MST algorithm 6.1.1 cannot be generalized to find loopy graphs and a completely new strategy must be devised. The assumption that a more general joint distribution can be approximated by a factorized form such as the above has led to the discovery, independently in various disciplines, of a class of inference algorithms known as *message passing*. Such algorithms, although exact only on trees, can be applied on loopy graphs as well and, under certain conditions, they turn out to be very good approximations.

We will first introduce the celebrated *Belief Propagation* (BP) algorithm, an inference algorithm suited for the direct problem, who operates in polynomial time. As we will show, this algorithm also minimizes the KL divergence of the true distribution with a trial distribution of the form (9.1). Then, using the fluctuation-response theorem, we will derive a set of equations involving the correlations and the couplings who can be used as an algorithm for the inverse problem, as we did for the mean field theories of chapter 7.

9.1 Belief Propagation

In the analysis that follows we will omit writing the normalization constants in equations involving distributions and use the symbol $\cong$ to denote equality up to a normalization. We will also need to introduce some notation. We will rewrite the distribution of the Ising model on a

tree graph $G = (V, E)$ as

$$P(\underline{s}) \cong \prod_{(ij)\in E} \psi_{ij}(s_i, s_j) \prod_{i\in V} \phi_i(s_i) \quad \text{with} \quad \psi_{ij}(s_i, s_j) = \mathrm{e}^{J_{ij}s_i s_j} \quad \text{and} \quad \phi_i(s_i) = \mathrm{e}^{H_i s_i} \ . \tag{9.1}$$

Belief propagation is based on the following idea: in tree graphical models there is always an efficient way of performing the sums needed to compute the marginal of some variable or set of variables. One has to start by summing the variables corresponding to the leaves of the tree, then proceed to the next level of variables and so on until he reaches the desired variables. That way, one always does sums of a single variable which is a computationally easier task. It turns out that the total computation time is proportional to the number of edges, *i.e.* it is of order $\mathcal{O}(N)$, instead of the exponential time required to compute marginals naively.

For example, let's say we have a 4-spins chain without local fields as the one depicted above. If we were to compute the marginal of s_4 we would do

$$P_4(s_4) \cong \sum_{s_3} \psi_{34}(s_3, s_4) \sum_{s_2} \psi_{23}(s_2, s_3) \sum_{s_1} \psi_{12}(s_1, s_2) \quad , \tag{9.2}$$

where it is understood that we begin by performing the rightmost sum and proceed towards the left. That way, we end up summing $3 \cdot 2$ terms, for each value of s_4, instead of 2^3.

Belief propagation is based on the view that, each time a variable is summed, a "message" is sent to the next variable containing information about the distribution of the latter based on the state of the former. More precisely, for a general tree-graph the messages are defined in the following way

$$\mu_{i\to j}(s_j) \cong \sum_{s_i} \phi_i(s_i)\psi_{ij}(s_i, s_j) \prod_{k\in\partial i\setminus j} \mu_{k\to i}(s_i) \quad , \tag{9.3}$$

where $\partial i \setminus j$ is the neighborhood of s_i except s_j. It can be checked that, if one starts form the leaves and iteratively computes all the messages towards the bulk, then the one and two variables marginals are given by

$$P_i(s_i) \cong \phi_i(s_i) \prod_{k\in\partial i} \mu_{k\to i}(s_i) \qquad \text{and} \tag{9.4}$$

$$P_{ij}(s_i, s_j) \cong \psi_{ij}(s_i, s_j)\phi_i(s_i)\phi_j(s_j) \prod_{k\in\partial i\setminus j} \mu_{k\to i}(s_i) \prod_{l\in\partial j\setminus i} \mu_{l\to j}(s_j) \quad . \tag{9.5}$$

Moreover, one benefits from the computational gain that was described in the 4-spins system example above.

The interesting feature of equations (9.3) is that they don't necessarily need to be restricted by the precise update schedule described above (from the leaves to the bulk). One can update every message at every time step, starting from any initial conditions, and then, after a number

of steps[i], compute all marginals from equations (9.4,9.5). In fact, this procedure doesn't even require the graph to be a tree. Messages can be defined on the edges of any graph and the equations (9.3) can be iterated until, hopefully, convergence is met. One can then compute a set of marginals that can be used as approximations of the true marginals.

Belief Propagation uses only local exchange of information, between variables, which concerns consistent marginalization of the variables with their neighbors, *i.e.* it guarantees that $P_i(s_i) = \sum_{s_j} P_{ij}(s_i, s_j)$ for any pair of neighbors. Locally consistent marginals, like these, don't necessarily correspond to a global distribution, hence they will be called *beliefs* hereon and will be denoted by the symbol $b(\,\cdot\,)$.

Since BP is exact only on trees, its effectiveness on loopy graphs relies on some kind of "resemblance" with tree-graphs. This idea can be described qualitatively in the following way: If, after the removal of a node of the graph, its neighbors become weakly correlated BP provides a good approximation for the marginals. A class of models where this property can be found is the class of sparse random graphs. In such graphs, because of the rarity of edges and the randomness of their position, small loops are rare. If, moreover, variables are weakly correlated in long distances, *e.g.* if, in the Ising model, the temperature is high enough, the above condition is fulfilled. It turns out that the breakdown of BP correctness occurs at the transition between the paramagnetic (replica symmetric) and spin-glass (broken replica symmetry) phases. The latter phase is characterized by the decomposition of the Boltzmann distribution in a great number of Gibbs states, an so BP tries to find marginals which are locally consistent with different such states but inconsistent in a global way.

But let's look at BP from our information theoretic point of view. As we have stressed numerous times already, a method whose aim is to compute approximate marginals (or equivalently moments, as in the mean field methods) tries to find the minimum KL divergence between some "easy" trial distribution and the true model distribution. In our case now, this trial distribution has the factorized form we found in trees, cf. distribution (9.1), except for the fact that, since BP only guarantees the local consistency of the inferred marginals, the true marginals are replaced by beliefs.

$$b(\underline{s}) = \prod_{(ij)\in E} b_{ij}(s_i, s_j) \prod_{i\in V} b_i(s_i)^{1-|\partial i|} \quad . \tag{9.6}$$

The minimization of the KL divergence between the above distribution and the true Ising distribution yields, as usual, a free energy minimization problem for the following distribution

$$\mathbb{F}[b] = \sum_{(ij)\in E} \sum_{s_i,s_j} b_{ij}(s_i, s_j) \log \frac{b_{ij}(s_i, s_j)}{\psi_{ij}(s_i, s_j)} + \sum_{i\in V} (1 - |\partial i|) \sum_{s_i} b_i(s_i) \log b_i(s_i) \tag{9.7}$$

It can be shown [MezardM 09] that the minima of the above free energy functional, under normalization and consistency constrains, $\sum_{s_i} b_i(s_i) = 1$ and $\sum_{s_j} b_{ij}(s_i, s_j) = b_i(s_i)$, correspond to the BP fixed points, *i.e.* to the sets of marginals given by equations (9.4,9.5) whose messages are such that the BP iteration given by eq. (9.3) wouldn't alter them. This doesn't say, however, that the BP iteration will converge in any case in one of those fixed points. As we said before, if we are in the paramagnetic phase the above free energy has one minimum and BP will converge to the correct marginals. But in the spin glass phase, convergence may not even occur.

i. In trees, the number of time steps needed is equal to the length of the longest path, *i.e.* the time needed so that the variables which are further way from each other exchange information.

9.2 Susceptibility Propagation

Belief Propagation, and its derivatives [ii], has been used with success in a number of contexts, like Bayesian inference [Pearl 88], decoding low density parity check codes [Gallager 62], turbo-codes [BerrouG 96] and satisfiability [BraunsteinMZ 05]. For the inverse Ising problem, Mora and Mézard [MoraM 09] have derived a message passing algorithm, based on BP, by means of the fluctuation-response theorem in the same line of thought we saw in the mean field theories. To outline the derivation of their algorithm we first need to introduce an alternative formulation of BP suited for models with binary variables, like spin systems. We define the *cavity fields* as

$$h_{i\to j} \equiv \frac{1}{2}\log\frac{\mu_{i\to j}(+1)}{\mu_{i\to j}(-1)} \quad . \tag{9.1}$$

At the BP fixed point these quantities are interpreted as the total effective field the spin i would feel if spin j would be removed from the graph. Using the cavity fields, the BP equations can be written as

$$h_{i\to j} = \sum_{k\in\partial i\setminus j} u_{k\to i} + H_i \quad \text{with} \tag{9.2}$$

$$u_{k\to i} = \operatorname{atanh}\left(\tanh J_{ki}\tanh h_{k\to i}\right) \quad . \tag{9.3}$$

For the messages $h_{i\to j}$ and $u_{k\to i}$ we define their derivatives

$$g_{i\to j,l} \equiv \frac{\partial h_{i\to j}}{\partial H_l} \quad \text{and} \quad v_{k\to i,l} \equiv \frac{\partial u_{k\to i}}{\partial H_l} \quad . \tag{9.4}$$

Then, from the BP equations (9.2,9.3), we have that

$$g_{i\to j,l} = \sum_{k\in\partial i\setminus j} v_{k\to i,l} + \delta_{il} \quad \text{with} \tag{9.5}$$

$$v_{k\to i,l} = g_{k\to i,l}\tanh J_{ik}\frac{1-\tanh^2 h_{k\to i}}{1-\tanh^2 u_{k\to i}} \quad . \tag{9.6}$$

From eq. (9.4), together with the definition of the cavity fields in eq. (9.1), we have the following expression for the magnetizations

$$m_i = \tanh\left(H_i + \sum_{k\in\partial i} u_{k\to i}\right) \tag{9.7}$$

form where we can derive an expression for the correlations, using once more the fluctuation-response theorem and the expressions for the messages derivatives in eqns. (9.5,9.6)

$$C_{ij} = g_{j\to i,j}\bar{C}_{ij} + g_{i\to j,j}(1-m_i^2) \quad \text{with} \tag{9.8}$$

$$\bar{C}_{ij} \equiv \frac{\tanh J_{ij} + \tanh h_{i\to j}\tanh h_{j\to i}}{\tanh J_{ij}\tanh h_{i\to j}\tanh h_{j\to i}} - m_i m_j \quad . \tag{9.9}$$

ii. BP solves marginalization problems. In different contexts, similar and equally intractable problems may arise, like finding the configuration that maximizes a distribution. Similar, message-passing procedures can be derived from the BP algorithm, suited for these problems, like the Max-Product and Min-Sum algorithms [MezardM 09].

The above formula can easily be inverted to yield a relation for the couplings, to be used in the inverse problem

$$\tanh J_{ij} = \frac{\tilde{C}_{ij} - \tanh h_{i\to j} \tanh h_{j\to i}}{1 - \tilde{C}_{ij} \tanh h_{i\to j} \tanh h_{j\to i}} \tag{9.10}$$

where we have used the disconnected part of $\bar{C}_{ij} = \tilde{C}_{ij} - m_i m_j$ given by eqn (9.8)

$$\tilde{C}_{ij} = \frac{C_{ij} - g_{i\to j,j}(1 - m_i^2)}{g_{j\to i,j}} + m_i m_j \quad . \tag{9.11}$$

Once the fixed point values of the cavity fields and cavity susceptibilities $h_{i\to j}$ and $g_{i\to j,k}$ are found the couplings can be computed using eqn (9.10). The authors propose the following iteration to find those fixed point values

Algorithm 9.2.1: SUSCEPTIBILITY PROPAGATION(m, C)

Initialize all u's to random values
Initialize all h's, v's and g's to zero
while $\delta h \neq 0$ **and** $\delta g \neq 0$
do
- **for** $(ij) \in V^2$ **do** $h_{i\to j} \leftarrow \text{atanh}(m_i) - u_{j\to i}$
- **for** $(ijk) \in V^3$ **do** update $g_{i\to j,k}$ using eqn (9.5)
- **for** $(ij) \in V^2$ **do** update J_{ij} using eqn (9.10)
- **for** $(ij) \in V^2$ **do** update $u_{i\to j}$ using eqn (9.3)
- **for** $(ijk) \in V^3$ **do** update $v_{i\to j,k}$ using eqn (9.6)

for $i \in V$
do $H_i \leftarrow \text{atanh}(m_i) - \sum_{j\in\partial i} u_{j\to i}$
return (J, H)

The above algorithm has been studied in the context of the inverse Ising problem [AurellOR 10, MarinariK 10, Huang 10b] and the general conclusion is that although it provides accurate results in high temperature settings, it suffers from convergence problems as the temperature is lowered. We will see, however, in the next section that these problems can be circumvented.

9.3 Bethe Approximation Method

Although we chose to introduce the subject from an algorithmic point of view, the BP algorithm is deeply connected with the well known *Bethe Approximation* of statistical physics. This approximation has been used to solve the Ising problem and uses the following idea: one

can replace the original model with a tree model that has the same local structure *i.e.* each spin has the same number of neighbors. This amounts in replacing the true free energy of the system with one that has the form of eq.(9.7) called *Bethe Free Energy*. Here a connexion with the naive mean field (NMF) and TAP theories can be made. We showed how these methods are obtained by the two first of the expansion of the free energy in small couplings at fixed magnetizations. Continuing the expansion naturally leads to loop terms, like $J_{ij}J_{jk}J_{ki}$, and higher powers of single couplings J_{ij}^n. It has been showed [GeorgesY 91] that these latter terms can be resummed and lead to the Bethe Approximation (BA). Thus the Susceptibility Propagation (SP) algorithm, clearly corresponding to the BA, can be seen as a further improving NMF and TAP by taking into account also the higher powers of single couplings in the expansion of the true but intractable free energy.

Concerning the inverse Ising problem, it has recently been shown [Ricci-Tersenghi 12] that there exist an analytical expression for the fixed point of the SP algorithm. These allows one to use the Bethe Approximation without having to cope with the serious convergence problems of SP. Here we will simply give the expression used to infer the couplings and refer the reader to the original paper for further details

$$\begin{aligned} J_{ij}^{\mathrm{BA}} &= -\mathrm{atanh}\left[\frac{1}{2(C^{-1})_{ij}}\sqrt{1+4(1-m_i^2)(1-m_j^2)(C^{-1})_{ij}^2} - m_i m_j - \right. \\ &\left. \frac{1}{2(C^{-1})_{ij}}\sqrt{(\sqrt{1+4(1-m_i^2)(1-m_j^2)(C^{-1})_{ij}^2} - 2m_i m_j (C^{-1})_{ij})^2 - 4(C^{-1})_{ij}^2}\right] . \end{aligned} \tag{9.1}$$

Latter, in chapter 12 it will be seen that the above formula is able to give results even for low temperatures, below the critical temperature of the systems in question, a region where SP clearly fails.

Chapter 10

Adaptive Cluster Expansion

Belief Propagation is based on the assumption that the probability distribution of the system in question can be approximated by a factorized distribution of the form of distribution (9.1), which is exact on trees. As we have seen in chapters 6 and 9, this form allows extensive quantities, such as the entropy, to be computed efficiently as they are decomposed in a sum of local terms. To illustrate this point we rewrite the entropy corresponding to the distribution (9.1)

$$S = -\sum_{(ij)\in E}\sum_{s_i,s_j} P_{ij}(s_i,s_j)\log P_{ij}(s_i,s_j) - \sum_{i\in V}(1-|\partial i|)\sum_{s_i} P_i(s_i)\log P_i(s_i) \quad . \tag{10.1}$$

One way of looking at the above relation is that we have approximated the true entropy by a sum of terms which are themselves correct expressions of the entropies of subsets of spins. These subsets, called *clusters*, are, in that case, of sizes one and two. They are combined in that particular way in order to avoid multiple counting of the single spin contributions: if we simply sum all the two spins entropies we are taking into account $|\partial i|$ times the contribution of the i^{th} spin [YedidiaFW 03].

This idea can be extended by using larger and larger clusters, where the entropy is computed exactly, and combine them in order to get better approximations of the entropy (or the free energy). This leads to the so called Kikuchi approximation [Kikuchi 51, YedidiaFW 03]. This approximation is able to provide better results than the Bethe approximation in cases where the graph deviates from the locally tree-like structure of sparse random graphs and may contain some small loops. It does so by using the exact form of the entropy or free-energy for clusters large enough to contain these small loops. Of course, a drawback of the method is its higher computational complexity, since the computation of those cluster entropies is exponential in the cluster size.

In many natural contexts, systems have structures which, despite being sparse, they are not typical examples of sparse random graph ensembles. They may contain regions of high connectivity, where lots of small loops are present, and regions of relative sparseness. In such cases a scheme using small clusters wouldn't be effective, as high connectivity regions would generate big errors, while a scheme using large clusters, able to account for the highly connected parts, would be inefficient due to the high computational complexity.

In order to optimize the performance of such an algorithm, while not missing the information contained in highly connected parts, S. Cocco and R. Monasson [CoccoM 11, CoccoM 12] have

devised an adaptive cluster expansion for the inverse Ising problem, where the algorithm adapts to the data and cleverly chooses different cluster sizes, to be used in the inference, for different parts of the system. The idea is that one starts form the set of all clusters of size one, and then selectively combines them to form larger clusters using a criterion that shows if the information gain is worth the extra computational effort.

Since they focus their work on the inverse Ising model, the starting point is the Legendre transform of the free-energy at fixed magnetizations and correlations, the same as the one used in chapter 8

$$S(m,C) = \min_{J,H} \left\{ \log Z(J,H) - \sum_{i<j} J_{ij}(C_{ij} + m_i m_j) - \sum_i H_i m_i \right\} \quad . \tag{10.2}$$

For any subset $\Gamma \subset V$ with $|\Gamma| = K$ one can compute the *subset entropy* $S_\Gamma(m,C)$ by restricting the variables in above formula accordingly. The complexity is dominated by the computation of the partition function $Z(J,H)$ and is therefore of order 2^K. A second notion is that of the *cluster entropy* $\Delta S_\Gamma(m,C)$, which is the remaining contribution to the subset entropy, once all other cluster entropies of smaller clusters have been subtracted. The two quantities are related through the identity

$$S_\Gamma(m,C) = \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'}(m,C) \quad . \tag{10.3}$$

Note that the sum runs over $2^K - 1$ clusters of sizes $|\Gamma'| < K$.

Using the Möbius inversion formula one can show that the cluster entropies can be recursively calculated from all subset entropies $S_{\Gamma'}$ with $\Gamma' \subset \Gamma$

$$\Delta S_\Gamma(m,C) = \sum_{\Gamma' \subset \Gamma} (-1)^{K'-K} S_{\Gamma'}(m,C) \tag{10.4}$$

which is a generalization of the idea, found also in the Bethe entropy form, cf. eq.(10.1), that in order to avoid multiple counting of smaller cluster contributions, one has to subtract from each subset entropy of size K the corresponding subset entropies of sizes $K-1$.

The authors maintain that, in many cases, a good approximation to the entropy can be achieved using only a well-chosen set of small clusters L,

$$S(m,C) \approx S_0(m,C) + \sum_{\Gamma \in L} \Delta S_\Gamma(m,C) \quad , \tag{10.5}$$

where the expansion is actually carried around a reference entropy which is given by mean field theory, $S_0(m,C) = \frac{1}{2} \log\det M$ with $M_{ij} \equiv \frac{C_{ij}}{\sqrt{m_i(1-m_i)m_j(1-m_j)}}$.

One obvious way to chose the set L is by not taking into account all cluster entropies below some threshold θ. In that way, however, one has to compute the $2^N - 1$ cluster entropies and then discard all those for which $|\Delta S_\Gamma(m,C)| < \theta$. This is clearly not efficient, so the authors propose a recursive method reminiscent of evolution. Starting from the set of all one spin clusters, a *combination* process constructs new clusters by combining old ones and a *selection* process eliminates the non relevant clusters, keeping only those for which $|\Delta S_\Gamma(m,C)| < \theta$. This criterion also makes information theoretic sense since, as the authors show, the cluster entropy

of a two spin system is equal to the Kullback-Leibler divergence between the joint probability of the two spins and the probability of two independent spins with magnetizations equal to the true ones. In other words, it is the mutual information of the two spins. The argument can be recursively iterated for larger clusters and therefore the cluster entropies represent, in some sense, the information gain of taking into account those larger clusters, instead of considering their sub-clusters to be independent. It is asserted that when two clusters, sharing some spins, are linked by additional paths in the true graphical model, then the cluster entropy of their combination will be important, as opposed to the case where all the information about actual edges can be already found in the separate clusters. This leads to the following algorithm for finding the set L

Algorithm 10.0.1: ADAPTIVE CLUSTER EXPANSION(m, C, θ)

Initialization: build the set of all clusters of size one $L_1 = \{1, \ldots, N\}$
while at least one cluster is selected
do {
  for $\Gamma_i, \Gamma_j \in L_K$ such that $|\Gamma_i \cup \Gamma_j| = K + 1$
  do {
    $\Gamma \leftarrow \Gamma_i \cup \Gamma_j$
    if $|\Delta S_\Gamma(m, C)| \geq \theta$
    then $L_{K+1} \leftarrow \{L_{K+1}, \Gamma\}$
  }
  $K \leftarrow K + 1$
}
return $(L = \bigcup_{l=1}^{K_{\max}} L_l)$

The selection step requires the computation of the cluster entropy from the cluster entropies of the previous iteration according to the formula

$$\Delta S_\Gamma(m, C) = S_\Gamma(m, C) - (S_0)_\Gamma(m, C) - \sum_{\substack{\Gamma' \subset \Gamma \\ \Gamma' \neq \Gamma}} \Delta S_{\Gamma'}(m, C) \tag{10.6}$$

Finally, once L has been found, the entropy can be computed from eq (10.5). The couplings and fields can then be found using eqns (8.4,8.5) [i].

i. The authors propose a more clever method which doesn't require symbolic differentiation of the above expression. We won't get into the details and refer the reader to the original paper for a full discussion [CoccoM 12].

Chapter 11

Inference in the $p < N$ regime

In chapters 7, 8 and 9 we presented a series of methods based on standard mean field theory. As we have mentioned there, one of the necessary conditions for those methods to yield correct results is that the graphical model, on which they will be applied, has to be fully connected. In such cases, and under some additional conditions related to the strength of the interactions, those methods are able to correctly predict the model parameters asymptotically for $p \to \infty$. Moreover, all the aforementioned methods rely explicitly on the inversion of the correlation matrix C so they are constrained to cases where the correlation matrix is invertible. This condition is fulfilled if the number of measurements p is at least equal to the number of spins N so that the rows and columns of C are linearly independent. If $p < N$ those methods will potentially lead to infinite couplings if no regularization is used. In fact other methods also, like Susceptibility Propagation or Adaptive Cluster Expansion, although not implicitly, rely on the invertibility of C since they actually solve a system of equations whose constant factors (determined by C and m) have to be linearly independent. So SP will also lead to infinite couplings if $p < N$.

On the other hand, the whole question of inferring in the $p < N$ is irrelevant in the fully connected case since the couplings are of order $\mathcal{O}(1/\sqrt{N})$ and estimation with p measurements yields errors of order $\mathcal{O}(1/\sqrt{p})$. Nonetheless, the situation is different when the graphical model in question is not fully connected. Intuitively, inferring a sparse system has to need less information than inferring a fully connected one. So if a system is sparse enough it could in principle be inferred using $p < N$ samples. However, the methods described previously would still be worthless, because of the implicit or explicit matrix inversion.

A standard approach to avoid infinities is the use of some regularizer, *i.e.* some sort of potential that forces the couplings to stay in the vicinity of zero. In this chapter we will review a particular class of regularizers, used extensively in the literature, the *ℓ_p-norm regularizers*. We will discuss how they can be used in the context of an inverse Ising algorithm through the example of a method introduced by P. Ravikumar, M.J. Wainwright and J.D. Lafferty, the *ℓ_1-regularized logistic regression* [RavikumarWL 10], which uses the particular value of $p = 1$ [i] for the ℓ_p-norm, a value that presents a number of benefits and is thus a standard choice. The introduction of ℓ_p-norm regularizers is not constrained in that particular method and can be

i. Unfortunately the standard choice for the norm parameter is p, the same symbol as the one we use for the number of samples. The probability of confusion is low however, since they never coexist in the same expression.

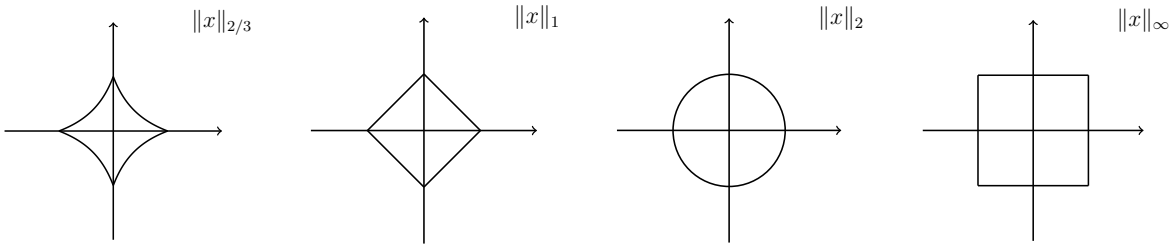


Figure 11.1: Unit circles using ℓ_p-norms for various values of p. From left to right $p = 2/3$, $p = 1$, $p = 2$, $p \to \infty$. Note that the circles are convex only for $p \geq 1$ and they are not smooth at $x_i = 0$ only for $p \leq 1$.

used in other contexts as well, *e.g.* S. Coco and R. Monasson discuss their use in their Adaptive Cluster Expansion algorithm in [CoccoM 12]. In part III, where we will present our new method for exactly inferring systems with asymmetric interactions, we will also show how this idea can be used to obtain an algorithm better suited for sparse systems in the regime $p < N$.

11.0.1 ℓ_p-norm regularization

The usual Euclidean norm of a vector $x = (x_1, x_2, \ldots, x_N)$ is an ℓ_p-norm with $p = 2$. The generalization to any p is defined as

$$\|x\|_p \equiv \sqrt[p]{|x_1|^p + |x_2|^p \cdots + |x_N|^p} \quad . \tag{11.1}$$

Some other interesting values of p are $p = 0$ where the norm equals the number on non-zero elements, $p = 1$ where it equals the sum of their absolute values and $p = \infty$ where it equals the value of the maximum element [ii]. In fig.(11.1) we see a series of unit circles in spaces where the ℓ_p-norm defines the metric, for some values of p.

In the context of machine learning or inverse problems, an important property of any regularizer is that it should be convex so that the corresponding optimization problem remains convex. Circles are convex only in metrics defined by ℓ_p-norms with $p \geq 1$, cf fig. (11.1). An other important feature of inference algorithms in such contexts is that they should set a fraction of the inferred parameters exactly to zero. This is, anyway, a very important aspect of every scientific explanation: data should be interpreted with the smallest set of arbitrary parameters. Moreover, in the inverse Ising problem it also has practical benefits since the goal is to create algorithms able to predict the actual structure of physical networks, where a null coupling is interpreted as a missing link. ℓ_p-norms which are able to set a fraction of the parameters exactly to zero are those which are not smooth at $x_i = 0$, *i.e.* those with $p \leq 1$. The only value of p having both features is $p = 1$ obviously. The idea of using the ℓ_1-norm in optimization was introduced by R. Tibshirani [Tibshirani 96]. Let us look at his example to see why the ℓ_1-norm is able to set a fraction of the parameters to zero.

Let's say we have the following optimization problem

$$(\eta^*, \theta^*) = \arg\min \left\{ \sum_{\mu}^{p} \left(y_\mu - \eta - \sum_{i=1}^{N} \theta_i x_{\mu i} \right)^2 \right\} \quad \text{subject to } \sum_{i=1}^{N} |\theta_i| \leq \lambda \tag{11.2}$$

ii. The $p = 0$ and $p = \infty$ are defined as limiting cases and hence the term *pseudo-norm* is often used.

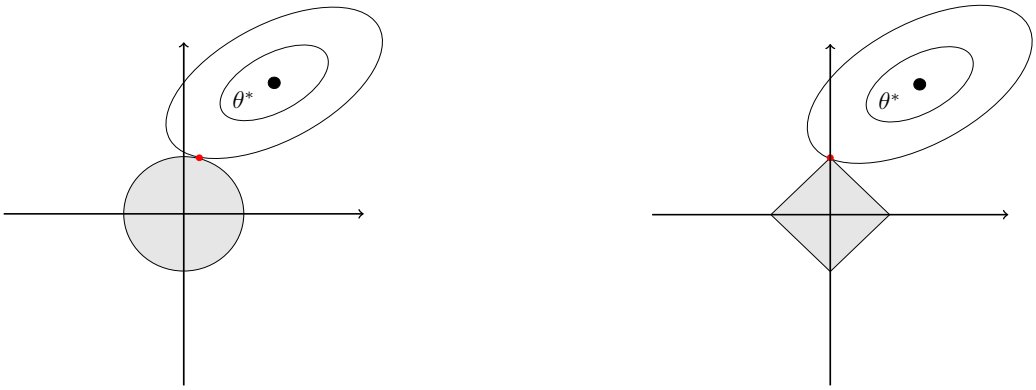


Figure 11.2: A quadratic form, with minimum at θ^*, is minimized subject to $\|\theta\|_2 \leq \lambda$ (left) and $\|\theta\|_1 \leq \lambda$ (right). The red dots indicate the solutions, *i.e.* the first place where the contour lines hit the circles.

for some data $(\mathbf{x}^{(\mu)}, y_\mu)$, $\mu = 1, \ldots, p$. The constraint is nothing but $\|\theta\|_1 \leq \lambda$, where λ is a tuning parameter which controls the number of parameters θ_i to be set to zero. The above problem amounts to minimize the quadratic form under the constrain. As can be seen in fig. (11.2), if the ℓ_2-norm is used, the contour will hit the circle at a zero almost never, whereas if the ℓ_1-norm is used this will sometimes occur at a corner.

So the ℓ_1 regularizer has the ability to infer sparse results. This will prove very useful for our discussion since we are interested in the case where inference is done with $p < N$, something that can be done in principle only when the graphical model is sparse. We will present a natural way of incorporating the ℓ_1 regularizer in our inverse Ising problem by turning to the Bayesian formulation of the problem. As we have seen in section 3.2, in stead of minimizing a Kullback-Leibler divergence, an equivalent way of formulating an inference problem is by means of Bayes theorem

$$P(J, H|\underline{s}) = \frac{P(\underline{s}|J, H)P(J, H)}{P(\underline{s})} \quad , \tag{11.3}$$

formulated here for the inverse Ising problem. In order to penalize dense models, the prior distribution can be set to decay exponentially with the number of non-zero elements of the couplings matrix $P(J, H) \cong \exp(-\lambda \|J\|_0)$. The problem with that is the non-convexity of the ℓ_0-norm, hence the ℓ_1-norm is preferred since it favors sparse models while being convex. The likelihood of the couplings and fields, using P measurements, now reads

$$P(J, H) \cong \left[\mathrm{e}^{\sum_{i<j} J_{ij}(C_{ij}+m_i m_j)+\sum_i H_i m_i - \log Z(J,H)} \, \mathrm{e}^{-\lambda \sum_{i<j} |J_{ij}|} \right]^p \tag{11.4}$$

Taking the negative logarithm we get the following log-likelihood function, rescaled by a factor $1/p$

$$\mathcal{L}(J, H) = \log Z(J, H) - \sum_{i<j} J_{ij} \widetilde{C}_{ij} - \sum_i H_i m_i + \lambda \sum_{i<j} |J_{ij}| \quad , \tag{11.5}$$

where we used the non-connected correlation $\widetilde{C}_{ij} \equiv \langle s_i s_j \rangle = C_{ij} + m_i m_j$. Note that the above function, without the regularizer, is identical to the entropy in eq. (8.1) that we encounter in chapters 8 and 10. The minimum of $\mathcal{L}$ yields a set of model parameters $(J^*, H^*) =$

arg min$\{\mathcal{L}(J,H)\}$ which are able to reproduce the correct correlations and magnetizations while being sparse. The regularizer, in this case, doesn't impose a hard constrain as in the quadratic form example, cf. eqn (11.2),but a probabilistic one. The parameter λ functions as a kind of chemical potential that controls the presence of non-zero couplings.

11.1 ℓ_1-regularized Logistic Regression

Minimizing eqn (11.5) leads to a Boltzmann machine learning algorithm, as in eqns (14.7,14.6) modified by the presence of the regularizer, thus it suffers from the usual intractability of the partition function Z. In stead, P. Ravikumar, M.J. Wainwright and J.D. Lafferty proposed in [RavikumarWL 10] a way to obtain a tractable algorithm by treating independently the neighborhood of each spin. By doing that they map the problem to a set of N independent problems which, being defined on star graphs, have partition functions which are easy to compute. The trade off is that their algorithm fails to provide correct results in the low temperature phase where the correlations are long-range. Nonetheless, it can be applied with success to sparse models with weak interactions (high temperature). More importantly, the presence of the regularizer cures the infinities caused by the high level of noise in the data and succeeds in providing correct results even in the $p < N$ regime.

In order to treat each neighborhood independently one has to infer the couplings, adjacent to each spin, given the combined knowledge of their state and the state of their neighbors, found in the data. For that, one has to start from the conditional probabilities for a spin given the remaining ones

$$P(s_i|\underline{s}_{\setminus i}) = \frac{\exp\left(s_i(H_i + \sum_{j\in V\setminus i} J_{ij}s_j)\right)}{2\cosh\left(H_i + \sum_{j\in V\setminus i} J_{ij}s_j\right)} \quad , \tag{11.1}$$

which leads to the the following set of negative log-likelihoods, one for each spin

$$\mathcal{L}^{(i)}(J_{\setminus i}, H_i) = \frac{1}{p}\sum_{\mu=1}^{p} f(J_{\setminus i}, \underline{s}_{\setminus i}^{(\mu)}) - H_i m_i - \sum_{j\in V\setminus i} J_{ij}\widetilde{C}_{ij} + \lambda\|J_{\setminus i}\|_1 \quad , \tag{11.2}$$

with

$$f(J_{\setminus i}, \underline{s}_{\setminus i}^{(\mu)}) \equiv \log 2\cosh\left(H_i + \sum_{j\in V\setminus i} J_{ij}x_j^{(\mu)}\right) \quad , \tag{11.3}$$

where $J_{\setminus i}$ is a shorthand for the vector $\{J_{ij} : j \in V \setminus i\}$.

Then N^2 independent couplings and N fields are found by solving the N convex minimization problems

$$(J^*_{\setminus i}, H^*_i) = \arg\min\left\{\mathcal{L}^{(i)}(J_{\setminus i}, H_i)\right\} \quad . \tag{11.4}$$

The actual couplings should then be inferred by taking into account the, *a priori* different, values found for each two symmetric entries of J, for example by taking their average.

Let us examine the computational complexity of the above program. There are N minimization problems in each of which N parameters have to be inferred. However, the computation of the log-likelihood (or its derivatives) requires summing over the p samples and summing over the $N-1$ remaining parameters. Thus the computational complexity of the algorithm is $\mathcal{O}(pN^3)$. In the mean field algorithms for fully connected graphical models of the previous sections the sample complexity was determined by the fact that the correlation matrix had to be invertible, thus p was at least equal to N in which case the above algorithm would be at least of order $\mathcal{O}(N^4)$. The authors of [RavikumarWL 10] have shown that, when the model is sparse, under some additional conditions concerning the couplings strength, correct inference can be done with a sample complexity of the order of $\mathcal{O}(d^3 \log N)$, where d is the maximum degree of the graph. The additional conditions for the success of the algorithm concern two things. First, the values of the non-zero couplings have to be bounded from below in absolute value, *i.e.* $\min_{(ij)\in E} |J_{ij}| \equiv J_{ij}^{(\min)} \neq 0$. If one lets the couplings to have values arbitrarily close to zero then one must have $p \to \infty$ in order to correctly distinguish them from the zero couplings since the inference errors are of the order of $\mathcal{O}(1/\sqrt{p})$. This is not proper to the particular algorithm, it is a general condition for any algorithm that wants to infer the structure of a sparse graphical model by distinguishing zero from non-zero couplings. The second condition is related with the region of validity of the basic assumption on which the method relies: the fact that the neighborhoods can be inferred independently. J. Bento and A. Montanari have shown in [BentoM 09] that there is a threshold in the temperature at which the data was generated beyond which the algorithm fails to reconstruct the correct graph with high probability, even for a great number of samples. This threshold is apparently related, but does not coincide, with the critical temperature of the model. It is natural to expect that, in the low temperature phase, the long-range correlations forbid the independent treatment of each spin neighborhood.

Chapter 12

Comparative simulations

In this section we will present the results of a number of simulations in order to compare the performances of the algorithms described earlier on the inverse Ising problem. All the examples take the following form: an Ising system is generated at random, the system's evolution is simulated using a Monte Carlo dynamics for a number of steps, the inverse Ising methods are used to infer the couplings from the configurations generated by the dynamics and then the error made on the couplings is computed and plotted against parameters like the inverse temperature at which the samples were generated (β), the number of samples p and others. Various definitions of the error have been used depending on the situation. The details will be fully explained for every case.

We didn't include two methods in the presentation for a number of reasons. The *Boltzmann Machine* (see chapter 5) and the *Adaptive Cluster Expansion* (see chapter 10). All other methods have a polynomial computational complexity[i] and it is therefore natural to compare them. On the other hand the Boltzmann Machine has an exponential complexity and it is well known that, given enough time, it will yield the correct results within the limits imposed by finite sampling, of course. The Adaptive Cluster Expansion has a computational complexity which is not well defined. Its running time depends on the choice of the parameter θ (see 10.0.1) and so is the inference quality. One can choose a very small θ in which case the algorithm tends to be exact and of exponential complexity. It is therefore not very relevant to compare them with the remaining algorithms which are designed to have a clear computational advantage with the drawback of an often higher error.

12.1 Mean Field methods on fully connected systems.

First we examine fully connected systems, the so called Sherrington-Kirkpatrick (SK) model (see chapter 2.2). As we have already mentioned, the couplings J_{ij} are Gaussian variables with mean 0 and variance $1/N$. For the local fields H_i we included two cases: one where all fields are zero (upper row of fig.12.1) and one where they are chosen at random uniformly between in $[-0.5, 0.5]$ (lower row of fig.12.1). In this first example we are interested in the quality of

i. Most of them are of the mean-field type which is the reason of their low complexity.

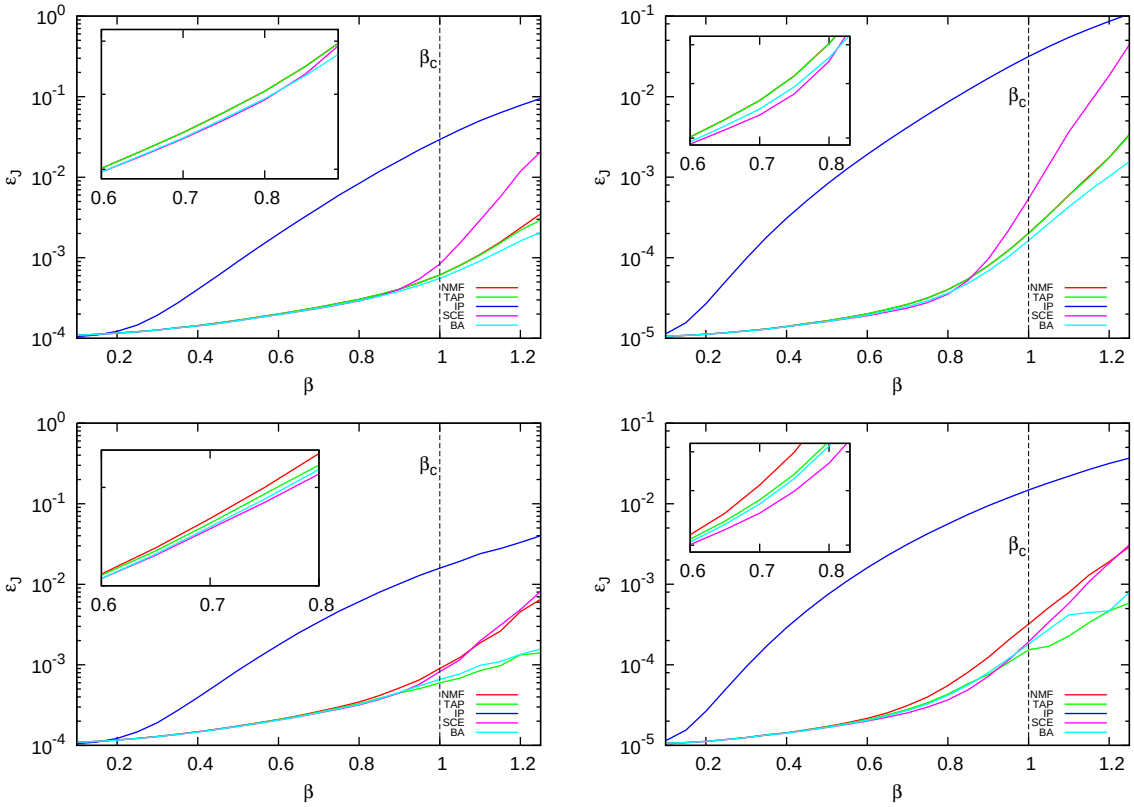


Figure 12.1: Couplings inference error (eq.(12.1)) versus beta for a $N = 100$ system for $p = 10^4$ (left) and $p = 10^5$ (right). The local fields are zero in the upper row and uniformly random in $[-0.5, 0.5]$ in the lower row.

the couplings inference hence we have used a simple mean squared error

$$\epsilon_J = \overline{(\beta J_{ij}^{\text{true}} - \beta J_{ij}^{\text{inferred}})^2} = \frac{2}{N(N-1)} \sum_{i<j} (\beta J_{ij}^{\text{true}} - \beta J_{ij}^{\text{inferred}})^2 \quad . \tag{12.1}$$

The first procedure is to simulate the system using Monte Carlo for a number of steps. Then the data are used to compute the magnetizations and correlations from where the couplings are inferred using the methods described in the previous chapters. The Monte Carlo procedure is done in the following way: starting from a random initial spin configuration, the system is simulated for $2Np$ steps in total. In each step, one spin is picked at random and is updated in place using the standard Metropolis-Hastings criterion. For every $2N$ steps, one configuration is memorized and only those resulting p configurations are used in the end.

For two values of $p = 10^4$ and 10^5 we plot the error of the above equation versus the inverse temperature β at which the simulation was conducted. The experiment is then repeated 50 times with different realizations of the couplings matrix and the local fields and the results are averaged. At this point it is important to stress that we didn't take into consideration the different thermalization times needed in different temperatures. In order to have a curve for the error as a function of β we kept the number of samples fixed even if in lower temperatures (especially beyond β_c) one would probably need a greater number of samples to correctly represent the distribution.

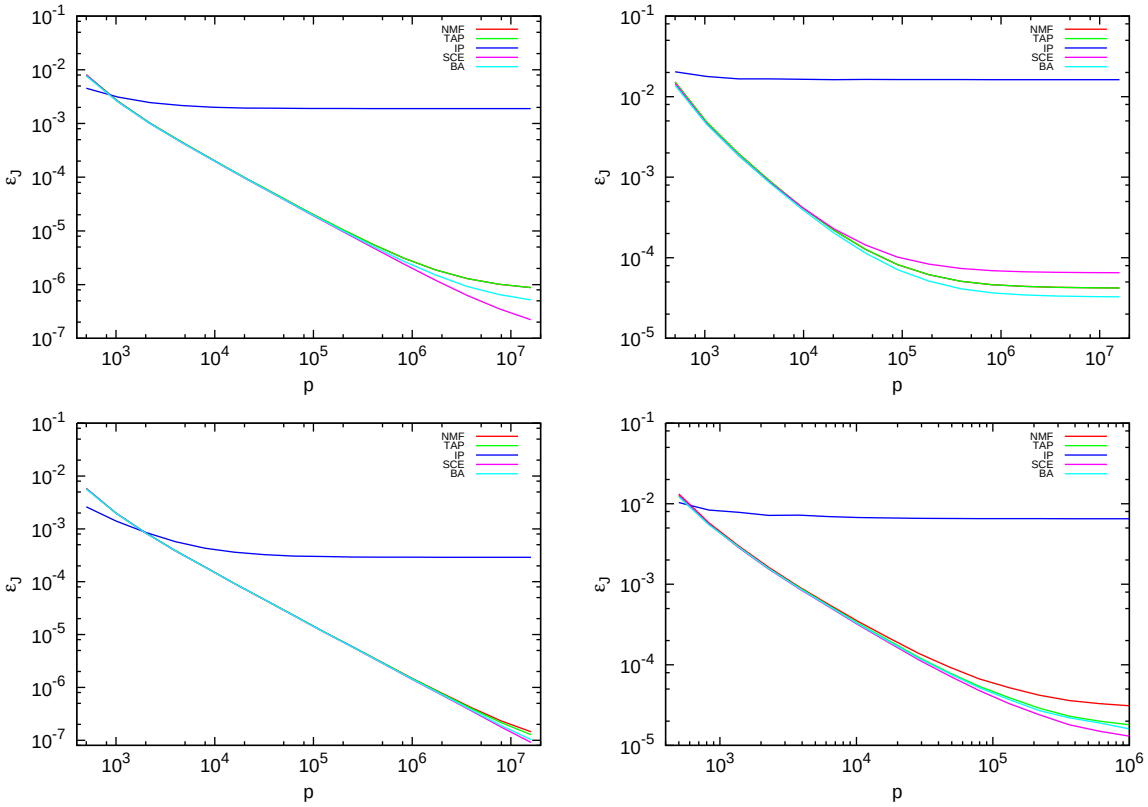


Figure 12.2: Upper row: Couplings inference error (eq.(12.1)) versus p for a $N = 100$ system with zero local fields, for $\beta = 0.6$ (left) and $\beta = 0.9$ (right). Lower row: The same but with non-zero local fields, uniformly drawn in $[-0.5, 0.5]$, for $\beta = 0.4$ (left) and $\beta = 0.8$ (right).

The different curves seen in fig.12.1 represent the results for the Naive Mean-Field method (red) found in chapter 7.1, the TAP method (green) of section 7.2, the Independent Pair method (blue) of section 6.2, the Small Correlation Expansion (magenta) of chapter 8 and the Bethe Approximation (cyan) of section 9.3.

The first remark concerns the error at low β. It can be clearly seen in fig.12.1 that for high temperatures (low β) all methods are equivalent and produce an error very close to $1/p$. Indeed it can be shown [AmariKN 92] that the error of the exact Boltzmann Machine algorithm becomes $1/p$ for high temperatures and since all the above methods are correct in the limit $\beta \to 0$ this coincidence is expected. This error is natural since when we estimate the correlations and magnetizations from a finite number of samples then we commit errors of order $\mathcal{O}(1/\sqrt{p})$.

In higher values of β the IP method is clearly by far the worst. This is not surprising since it is also by far the simplest [ii]. The remaining methods are all very close to the value $1/p$ with SCE and BA being the best followed by TAP and then NMF. Especially in the case of zero external fields all methods are almost indistinguishable up to $\beta \approx 0.8$. Then however the curves begin to deviate one from the other. As we enter the spin-glass phase the error becomes much more important. The BA method seems to perform quite well for low temperatures while the SCE who had the lowest error in higher temperatures becomes a really bad choice.

ii. It has a computational complexity of $\mathcal{O}(N^2)$ while all other methods are $\mathcal{O}(N^3)$.

Let us now see how the error behaves as a function of the number of samples (fig.12.2). Again IP performs much worse than the rest in general, although it can outperform the other methods if the number of samples is really low. This is a general principle: the simpler a method is, the smaller the errors it produces when the data is very noisy. The reason is that in more complex methods the larger number of operations involved magnifies the already important errors in the estimation of the correlations due to the small number of samples.

Another important remark is that the rule $\epsilon_J \sim 1/p$ doesn't apply to large values of p. As was shown in the analysis of the NMF and TAP methods in [RoudiH 11a] [iii] the error takes the form $\epsilon_J = \epsilon + \epsilon^\infty$ where $\epsilon \sim 1/p$ and $\epsilon^\infty \sim 1/N$. As we can see in the figure these finite size effects are more important for higher values of β. Moreover, we note that the asymptotic term ϵ^∞ has different constants for each method and that the rank of the methods from better to worse is compatible with what we see in fig.12.1. Also note that the asymptotic term is more important in low temperatures.

12.2 Mean field methods on sparse systems.

Let's see now what happens when the model that we want to infer is sparse, *i.e.* a lot of its couplings are zero. The model used here for the underlying graphs is the Erdős-Rényi model where each of the $N(N-1)/2$ possible links appears in the graph with an independent probability such that the average degree is d. In the upper row we have again the error given by eq.(12.1) versus β and the picture is the same as the one we had in the fully connected case although the difference between the methods is more pronounced with NMF and TAP being clearly the worse choices, within the mean field methods, for sparse graphs. We known that NMF and TAP are good approximations in infinite dimensions or infinite range systems so this is not surprising. On the other hand we see that the Bethe Approximation is the better choice although SCE slightly outperforms it for a small range of β. This is expected since sparse Erdős-Rényi graphs are locally tree-like and the BA is well suited in such cases. The same remarks apply to the lower left frame where the error is plotted against the number of samples.

In sparse systems we have an additional parameter to vary, the average degree d. It is interesting to see the behavior of the inference error as a function of d also. However, we judged that the error given by eq.(12.1) is not a good measure for this case, as it would increase with d anyway because of the greater number of non-zero couplings. Instead we used a normalized version

$$\epsilon_J = \sqrt{\frac{\sum_{i<j}(\beta J_{ij}^{\text{true}} - \beta J_{ij}^{\text{inferred}})^2}{\sum_{i<j}(\beta J_{ij}^{\text{true}})^2}} \quad . \tag{12.1}$$

In the lower right frame of fig.12.3 we see the inference error, computed with the above formula, of systems of size $N = 100$ with d ranging from 1 to 20. It is interesting to note that the mean field methods get better with higher d while the IP, which is exact when applied to the already known edges of a tree, gets worse. We even see that for d close to 1 [iv] it outperforms NMF

iii. The analysis concern the non-equilibrium versions of the algorithms but the same results apply to the equilibrium case.

iv. Any connected tree has average degree 1. Here however it doesn't mean that for $d = 1$ the graph is a tree. It is however a collection of tree-like subgraphs.

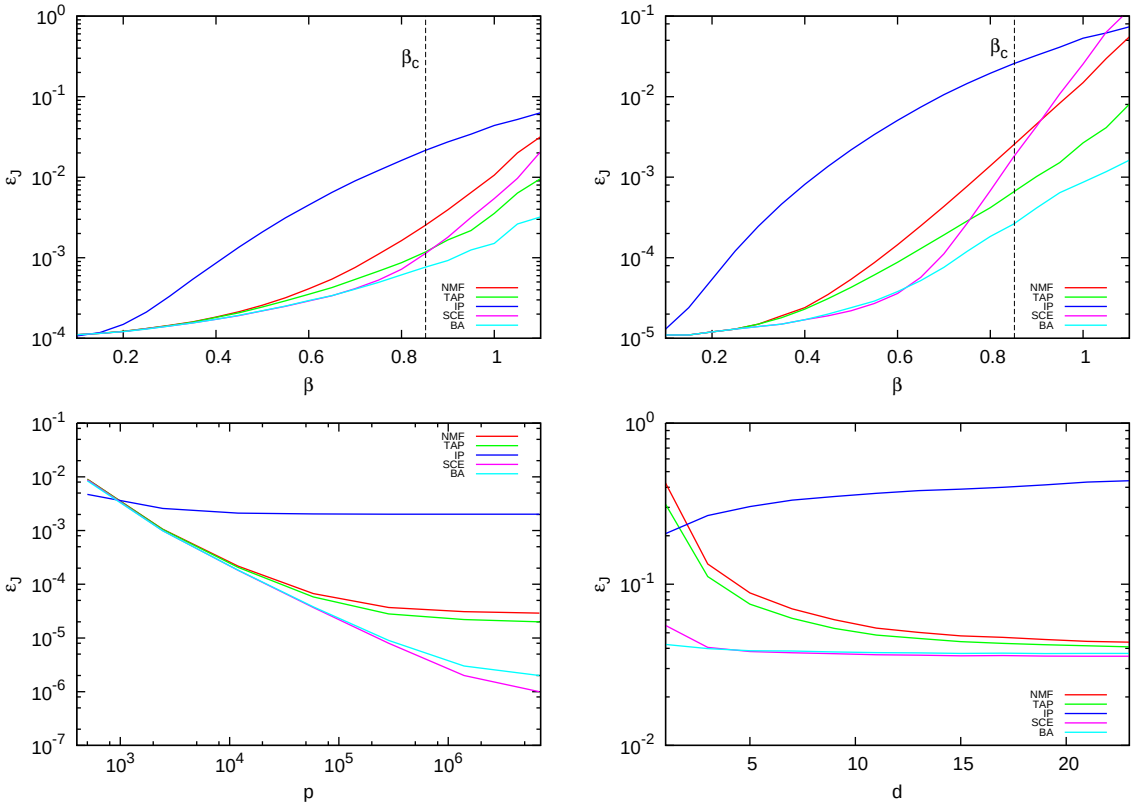


Figure 12.3: Upper row: Couplings inference error (eq.(12.1)) versus beta for a sparse $N = 100$ system with average degree $d = 10$ for $p = 10^4$ (left) and $p = 10^5$ (right). Low left: Couplings inference error (eq.(12.1)) versus p for a sparse $N = 100$ system with average degree $d = 10$ at $\beta = 0.5$. Low right: Couplings inference error (this time given by eq.(12.1)) versus the average degree d for systems of size $N = 100$ simulated at $\beta = 0.5$ using $p = 10^4$ samples.

and TAP. Interestingly, the BA method seems to perform equally good in sparse and dense situations.

12.3 ℓ_1-regularized Logistic Regression

In the above simulations we left out the last algorithm presented in this part of the thesis, the ℓ_1-regularized Logistic Regression (see chapter 11). The reason is that, as it is explained in the original paper [RavikumarWL 10], the purpose of this algorithm is slightly different. The presence of the ℓ_1 norm in the log-likelihood in eq.(11.5) has the following advantages: it allows sparse inference, *i.e.* it discerns the couplings who are exactly zero from others who might just be small, and it manages noisy data much better than the mean field methods allowing efficient inference in the $p \approx N$ regime. On the other hand there is a drawback, the ℓ_1 norm displaces the global minimum of the log-likelihood so that couplings, who might otherwise correctly be identified as non-zero, are inferred with wrong values. The result is that this algorithm is better suited for identifying the structure of the underlying graph, *i.e.* its edge-set, and not so much the actual values of the couplings.

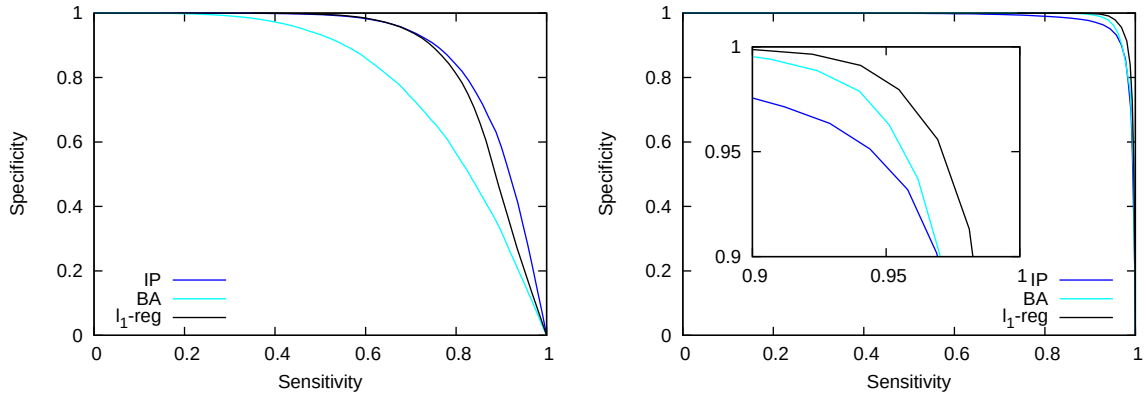


Figure 12.4: Receiver Operating Characteristic obtained by ℓ_1-regularized Logistic Regression and by thresholding the BA and IP methods for a system of size $N = 200$, average degree $d = 5$, simulated at $\beta = 0.4$ using $p = 400$ (left) and $p = 4000$ (right) samples.

In order to evaluate its performance we introduce an other kind of error better describing the quality of the graph structure inference. We introduce two quantities, the *True Positive Rate* (TPR) (aka the *sensitivity*) and the *True Negative Rate* (TNR) (aka the *specificity*) defined in the following way.

$$\mathrm{TPR} \equiv \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad , \tag{12.1}$$

$$\mathrm{TNR} \equiv \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \quad , \tag{12.2}$$

where

$$\begin{aligned} \mathrm{TP\ (True\ Positives)} &\equiv \text{Number of non-zero couplings correctly identified} \\ \mathrm{TN\ (True\ Negatives)} &\equiv \text{Number of zero couplings correctly identified} \\ \mathrm{FP\ (False\ Positives)} &\equiv \text{Number of non-zero couplings identified as zero} \\ \mathrm{FN\ (False\ Negatives)} &\equiv \text{Number of zero couplings identified as non-zero} \quad . \end{aligned}$$

Obviously the optimal inference corresponds to $\mathrm{TPR} = 1$ and $\mathrm{TNR} = 1$.

By varying the parameter λ in the algorithm described in section 11.1 we get the parametric curve $\mathrm{TPR}(\lambda), \mathrm{TNR}(\lambda)$, call a *Receiver Operating Characteristic* (ROC). In order to compare with the other methods in the task of inferring the graph of the model we chose to use the BA method, which has the best overall performance, and the IP method, which performs well in sparse systems when the number of samples is very low, and use a thresholding procedure in order to obtain a parametric curve similar with the one described above. By varying the parameter $J^{\min}$ and by setting equal to zero all couplings such that $\|J_{ij}\| < J^{\min}$ we obtain a ROC for the BA method and one for the IP. The results are plotted in fig.12.4.

In the left frame the number of samples is really low, $p = 400$. Interestingly we see that the ℓ_1-regularized Logistic Regression is able to infer the graph much better than BA but the

simpler IP method provides even better results. When the number of samples becomes larger, however, as in the right frame ($p = 4000$) the IP method becomes the worse choice, as expected, and the ℓ_1-regularized Logistic Regression the best one.

Part III

Exact Mean Field Theory in the Asymmetric Ising Model

Chapter 13

Asymmetric infinite-range model

As we have already discussed in part I, many biological systems are composed of a great number of elementary components interacting in a complex, non-regular way. In two archetypical examples, neural networks (NN) and gene regulatory networks (GRN), the components influence each other in a one-way fashion. A nerve signal is transmitted from the *soma* to the *dendrites* through the *axon* and in GRNs causal links are usually directed. Moreover, the variation of external stimuli (modeled as time-varying local fields) makes that the system cannot be described by an equilibrium measure. This has motivated the modeling of such systems by means of asymmetric kinetic networks. Asymmetric spin-glasses have been studied in a number of contexts [Derrida 87, DerridaGZ 87, GutfreundM 88, HertzGS 87, Parisi 86]. Although it has been shown that a dynamic phase transition occurs [Derrida 87, DerridaGZ 87] it is accepted that no "spin-glass" phase exists and therefore the individual spins never get locked in some particular configuration. As we will see soon this, together with the asymmetry of the couplings, allows an important simplification: the effective field felt by each spin acquires a Gaussian distribution and thus thermal averages can be replaced by a simple Gaussian integral. This was already noted in [GutfreundM 88].

We will examin a widespread variation of the asymmetric kinetic Ising model. One where time is discrete and spins are updated in parallel in every time step. We will establish two equations, one governing the time evolution of magnetizations, *i.e.* the ensemble averages of every spin, and an other governing the time evolution of correlations. These results have been presented in [MezardS 11, SakellariouRMH 12] reprinted in part IV. These equations relate the model parameters to the observable quantities in a simple way thus, they can be used in two ways. First, one can predict the observables at any time step, given an initial state and the values of the model parameters, *i.e.* one can solve the *direct* problem. Moreover, the relations can be inverted, and solved with respect to the model parameters, allowing one to solve the *inverse* problem, that is to infer the model that generated a given set of magnetizations and correlations.

13.1 The direct problem

First, the fields and couplings are considered to be time independent which leads to a stationary distribution. Including time dependent parameters in the direct problem would

yield equations of the same form, but with a slightly different interpretation. Still, we will present the two cases separately for clarity, since they yield different algorithms for the inverse problem.

13.1.1 The model

The model used throughout this section is the same as the one used in [RoudiH 11a, RoudiH 11b]. In these papers the authors adapt the naive mean-field and TAP theories, found in sections 7.1 and 7.2, to the case of asymmetric couplings. We will use these approximations in our simulations and compare their performances with our method. Interactions are infinite-range, as in the well known Sherrington-Kirkpatrick (SK) model, $J_{ij} \neq 0, \forall i, j$. Time is discrete and the N spins $s(t) = \{s_1(t), \ldots, s_N(t)\}$ evolve according to the following dynamics

$$P(s(t)|s(t-1)) = \prod_{i=1}^{N} \frac{1}{2\cosh(\beta h_i(t))} e^{\beta s_i(t) h_i(t)} \quad , \tag{13.1}$$

where

$$h_i(t) = H_i + \sum_j J_{ij} s_j(t-1) \quad . \tag{13.2}$$

The couplings J_{ij} are asymmetric, *i.e.* $J_{ij} \neq J_{ji}, \forall i \neq j$ and independent identically distributed according to a Gaussian distribution with mean 0 and variance $1/N$, as in the symmetric SK model

$$\rho(J) = \sqrt{\frac{N}{2\pi}} e^{-N\frac{J^2}{2}} \quad . \tag{13.3}$$

The local fields are also independent and distributed uniformly in $[-1, 1]$. The inverse temperature β can be seen simply as a parameter controlling the degree of independence of the spins, since it doesn't really correspond to the inverse physical temperature of the kinds of systems we are interested in.

In this section we use parallel Glauber dynamics for the numerical simulations. Starting from some random initial condition, in our case $s_i(t=0) = \pm 1$ with probability $1/2$, the spins are updated in parallel for p time steps. Then we use every configuration of the in $s \equiv \{s_i(t) : i = 1, \ldots, N; t = 1, \ldots, p\}$ to compute the *empirical* magnetizations and correlations, defined later in this chapter.

For comparison, we will also present in this chapter numerical results from a symmetric system. This system is exactly the same as the one described above with the only difference being its symmetric couplings matrix, $J_{ij} = J_{ji}$.

13.1.2 Magnetizations

We begin by deriving the mean-field equations for the magnetizations

$$m_i \equiv \langle s_i(t) \rangle = \langle \tanh \beta h_i(t) \rangle \quad . \tag{13.4}$$

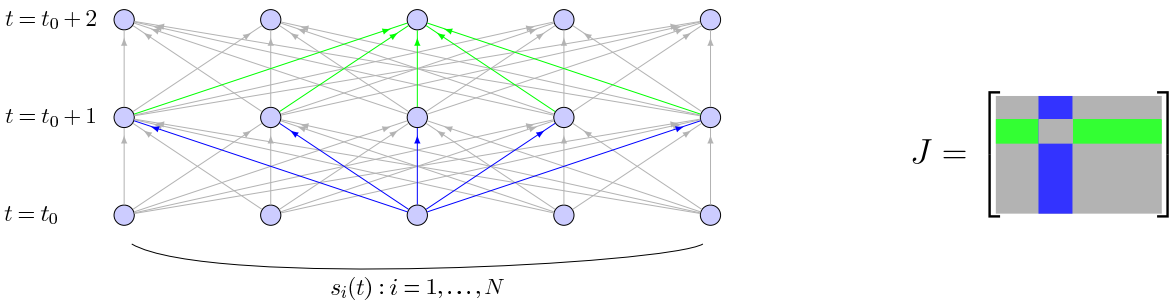


Figure 13.1: When the couplings matrix is asymmetric the in-coming and out-going couplings of one spin are uncorrelated and thus the terms in the local field due to the other spins are independent random variables, distributed according to the distribution of the couplings.

The local field on spin i due to the other spins, $\sum_j J_{ij} s_j(t-1)$, is the sum of a large number of terms. In the symmetric couplings case, these terms are correlated via the influence of the spins at time $t-2$. However, in the asymmetric case, the terms of the sum are independent due to the independence of J_{ij} and J_{ji} (see fig.13.1). Thus, as a sum of a large number of independent random variables, $\sum_j J_{ij} s_j(t-1)$ is a Gaussian random variable with mean

$$g_i \equiv \left\langle \sum_j J_{ij} s_j(t) \right\rangle = \sum_j J_{ij} m_j \tag{13.5}$$

and variance

$$\Delta_i \equiv \left\langle \left(\sum_j J_{ij} s_j(t) \right)^2 \right\rangle - \left\langle \sum_j J_{ij} s_j(t) \right\rangle^2 \tag{13.6}$$

$$= \sum_{j,k} J_{ij} J_{ik} \left[\langle s_j(t) s_k(t) \rangle - m_j m_k \right] \quad . \tag{13.7}$$

The sum in eq.(13.7) is dominated by the diagonal elements , therefore the variance can be written as

$$\Delta_i = \sum_j J_{ij}^2 (1 - m_j^2) \quad . \tag{13.8}$$

In fig.(13.2) we present some experimental evidence about the above claims. A $N = 100$ system is simulated according to the dynamics described by eq.(13.1,13.2) for $p = 10^5$ time-steps, first with symmetric (top row) and then with asymmetric couplings (bottom row), at three different values of β. What is plotted in every frame is a histogram of the different values the effective local field of the first spin, $\sum_j J_{1j} s_t(t)$, takes in every time-step, as well as the theoretical prediction of its distribution, *i.e.* a Gaussian with the appropriate mean and variance (eq. 13.5 and 13.8). In the asymmetric case the curve matches really well the experimental data: the contribution of every spin in the effective local field is independent of the others. In the symmetric case, however, the prediction is clearly wrong. At high temperatures (small β) the error is small because the spins still have some degree of independence due to the thermal noise. On the other hand at low temperatures (high β) the Gaussian fails completely to

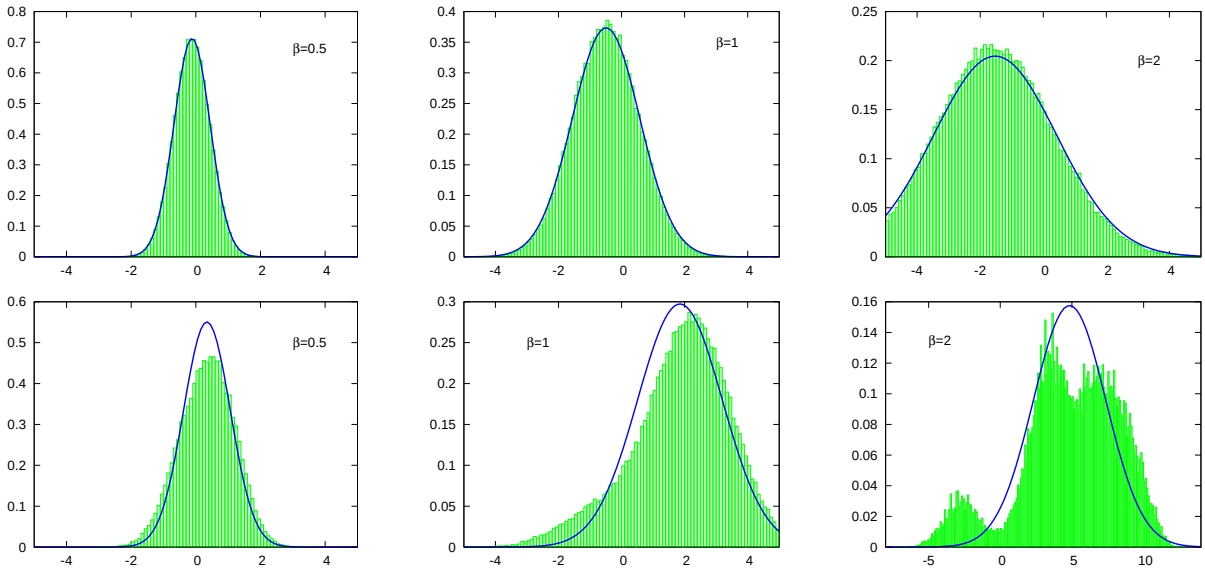


Figure 13.2: Evidence for the fact that the effective local field acting on a spin, due to the presence of the other spins, has a Gaussian distribution in the asymmetric case. Top row: a asymmetric system of $N = 100$ spins is simulated, according to the dynamics described by eq.(13.1) for $p = 10^5$ time-steps at three inverse temperatures $\beta = 0.5, 1, 2$ from left to right. The histograms show the effective local field of the first spin scaled by β, $\beta \sum_j J_{1j} s_j(t)$, for the whole time series. The blue curve is the theoretical prediction $\frac{1}{\sqrt{2\pi\beta^2\Delta_1}} \exp(-(x - \beta g_1)^2/(2\beta^2\Delta_1))$, where g_1 and Δ_1 are given by eq.(13.5,13.8). Bottom row: the same but for a system with symmetric couplings. At high β the distribution is clearly not Gaussian and has multiple modes, indicating the existence of metastable states, typical of the glassy phase.

describe the true distribution. The distribution develops multiple modes beyond $\beta_c = 1$, which is the inverse of the critical temperature of the SK model, indicating an ergodicity breaking, typical of the glassy phase. The system is stuck in some local minimum of the free energy and only explores a small sub-region of phase space. This creates one mode in the distribution of the effective local field. Eventually the system escapes the local minimum and finds itself trapped in some other region and so on. This creates the complex profile seen in the bottom right frame of fig.(13.2). Lowering the temperature even more will eventually lead to a series of delta peaks (not shown).

Using the above remarks we can replace the ensemble average of eq.(13.4) by a Gaussian integral and thus obtain the magnetizations of each spin as a function of the remaining magnetizations

$$m_i = \int Dx \tanh\left[\beta\left(H_i + g_i + x\sqrt{\Delta_i}\right)\right] \quad , \tag{13.9}$$

where $Dx \equiv \frac{dx}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$ is the measure of a Gaussian variable x with zero mean and variance unity.

Equations (13.5,13.8,13.9) are our mean field equations. The abbreviation MF will be used hereafter for any reference to these equations as well as others derived from those, as those

describing the correlations and those used in the inverse problem found later in this chapter. They rely only in the central limit theorem and are exact for asymmetric systems in the limit $N, p \to \infty$. They can also be used as an approximation for symmetric systems, as long as the temperature is not too low.

It is important to state that in this form they have no predictive power. One cannot compute any magnetization if he doesn't have a priori knowledge of all the magnetizations. However they are useful for the inverse problem since they relate the magnetizations with the parameters of the model, and thus can be inverted to infer the model.

It is instructive to compare eq.(13.9) with the corresponding equations of naive mean field theory (NMF) and the TAP approximation found in [RoudiH 11a, RoudiH 11b]. We rewrite the two equations here using our notation.

$$m_i = \tanh\left[\beta\left(H_i + g_i\right)\right] \tag{13.10}$$

and

$$m_i = \tanh\left[\beta\left(H_i + g_i - m_i\beta\Delta_i\right)\right] \tag{13.11}$$

We expand the quantity $\phi_i \equiv \frac{1}{\beta}\text{atanh}m_i$ in powers of Δ_i using the Gaussian mean field expression for m_i from eq.(13.9).

$$\phi_i = u_i - \beta\Delta_i \tanh(\beta u_i) + 2\beta^3\Delta_i^2 \tanh(\beta u_i)[1 - \tanh^2(\beta u_i)] + \mathcal{O}(\Delta_i^3) \quad , \tag{13.12}$$

where $u_i \equiv H_i + g_i$. We see that the first term is just the naive mean field equation (13.10). This is a well known fact of mean field theory: by taking $\left\langle \tanh(H_i + \sum_j s_j)\right\rangle \approx \tanh(H_i + \sum_j \langle s_j\rangle)$ we completely threw away the variance, *i.e.* Δ. We now do the same for the TAP equation (13.11) and obtain

$$\phi_i = u_i - \beta\Delta_i \tanh(\beta u_i) + \beta^3\Delta_i^2 \tanh(\beta u_i)[1 - \tanh^2(\beta u_i)] + \mathcal{O}(\Delta_i^3) \quad . \tag{13.13}$$

They differ at order Δ_i^2. Equations (13.12,13.13) are high temperature expansions, therefore both NMF and TAP are high temperature (or weak couplings) approximations to the correct result.

The interpretation of eq.(13.11) raises a small paradox. The improvement of TAP over NMF is due to the Onsager *reaction term*, $m_i\beta\sum_j J_{ij}^2(1 - m_j^2)$. In symmetric systems there is a physical, cavity-type, argument leading to this term. The idea is the following. In eq.(13.10) the magnetizations appearing in $g_i = \sum_j J_{ij}m_j$ are not the actual ones, but the magnetizations the spins would have in the absence of spin i. Eq.(13.10) would be correct if the spin i is influenced by but does not influence the remaining spins. What actually happens is that the magnetizations of the remaining spins are shifted by $\chi_{jj}J_{ij}m_i$, as we explained in section 7.2, where

$$\chi_{jj} \equiv \left.\frac{\partial m_j}{\partial h_j}\right|_{h_j=0} = \beta(1 - m_j^2) \quad , \tag{13.14}$$

hence the form of eq.(13.11). Here, χ_{jj} is the magnetic susceptibility, *i.e.* the answer of m_j to a small change of the effective field h_j. This is true for symmetric systems. However, in the asymmetric case the reaction term, following the same cavity argument, would be

$m_i\beta\sum_j J_{ij}J_{ji}(1-m_j^2)$. Since $J_{ij}\neq J_{ji}$ and $J_{ij}\sim\mathcal{O}(1/\sqrt{N})$ this term vanishes for $N\to\infty$ and there is no reaction phenomenon. However, there is an alternative way of deriving the TAP equations without the use of a cavity argument. This is done by a second order Plefka-type expansion which yields the term $m_i\beta\sum_j J_{ij}^2(1-m_j^2)$ for both symmetric and asymmetric systems. In conclusion, although the physics that led originally to the TAP equations is not valid for asymmetric systems, they are still a weak couplings expansion of the exact equations.

In fig.(13.3) we present a series of scatter plots depicting the results of the three mean field methods, for both asymmetric (top row) and symmetric (bottom row) systems, at inverse temperatures β =0.5, 1 and 2 (from left to right). In the asymmetric case Gaussian mean field theory (denoted simply MF hereon) is always correct while NMF and TAP become worse at high β. In the symmetric case all three methods are inexact. However both MF and TAP are descent approximations in the high temperature region, $\beta<1$. Note that in the bottom right frame ($\beta=2$) TAP predictions are closer to the true values of the magnetizations than those predicted by MF. A more detailed numerical study comparing the results of the three mean-field methods between symmetric and asymmetric systems can be found in [SakellariouRMH 12].

The behavior of the errors as a function of β can be better appreciated in the next set of figures (13.4). What is shown is the mean squared error of the magnetizations $\overline{(m_i^{true}-m_i^{inferred})}$ made by the three mean field methods, as a function of β, for asymmetric (left) and symmetric systems (right). The systems used have $N=100$ spins and are simulated for $p=10^5$ times. The whole procedure is repeated 50 times for different realizations of the system and the curves are the average over these realizations.

13.1.3 Correlations

We now turn to the problem of computing the correlations. We begin by defining the two types of correlations suited for the study of non-equilibrium systems. Equal-time correlations and time-delayed correlations,

$$C_{ij}\equiv\langle\delta s_i(t)\delta s_j(t)\rangle \tag{13.15}$$

and

$$D_{ij}\equiv\langle\delta s_i(t+1)\delta s_j(t)\rangle \quad , \tag{13.16}$$

where $\delta s_i(t)$ is the fluctuation of the magnetization $\delta s_i(t)=s_i(t)-\langle s_i(t)\rangle$. We shall establish a mean field relation between the matrices C and D using the same arguments as before, about the Gaussianity of the effective field.

We start by writing $\sum_j J_{ij}s_j(t)=g_i+\delta g_i(t)$, where $\delta g_i(t)$ is Gaussian distributed with mean 0 and variance Δ_i. Now, by definition of D_{ij} we have

$$D_{ij}=\langle s_j(t)\tanh\left[\beta\left(H_i+g_i+\delta g_i(t)\right)\right]\rangle-\langle s_j(t)\rangle\langle\tanh\left[\beta\left(H_i+g_i+\delta g_i(t)\right)\right]\rangle \quad . \tag{13.17}$$

Multiplying by the couplings matrix J we have

$$\begin{aligned}\sum_k J_{jk}D_{ik}&=\langle\left(g_j+\delta g_j\right)\tanh\left[\beta\left(H_i+g_i+\delta g_i\right)\right]\rangle-g_j\langle\tanh\left[\beta\left(H_i+g_i+\delta g_i\right)\right]\rangle\\&=\langle\delta g_j\tanh\left[\beta\left(H_i+g_i+\delta g_i\right)\right]\rangle\end{aligned} \tag{13.18}$$

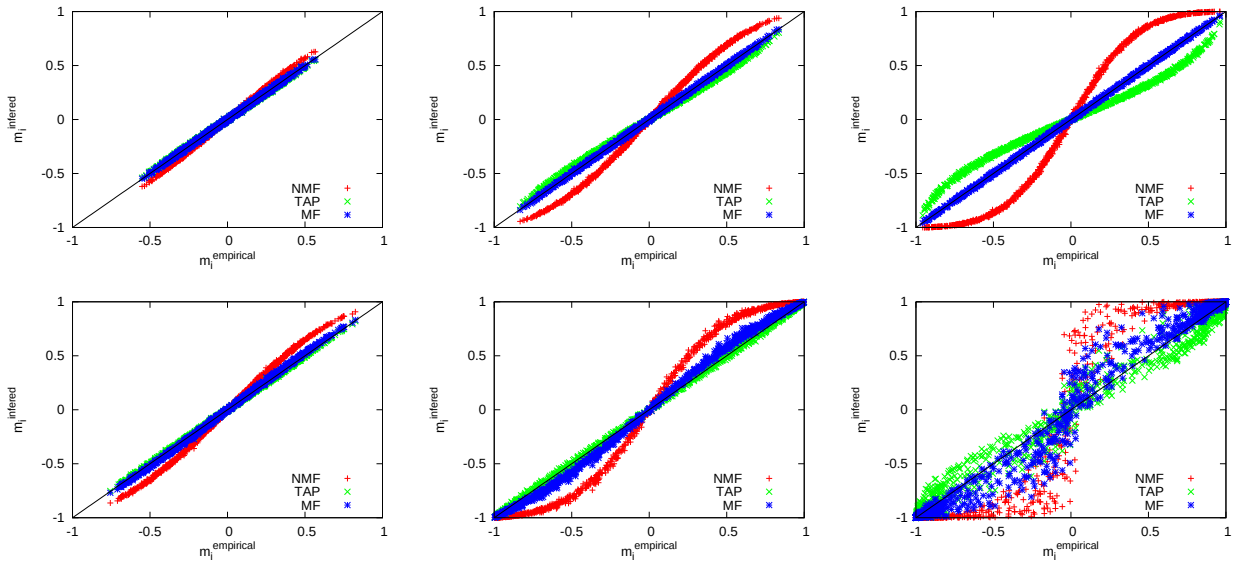


Figure 13.3: Scatter plots of inferred versus predicted magnetizations. Top row: a $N = 10^3$ spins system with asymmetric couplings matrix is simulated for $p = 10^4$ time-steps according to the dynamics defined in eq.(13.1,13.2) at inverse temperatures β =0.5, 1, 2 (from left to right). What is plotted are the magnetizations, inferred according to the three mean field equations (13.9,13.10,13.11). The color scheme is: NMF red, TAP green, MF blue. The figures strongly suggest that eq.(13.9) is exact for any choice of β, whereas NMF and TAP are only high temperature approximations with TAP performing better than NMF. Bottom row: the same but for a system with symmetric coupling matrix. At high temperatures (left panel) MF and TAP are close to the correct result and they both outperform NMF. However, as the temperature is lowered, they both give incorrect results, with TAP being a better approximation.

In order to evaluate the average we need the joint distribution of δg_i and δg_j. The crucial point to keep in mind is that, as the couplings are of order $1/\sqrt{N}$, each matrix element of C and D is also small, of order $1/\sqrt{N}$. Their covariance is therefore small:

$$\begin{aligned} \langle \delta g_i \delta g_j \rangle &= \left\langle \sum_k J_{ik} \left(s_k - \langle s_k \rangle \right) \sum_l J_{jl} \left(s_l - \langle s_l \rangle \right) \right\rangle \\ &= \sum_{k,l} J_{ik} J_{jl} C_{kl} = \left(J C J^T \right)_{ij} \equiv \varepsilon \ , \end{aligned} \tag{13.19}$$

where ε is typically of order $1/\sqrt{N}$. So the joint distribution of $x = \delta g_i$ and $y = \delta g_j$ takes the form, in the large N limit (omitting terms of order ε^2):

$$P(x,y) = \frac{1}{2\pi\sqrt{\Delta_i \Delta_j}} \exp\left(-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j} + \varepsilon \frac{xy}{\Delta_i \Delta_j} \right) \tag{13.20}$$

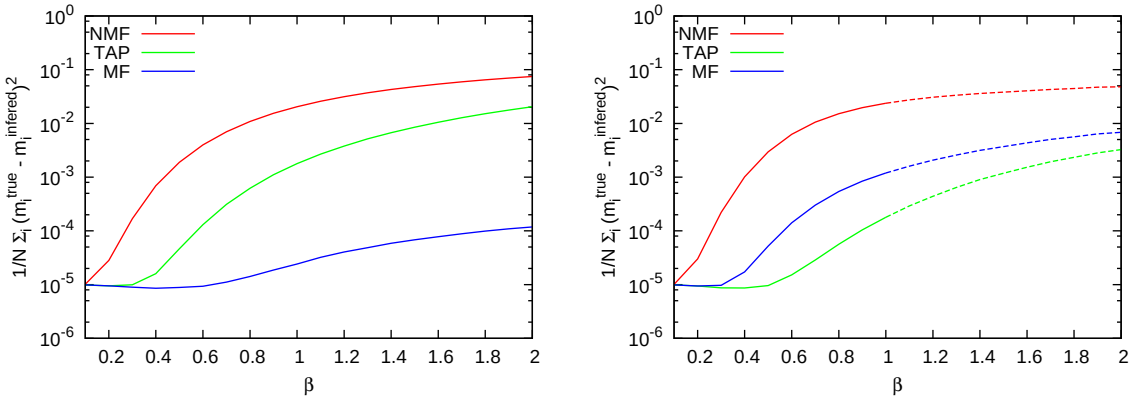


Figure 13.4: Mean squared error in predicting the magnetizations $\overline{(m_i^{true} - m_i^{inferred})}$, for the three methods MF (blue), TAP (green) and NMF (red), as a function of β in a $N = 100$ asymmetric (left) and symmetric (right) system. $p = 10^5$ samples were used. Note: The error bars for the asymmetric case are negligible. The same applies in the symmetric case for $\beta < 1$. However in the glassy phase ($\beta > 1$) fluctuations become important, as the ergodicity of the phase space trajectory, and thus the sampling quality, fluctuate a lot. We did not include the large error bars as they do not show well in logarithmic scale, but used dashed lines to stress that the results are only indicative. Despite that, the curves are still relevant because they show the relative distance between the errors.

Using the small ε expansion of eq. (13.20) we can rewrite eq. (13.18) as

$$\begin{aligned}\sum_k J_{jk} D_{ik} &= \frac{\varepsilon}{\Delta_i \Delta_j} \int \frac{dx}{\sqrt{2\pi\Delta_i}} \frac{dy}{\sqrt{2\pi\Delta_j}} \mathrm{e}^{-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j}} xy^2 \tanh\left[\beta\left(H_i + g_i + x\right)\right] \\ &= \varepsilon\beta \int \frac{dx}{\sqrt{2\pi\Delta_i}} \exp^{-\frac{x^2}{2\Delta_i}} \left(1 - \tanh^2\left[\beta\left(H_i + g_i + x\right)\right]\right) \ . \end{aligned} \tag{13.21}$$

Combining eq. (13.19) and eq. (13.21) we get:

$$\left(DJ^T\right)_{ij} = \left(JCJ^T\right)_{ij} \beta \int \frac{dx}{\sqrt{2\pi\Delta_i}} \mathrm{e}^{-\frac{x^2}{2\Delta_i}} \left(1 - \tanh^2 \beta\left(H_i + g_i + x\right)\right) \ , \tag{13.22}$$

which gives the explicit mean-field relation between C and D

$$D = A\, J\, C \ , \tag{13.23}$$

where A is a diagonal matrix: $A_{ij} = a_i \delta_{ij}$, with:

$$a_i = \beta \int Dx \ \left[1 - \tanh^2 \beta\left(H_i + g_i + x\sqrt{\Delta_i}\right)\right] \ . \tag{13.24}$$

Again, as in the case of magnetizations, the result takes the same form as the ones found by the naive mean field and TAP approach. The difference is found in the diagonal matrix A and comes from the evaluation of the ensemble average in eq.(13.16). The NMF and TAP results are

$$a_i^{nMF}(t) = \beta\left[1 - m_i(t+1)^2\right] \ , \tag{13.25}$$

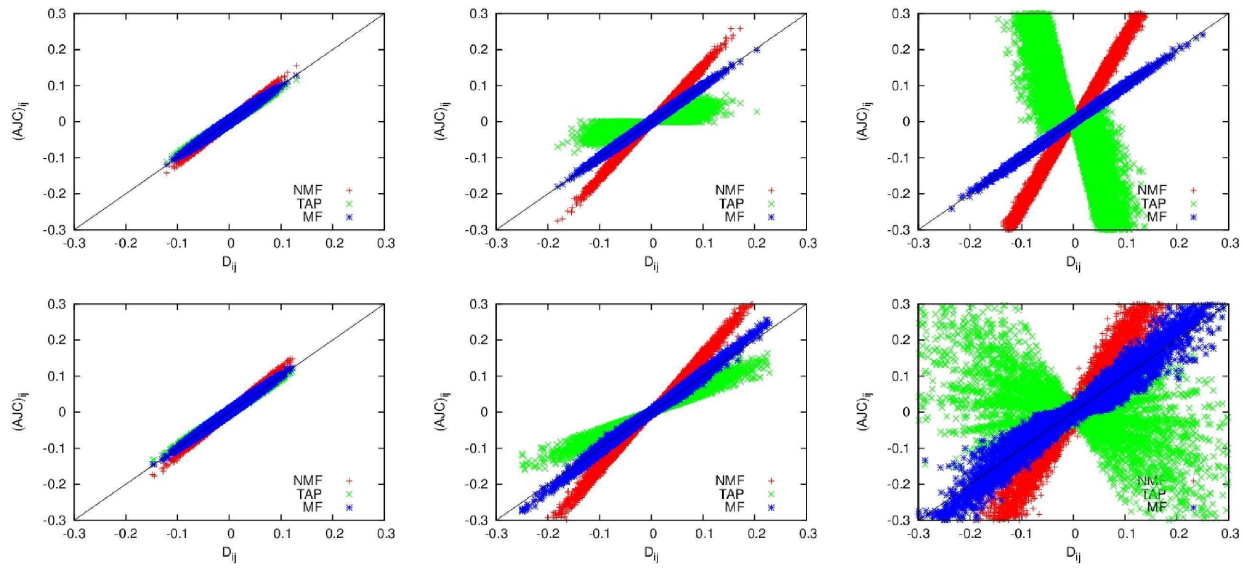


Figure 13.5: Top row: A $N = 100$ asymmetric system is simulated for $p = 10^5$ time steps at β=0.5, 1 ,2 form left to right. The plot shows AJC versus D for NMF (red), TAP (green) and MF (blue), where the matrices C and D are estimated from Monte Carlo simulation. Bottom row: the same but with symmetric couplings.

the 'TAP' approximation gives:

$$a_i^{TAP}(t+1) = \beta\left[1 - m_i(t+1)^2\right]\left[1 - (1 - m_i(t+1)^2)\beta^2 \sum_k J_{ik}^2 (1 - m_k(t)^2)\right] \tag{13.26}$$

As in the case of magnetizations, eq.(13.23,13.24) are exact in the large N limit when the couplings matrix is asymmetric. This result also relies solely in the central limit theorem. Figure 13.5 shows scatter plots of the three methods predictions for the correlations. A $N = 100$ spins system with asymmetric couplings (top row) is simulated for $p = 10^5$ time steps at inverse temperatures β =0.5, 1 and 2 and we plot the predicted values of the matrix $D = AJC$ versus the empirical, found in the simulation. We note again that MF is exact for all choices of the temperature, whereas NMF and TAP are only high temperature approximations. TAP outperforms NMF in high temperatures but overshoots at smaller ones making it completely unsuitable as an approximation. In the bottom row we also give plots of symmetric systems for the same set of N, p and β. Interestingly, unlike in the case of magnetizations, MF remains by far the better approximation even in this case because of the overshooting of TAP. Its predictions are in no way exact, but they are a descent approximation, especially in the $\beta < \beta_c$ phase.

The above remarks can be better appreciated in fig. 13.6 where we plot the mean squared error of the prediction of D, $1/N^2 \sum_{i,j}(D_{ij} - (AJC)_{ij})^2$ as a function of β. The system has size $N = 100$ and the simulation spans $p = 10^5$ time steps. The curves are averaged over 50 realizations of the system.

The results for the correlations, presented above, are very important when it comes to the

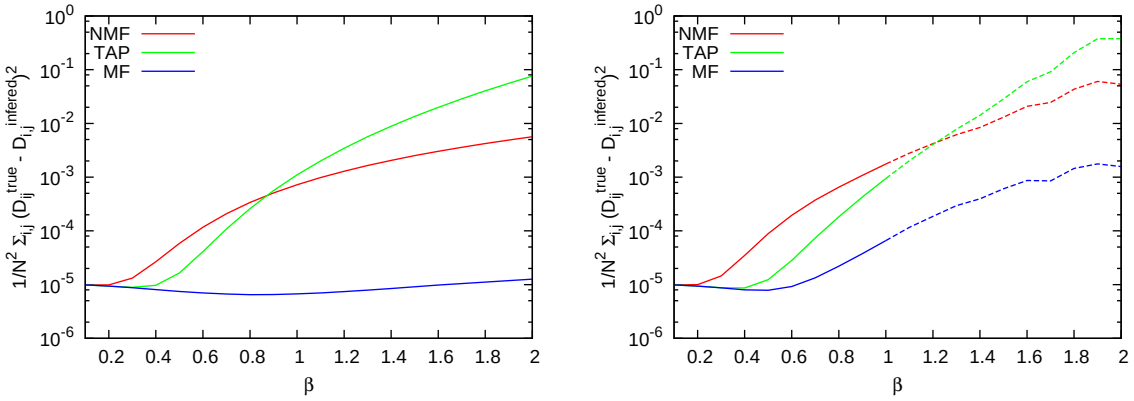


Figure 13.6: Mean squared error in predicting the time-delayed correlations $\overline{(D_{ij}^{true} - D_{ij}^{inferred})}$, for the three methods MF (blue), TAP (green) and NMF (red), as a function of β in a $N = 100$ asymmetric (left) and symmetric (right) system. $p = 10^5$ samples were used. See caption of fig. 13.4 about the error bars.

inverse problem. As we will see in the next section, we can have an efficient algorithm for inferring the couplings matrix of an unknown model by inverting the matrix relation (13.23). But before turning to the inverse problem let's make a few comments on the applicability of the above results in the non stationary case, where the model parameters are varying with time.

13.1.4 Non Stationary Case

In the general setting both the couplings J and the local fields H can vary with time. In our simulations, however, we focused on the particular case where the couplings are time independent and the local fields vary uniformly as

$$H_i(t) = H_0 \sin(\omega t) \quad . \tag{13.27}$$

This can be used to model some interesting experimental situations in the context of neural networks, where one observes a fixed structure (e.g. the retina) subject to a varying external stimulus (e.g. an image of varying intensity). The generalization of time varying couplings is straightforward and will be omitted from this presentation.

The dynamics used is, as before, given by eq.(13.1) with eq.(13.27) substituted in eq.(13.2). The procedure, however, is slightly altered: Starting from some initial configuration $s(0)$ we run the parallel Glauber algorithm for $t_{\max}$ time steps. This will produce a time series $\{s(t) : t = 1, \ldots, t_{\max}\}$, referred to as a *run* hereafter. Then, we set $t = 0$, initialize the spins back to the same initial configuration and make another run, and so on until we have p runs. We end up with a set of data which can be written $s = \{s_i^r(t) : i = 1, \ldots, N; t = 1, \ldots, t_{\max}; r = 1, \ldots, p\}$. The averages are then taken separately for each time t with respect to the index r.

$$m_i(t) \equiv \langle s_i^r(t) \rangle_{r=1,\ldots,p} \tag{13.28}$$

$$C_{ij}(t) \equiv \left\langle s_i^r(t) s_j^r(t) \right\rangle_{r=1,\ldots,p} - m_i(t) m_j(t) \tag{13.29}$$

$$D_{ij}(t) \equiv \left\langle s_i^r(t+1) s_j^r(t) \right\rangle_{r=1,\ldots,p} - m_i(t+1) m_j(t) \tag{13.30}$$

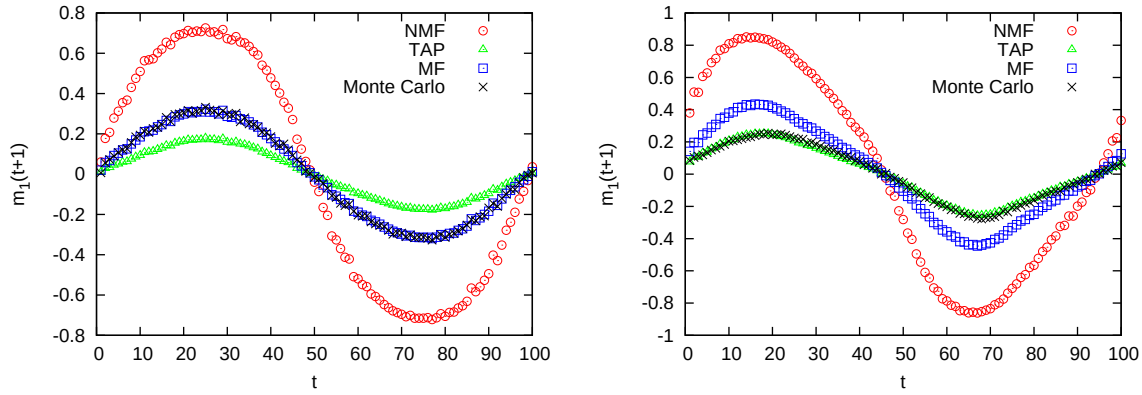


Figure 13.7: Time evolution of the magnetization of one spin for a $N = 100$ spins system computed from $p = 10^4$ Monte Carlo runs of length $t_{\max} = 100$ at inverse temperature $\beta = 2$ (black x's), and their theoretical predictions based on NMF (red), TAP (green) and MF (blue) theories. The external field is of the form of eq.(13.27) with $H_0 = 0.5$ and $\omega = 2\pi/100$ Left: asymmetric couplings. Right: symmetric couplings.

The r index will be dropped hereafter.

The derivation of the mean field equations is straightforward. One only needs to keep track of the time indices. The result for the magnetizations is

$$m_i(t+1) = \int Dx \tanh\left[\beta\left(H_i(t) + g_i(t) + x\sqrt{\Delta_i(t)}\right)\right] \tag{13.31}$$

and for the correlations

$$D(t) = A(t)\; J\; C(t)\;, \tag{13.32}$$

with $A_{ij}(t) = a_i(t)\delta_{ij}$ and

$$a_i(t) = \beta \int Dx\;\left[1 - \tanh^2\beta\left(H_i(t) + g_i(t) + x\sqrt{\Delta_i(t)}\right)\right]\;. \tag{13.33}$$

In figure 13.7 we can see the predictions of the three methods in the non stationary case. Not surprisingly MF matches perfectly the experimental data in the asymmetric case where NMF and TAP fail to do so and fails in the symmetric one, where TAP performs better. Note however that in the low temperature phase of the symmetric case, as in the right frame of figure 13.7, the results vary a lot from one example to another and TAP is just on average better that MF, as opposed to the asymmetric case where MF performs always better than TAP.

13.2 The inverse problem

As mentioned in the beginning of this chapter, unlike in the traditional realm of statistical mechanics where systems of atoms or molecules have always symmetric interactions, in many biological systems components rarely interact in a symmetric way. Traditionally, in physics, the microscopic details of systems of interest are postulated on the basis of symmetry and

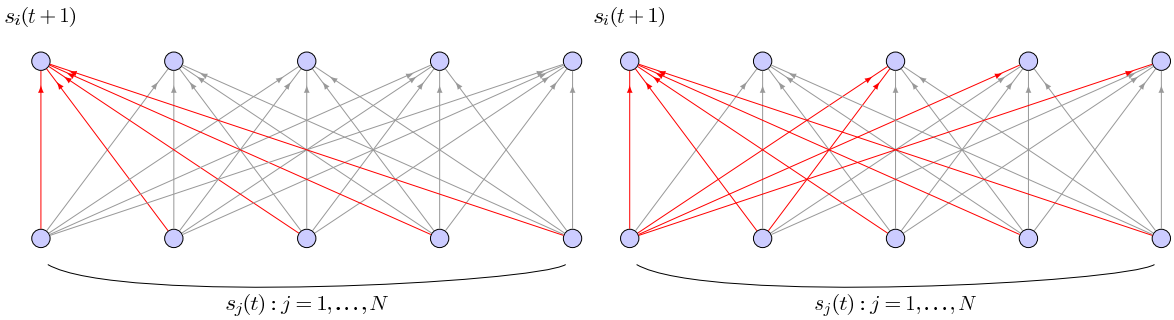


Figure 13.8: Graph representing 2 time steps of a 5 spins system. Highlighted in red are the couplings corresponding to one row of J. In the asymmetric case (left) they represent all in-coming couplings in one spin and can be inferred independently of the rest. In the symmetric case however (right) they are related to the corresponding column of J and the in-coming neighborhoods are not independent.

homogeneity arguments and on the understanding of the fundamental physical laws. In biology, however, such a reasoning would be catastrophic since the complex details of living organism are essential to their functioning and behavior. Hence the need for a method for inferring the complex network of interacting components based on observations of their collective behavior.

In the previous section we introduced a set of relations describing the behavior of magnetizations and correlations which are asymptotically exact for asymmetric systems. These relations will help us establish an algorithm for solving the inverse problem which will be both numerically exact and efficient in terms of its time complexity.

13.2.1 Stationary case

Let's start with the stationary case, where both couplings and local fields are independent of time. We begin by rewriting the two main equations of the previous section in a simplified way. First, the factor β representing the inverse temperature at which the data are generated, will be absorbed in the couplings and local fields. It's just a global scaling factor and cannot be inferred separately. Second, we will notice that the inverse problem decouples in a set of smaller problems: the in-coming couplings for each spin can be inferred independently. The reason for that lies again in the asymmetry of the couplings. In the symmetric case the in-coming couplings of one spin take the same values as the out-going ones. But the out-going couplings of one spin are the in-coming of other spins and so on. In the asymmetric case, however, this is not true and all the couplings of the system can be viewed as a collection of independent sets of in-coming couplings, one for every spin. See figure 13.8. We are going to further simplify the notation by dropping also the i index and write, for example, $m = m_i$, $\Delta = \Delta_i$, $J_j = J_{ij}$ and so on. The equations for the magnetizations now read

$$m = \int Dx \ \tanh\left[H + g + x\sqrt{\Delta}\right] \ . \tag{13.1}$$

For the correlations some additional notation is necessary. We can obviously rewrite eq.(13.23) as

$$J = [A(J)]^{-1} D C^{-1} \tag{13.2}$$

where we have stressed the J dependence of A through g_i and Δ_i. We introduce the matrix $B = DC^{-1}$ with column vectors $b_j^{(i)} = \sum_k D_{ik} C_{kj}^{-1}$. Now, dropping the i index, one can infer each row of the J from

$$J_j = b_j/a \quad , \tag{13.3}$$

with

$$a = \int Dx \left(1 - \tanh^2 \left[H + g + x\sqrt{\Delta}\right]\right) \ . \tag{13.4}$$

Additionally we have the following link between a and Δ

$$\Delta = \frac{1}{a^2} \sum_j b_j^2 (1 - m_j^2) \equiv \frac{\gamma}{a^2} \tag{13.5}$$

Obviously one cannot compute J_j directly from eq.(13.3) because a depends on the J_j's. We therefore propose the following iterative procedure

Algorithm 13.2.1: MEAN FIELD INVERSE ISING, STATIONARY CASE(m, C, D)

Invert C
for $i \leftarrow 1$ **to** N
{
  for $j \leftarrow 1$ **to** N
    do $b_j \leftarrow \sum_k D_{ik} C_{kj}^{-1}$
  $\gamma \leftarrow \sum_j b_j^2 (1 - m_j^2)$
  $\hat{\Delta} \leftarrow \Delta_0$
  while $\Delta \neq \hat{\Delta}$
    do {
      $\Delta \leftarrow \hat{\Delta}$
      Using m_i compute $u \equiv H_i + g$ by inverting eq.(13.1)
      Using u and Δ compute a using eq.(13.4)
      $\hat{\Delta} \leftarrow \gamma / a^2$
    }
  for $j \leftarrow 1$ **to** N
    do $J_{ij} = b_j / a$
  $g \leftarrow \sum_j J_{ij} m_j$
  $H_i \leftarrow u - g$
}
return (J, H)

The above algorithm takes as inputs the vector m and the matrices C and D, found experimentally, and returns the parameters of the corresponding Ising model H and J. For every spin it first computes γ and then, starting from some initial value for Δ, it iterates equations (13.1,13.4,13.5) until the value of a that satisfies eq.(13.23) is found. One then founds the couplings simply by using eq.(13.3) and the local fields by subtracting $g = \sum_j J_{ij} m_j$ from the total effective field u, also found during the iteration.

It is also important to realize that, in the $N \to \infty$ limit, Δ becomes independent of index i. Therefore, for large systems, one can modify the above algorithm to run the while loop only once and then use the same value for Δ in the calculations of the remaining a_i's.

Let's now look at the *time complexity* of the algorithm. The inversion of C and the computation of B, through matrix multiplication, dominate the time complexity of the algorithm. The while loop converges exponentially fast when some conditions are met, as we will see shortly. The time complexity is therefore $\mathcal{O}(N^3)$. (Actually, there exist $\mathcal{O}(N^{2.807})$ and $\mathcal{O}(N^{2.376})$ algorithms for matrix multiplication and inversion, but we are not interested here). This is the case when we already have m, C and D. In applications, however, one usually starts form the raw data $s = \{s_i(t) : i = 1, \ldots, N; t = 1, \ldots, p\}$. Then, the computation of m takes $\mathcal{O}(Np)$ time and that of C and D $\mathcal{O}(N^2 p)$. Since $p > N$, as we will see shortly, this is the part that actually dominates the whole procedure.

The second important remark concerns the convergence condition which is connected with the *sample complexity*. First, the number of samples cannot be less than N, otherwise the matrix C wouldn't be invertible. This is a well known fact and has to do with the linear independence of its row vectors and column vectors. One cannot construct N linearly independent vectors of the form $v_j^{(i)} = \frac{1}{P}\sum_{t=1}^{p} s_i^t s_j^t$ with $p < N$. That's a necessary condition but, unlike in the NMF and TAP derived algorithms, it is not sufficient. A stronger condition must be imposed and concerns the while loop. The procedure inside the while loop maps a value of Δ to a new one $\hat{\Delta}$ and halts when a fixed point is found. This can be described by a function $\hat{\Delta} = f(\Delta)$. The fixed point exists if f intersects the diagonal $y = x$ at some finite point and it is stable if $df/d\Delta \in]-1, 1[$. The function f can be written in a compact way as

$$f(\Delta) = \frac{\gamma}{a(u(\Delta), \Delta)^2} \tag{13.6}$$

where a is given by eq.(13.4) and $u(\Delta)$ satisfies eq.(13.1) for given m and Δ. The value at the origin is easily found $f(0) = \frac{\gamma}{(1-m^2)^2}$, which is, as expected, nothing but the NMF result. We can also compute the asymptotic behavior of f for $\Delta \to \infty$ by noting that

$$\tanh(u + x\sqrt{\Delta}) \sim 2\theta(u + x\sqrt{\Delta}) - 1 \quad \text{and} \tag{13.7}$$

$$1 - \tanh^2(u + x\sqrt{\Delta}) \sim \frac{2}{\sqrt{\Delta}}\delta(u + x\sqrt{\Delta}) \quad , \tag{13.8}$$

where θ and δ are the usual Heaviside step and Dirac delta functions. Therefore we have that equations (13.1,13.4) behave for large Δ as

$$m \sim \operatorname{erf}(\frac{u}{\sqrt{\Delta}}) \quad \text{and} \tag{13.9}$$

$$a \sim \sqrt{\frac{2}{\pi\Delta}} \mathrm{e}^{-u^2/2\Delta} \quad . \tag{13.10}$$

Combining all that we finally have

$$f \sim \frac{\pi}{2}\gamma \mathrm{e}^{\hat{u}^2}\Delta \quad , \tag{13.11}$$

where we introduced the auxiliary variable $\hat{u} \equiv \frac{u}{\sqrt{\Delta}}$ which is such that $m = \operatorname{erf}(\hat{u}/\sqrt{2})$.

Since, for finite temperature, $\gamma > 0$ we must have $f'_\infty = \frac{\pi}{2}\gamma \mathrm{e}^{\hat{u}^2} \in]0, 1[$ in order for f to have a stable fixed point. The term $\gamma \equiv \sum_j b_j^2(1 - m_j^2)$ can be easily computed from the data and

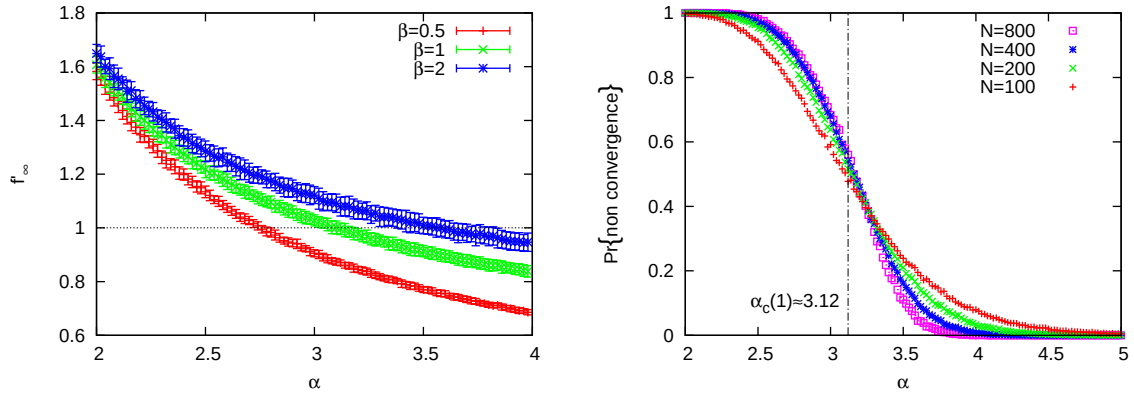


Figure 13.9: Left: Average of the asymptotic value of the slope f'_∞ as a function of the parameter $\alpha = p/N$ for three values of the inverse temperature β =0.5 (red), 1 (green) and 2 (blue). The system in question has $N = 100$, the simulation is repeated for 50 different systems and the results are averaged over each spin and each system. Right: The probability that the algorithm 13.2.1 fails to converge as a function of α. The systems used have sizes N =100 (red), 200 (green), 400 (blue) and 800 (magenta) and are simulated at $\beta = 1$. The vertical line is drawn at $\alpha \approx 3.12$ which is the point at which f'_∞ becomes smaller than 1 (see frame to the left).

so can $\hat{u}$. The left frame of figure 13.9 show numerical results for the asymptotic slope as a function of the ratio of samples to spins $\alpha \equiv p/N$. A system of size $N = 100$ is simulated at β =0.5, 1 and 2 and then the value of f'_∞ is averaged over all spins and over 50 repetitions of the experiment. We see that the curves cross the line $y = 1$ at different values of α for different inverse temperatures. We will name these values $\alpha_c(\beta)$ since they are the critical values of samples to spins ratio beyond which the algorithm converges. $\alpha_c(\beta)$ is an increasing function of β. This means that at lower temperatures we need more samples, which is logical since the samples become more and more correlated and so their information content is smaller.

In the right frame of figure 13.9 we see the probability that the while loop in algorithm 13.2.1 fails to converge as a function of α. The system into consideration is simulated at $\beta = 1$. The cross-over becomes steeper for larger systems and we expect that a phase transition takes place for $N, p \to \infty$. The vertical dashed line corresponds to the value of $\alpha_c(1)$ found by numerically solving $f'_\infty(\alpha) = 1$. The figure suggests that the sample complexity is linear in the system size $P = \mathcal{O}(N)$, at least for the algorithm to yield some result. This doesn't say anything about the error of the inferred couplings which behaves as $1/\sqrt{p}$ at high temperature. The above table shows some values of α_c for different β.

β	$\alpha_c(\beta)$
0.5	2.758±.002
1.0	3.116±.002
1.5	3.420±.004
2.0	3.582±.008
2.5	3.696±.016

Table 13.1: α_c for some values of β

Let us now see how algorithm 13.2.1 actually reconstructs a given the model. For the simulations of the inverse problem we first simulate a N spins system using Glauber dynamics for p samples at inverse temperature β and compute m, C and D from the data. We then run algorithm 13.2.1, along with the corresponding algorithms derived from NMF and TAP approximations, and compare the output with the original model. One obvious way to do that is to look at the scatter plots as we did for the magnetizations and correlations. In figure

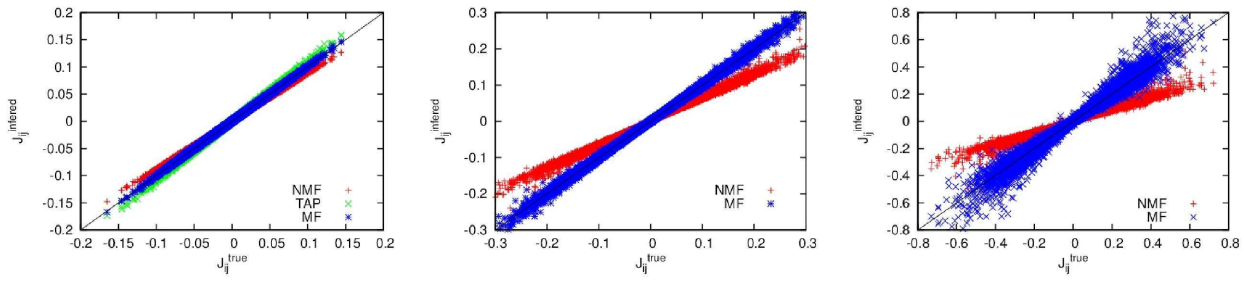


Figure 13.10: Scatter plots of inferred versus true couplings for a system of size $N = 100$ using the three mean-field algorithms. The algorithm used $p = 10^6$ samples generated at $\beta = 0.4$, 1 and 2 from left to right. Note that at high β TAP fails.

13.10 the output of this procedure is shown for a $N = 100$ system inferred from $p = 10^6$ samples generated at three different temperatures. As expected, our algorithm gives results well centered around the diagonal while the other two mean field algorithms are displaced. In fact the result of TAP doesn't even exist in the last to frames since it fails at high β (see chapter 7).

As a measure of the error, as we did for the correlations in the previous section, we take the mean squared error $\epsilon_J \equiv 1/N^2 \sum_{i,j} (\beta J_{ij}^{true} - J_{ij}^{inferred})^2$. In figure 13.11 we can see ϵ_J as a function of the inverse temperature β (left frame) and as a function of p (right frame). At high temperature the error is $1/p$. On the other hand, at low temperature finite size effects begin to appear and the error stops to improve after some value of p. Note also that the results for TAP are only present for high temperatures since it fails at lower ones.

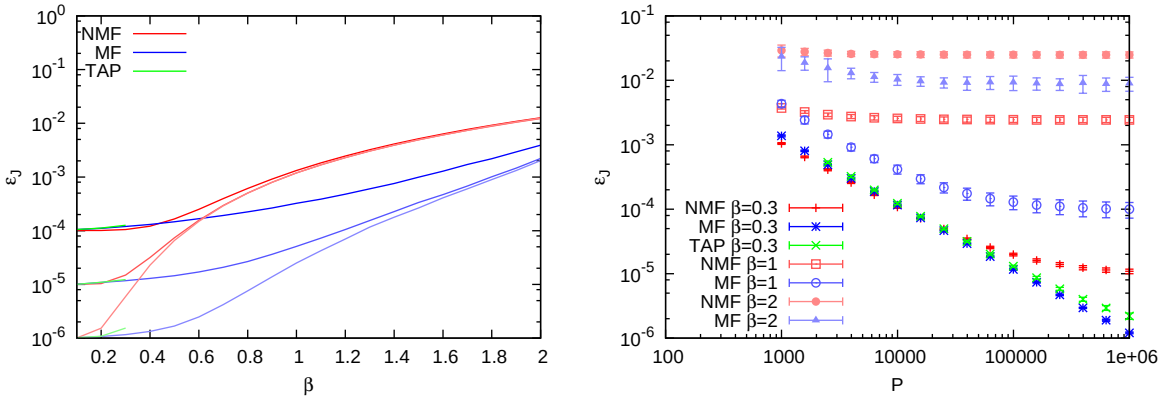


Figure 13.11: Asymmetric couplings. LEFT: ϵ_J as a function of the inverse temperature at which the samples where generated. The system has size $N = 100$ and the number of samples is 10^4, 10^5 and 10^6 from top to bottom. RIGHT: ϵ_J as a function of p for inverse temperatures $\beta = 0.3$, 1 and 2.

As we have mention earlier, the correlations predictions of our Gaussian mean field theory, in the case of symmetric couplings where effective local fields are not strictly Gaussian, are still better than that of the other mean field theories, although this is not true for the prediction of

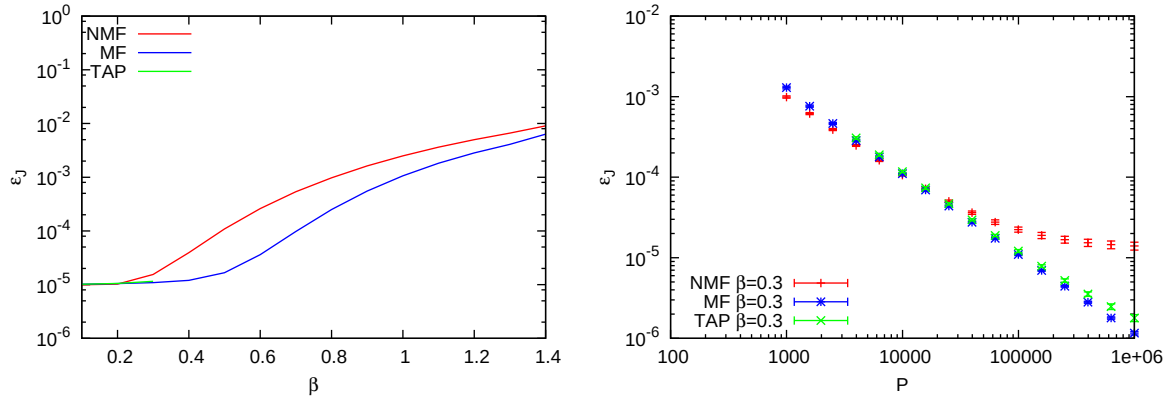


Figure 13.12: Symmetric couplings. LEFT: ϵ_J as a function of the inverse temperature at which the samples where generated. The system has size $N = 100$ and the number of samples is 10^5. RIGHT: ϵ_J as a function of p for inverse temperature $\beta = 0.3$. In both cases MF outperforms NMF and TAP

magnetizations. The correlations relations are the corner stone of every inverse problem algorithm because they characterize the structure of the interactions between spins. We therefore expect that our MF algorithm outperforms the other two mean field algorithms even in the case of symmetric couplings. This is indeed the case as can be seen in figure 13.12. The errors of MF are higher than in the asymmetric case, especially in the low temperatures, but still considerably lower than the NMF ones. It is interesting to note that TAP, even in the high temperature region, performs slightly worse than MF.

13.2.2 Non stationary case

For the non-stationary model that we used in section 13.1.4, *i.e.* a model with constant couplings but time varying local fields, like the ones in eq.(13.27), we need to modify the algorithm. Had both couplings and local fields been time varying, then the same algorithm could be used for every time step $t = 1, \ldots, t_{\max}$ separately. However, since the couplings don't change with time, we can combine the information of every time step of every sample to achieve a better precision than just applying algorithm 13.2.1.

We begin the description of the modified algorithm by taking time averages on both sides of eq.(13.32)

$$\langle D_{ij}(t) \rangle_t = \sum_k J_{ik} \langle a_i(t) C_{kj}(t) \rangle_t \tag{13.12}$$

which gives

$$J_{ij} = \sum_k \langle D_{ik}(t) \rangle_t [(K^{(i)})^{-1}] \quad , \tag{13.13}$$

where $K^{(i)}_{kj} \equiv \langle a_i(t) C_{kj}(t) \rangle_t$. The averages here are taken over $t = 1, \ldots, t_{\max}$. Based on the

above relation we derive the following algorithm

Algorithm 13.2.2: MEAN FIELD INVERSE ISING, NON-STATIONARY CASE(m, C, D)

$\overline{D} \leftarrow \langle D(t) \rangle_t$
for $i \leftarrow 1$ **to** N
{
  for $t \leftarrow 1$ **to** $t_{\max}$
    do $\hat{\Delta}(t) \leftarrow \Delta_0$
  while $\delta\Delta \neq 0$
    do {
      for $t \leftarrow 1$ **to** $t_{\max}$
        do {
          $\Delta(t) \leftarrow \hat{\Delta}(t)$
          Using $m_i(t+1)$ compute $u(t)$ by inverting eq.(13.1)
          Using $u(t)$ and $\Delta(t)$ compute $a(t)$ using eq.(13.4)
        }
      for $j \leftarrow 1$ **to** N
        do {
          for $k \leftarrow 1$ **to** N
            do $K_{kj} \leftarrow \langle a(t) C_{kj}(t) \rangle_t$
        }
      Invert K
      for $j \leftarrow 1$ **to** N
        do $J_{ij} \leftarrow \sum_k \overline{D}_{ik} [K^{-1}]_{kj}$
      for $t \leftarrow 1$ **to** $t_{\max}$
        do $\hat{\Delta}(t) \leftarrow \sum_j J_{ij}^2 (1 - m_j(t)^2)$
      $\delta\Delta \leftarrow \max_t |\Delta(t) - \hat{\Delta}(t)|$
    }
  for $t \leftarrow 1$ **to** $t_{\max}$
    do {
      $g \leftarrow \sum_j J_{ij} m_j(t)$
      $H_i(t) \leftarrow u(t) - g$
    }
}
return (J, H)

The non stationary problem is more complex and this is reflected to the fact that one has to invert a matrix for every i. Additionally, since the computation of $u(t)$, $a(t)$ and $\Delta(t)$ has to be done for every $t = 1, \ldots, t_{\max}$ separately the computational complexity will have also a factor $t_{\max}$. Thus complexity of algorithm 13.2.2 is $\mathcal{O}(t_{\max} N^4)$.

For the simulations we used a uniform external field

$$H_i(t) = H_0 \sin(\omega t) \quad , \tag{13.14}$$

with $\omega = 2\pi/100$ and $H_0 = 0.5$. In figure 13.13 we can see the predictions for $H_i(t)$ averaged over all the spins, for the MF and NMF methods (TAP fails at $\beta = 0.5$). In the left frame the couplings are asymmetric and MF is exact while in the right one the couplings are symmetric and MF is just an approximation, although a better one than NMF.

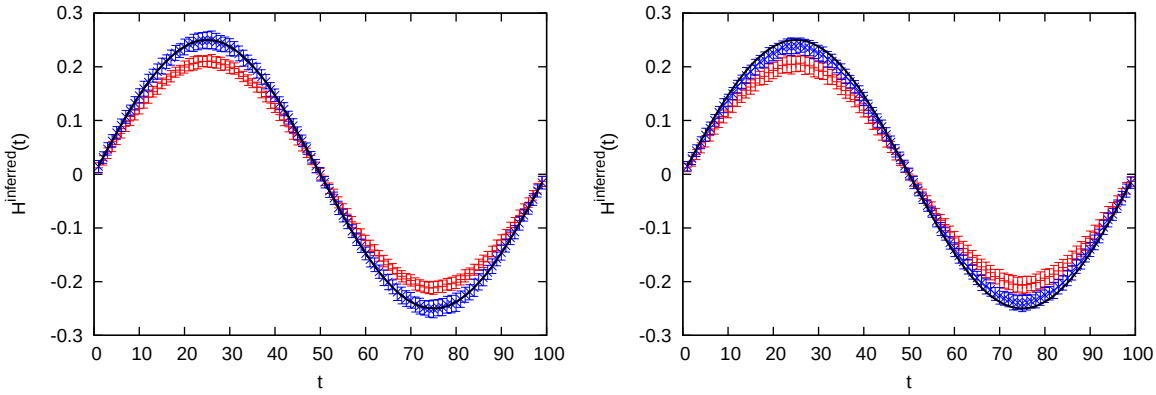


Figure 13.13: Varying external field predictions according to NMF (red) and MF (blue) methods averaged over all the spins. The black line shows the true value of the external field. LEFT: The system has asymmetric couplings, size $N = 100$ and is inferred using $p = 10^4$ samples obtained at inverse temperature $\beta = 0.5$. RIGHT: The same but with symmetric couplings.

Chapter 14

Sparse models

All the previous results, in this chapter, actually rely only on one assumption. The Gaussianity of the effective field $\sum_j J_{ij} s_j(t)$. As we have seen, when the couplings are asymmetric, this assumption is true and it can even provide some good results when they are not, as long as the temperature is not too low. In this section we will apply this idea to a different approach for solving the inverse problem, the so called Bayesian one. We have seen in chapter 5 that this approach leads to the original Boltzmann machine algorithm which is, in general, inefficient due to the intractability of the magnetizations and correlations in the direct problem. In the light of the previous results, however, it is clear that this is the case only when the couplings are symmetric. In the previous chapters we showed that there are exact and efficient ways of computing both quantities when the couplings are asymmetric. In this chapter we will show how, starting form the Bayesian formulation of the inverse problem, the Gaussianity of the effective field can be used to yield an efficient version of the Boltzmann machine algorithm. Subsequently, we will show that the two methods are equivalent in the sense that they lead to the same results. However in the Bayesian approach, there is a natural way to include the l_p-norm regularization, introduced in chapter 11, so useful in a *sparse network* context. Therefore we will focus more on sparse, rather than fully connected, systems for the simulations. As before, we divide this section into *stationary* and *non-stationary* subsections.

14.1 Stationary case

We start from the full distribution of the spin trajectories $s = \{s_i : i = 1, \dots, N\}$ where $s_i = \{s_i(t) : t = 1, \dots, p\}$

$$P(s_1, \dots, s_N) = \prod_{i=1}^{N} \prod_{t=1}^{p} \frac{e^{\beta s_i(t+1) h_i(t)}}{2\cosh(\beta h_i(t))} \prod_i P_i^0(s_i(0)) \quad , \tag{14.1}$$

where

$$h_i(t) \equiv H_i(t) + \sum_j J_{ij} s_j(t) \quad . \tag{14.2}$$

We then use Bayes' theorem to compute the likelihood of a model

$$P(\mathcal{M}|s) = \frac{P(s|\mathcal{M}) P_0(\mathcal{M})}{P(s)} \quad , \tag{14.3}$$

where $\mathcal{M}$ is a shorthand for the model parameters J and H. For the time being we will ignore the *a priori* probability over the models $P_0(\mathcal{M})$ and just attribute uniform probability over all model space. The prior of the spins is just a constant factor, absorbed in the normalization. We therefore rewrite the likelihood of a model given a set of measured spin configurations as

$$P(\mathcal{M}) \cong \prod_{i=1}^{N}\prod_{t=1}^{p} e^{s_i(t+1)h_i(t)-\log 2\cosh(h_i(t))} \prod_i P_i^0(s_i(0)) \quad , \tag{14.4}$$

where the $\cong$ symbol means equal up to a normalization. Taking then the log we have the log-likelihood

$$\mathscr{L} = \sum_{i=1}^{N}\sum_{t=1}^{p} [s_i(t+1)h_i(t) - \log 2\cosh h_i(t)] \tag{14.5}$$

by which, performing a gradient ascent, we get the following learning rules

$$\delta H_i = \varepsilon \left(\langle s_i(t+1) \rangle_t - \langle \tanh h_i(t) \rangle_t \right) \quad , \tag{14.6}$$

$$\delta J_{ij} = \varepsilon \left(\langle s_i(t+1) s_j(t) \rangle_t - \langle \tanh h_i(t) s_j(t) \rangle_t \right) \quad . \tag{14.7}$$

The idea, as in the usual Boltzmann machine, is then to substitute the first term of the left hand side of the above equations with the experimental values of the magnetizations m_i and time-delayed correlations D_{ij} and reevaluate their theoretical predictions at each step of the algorithm until they become equal. The evaluation of the magnetizations and correlations given a model is the time consuming part of the original Boltzmann machine algorithm as it is usually done by Monte Carlo simulation. However, and given that we are still interested mostly in systems with asymmetric couplings, we can replace the time averages with Gaussian integrals as we did in the previous chapter. The resulting learning rules can now be written

$$\delta H_i = \varepsilon \left(m_i - \int Dx \tanh \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) \tag{14.8}$$

$$\begin{aligned} \delta J_{ij} = \varepsilon \Bigg(& \tilde{D}_{ij} - m_j \int Dx \tanh \left[H_i + g_i + x\sqrt{\Delta_i} \right] \\ & - [JC]_{ij} \int Dx \left(1 - \tanh^2 \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) \Bigg) \quad . \end{aligned} \tag{14.9}$$

The tilde above the time-delayed correlations means that we use the non-connected ones $\tilde{D}_{ij} \equiv \langle s_i(t+1) s_j(t) \rangle_t$, as opposed to the connected correlations $D_{ij} \equiv \langle s_i(t+1) s_j(t) \rangle_t - \langle s_i(t+1) \rangle \langle s_j(t) \rangle_t$. Notice that if we manage to reach the point where $\delta H_i = 0$ we can drop the second term in eq.(14.9) and use the connected correlations instead, since that term cancels out with the difference $\tilde{D}_{ij} - D_{ij}$. This, simplified version of the equation is written

$$\delta J_{ij} = \varepsilon \left(D_{ij} - [JC]_{ij} \int Dx \left(1 - \tanh^2 \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) \right) \quad . \tag{14.10}$$

It is clearly visible that when the maximum of the log-likelihood is reached both equations (13.9) and (13.23) are verified. Thus, the procedure described above has the same fixed point as algorithm 13.2.1. The advantage, however, of the above procedure is that we can restrict the search in some subregion of model space by choosing a particular prior distribution $P_0(\mathcal{M})$ in eq.(14.3). One particularly relevant, for real world applications, class of models are sparse models. That is models where only some random subset of couplings are non zero. As we have seen in chapter 11 a natural choice for a prior, in such cases, is the distribution

$$P_0(\mathcal{M}) \cong \mathrm{e}^{-\lambda \|J\|_1} \quad . \tag{14.11}$$

This modifies the equation for the couplings update (14.10) as

$$\delta J_{ij} = \varepsilon \left(D_{ij} - [JC]_{ij} \int Dx \left(1 - \tanh^2 \left[H_i + g_i + x\sqrt{\Delta_i} \right] \right) - \lambda \, \mathrm{sign} J_{ij} \right) \quad , \tag{14.12}$$

where it is understood that $\mathrm{sign}0 = 0$.

As before, the different in-coming neighborhoods of every spin decouple and the rows of the matrix J can be inferred independently. Inspired by the algorithm 13.2.1 we propose the following one

Algorithm 14.1.1: GAUSSIAN BOLTZMANN MACHINE, STATIONARY (m, C, D, λ)

$\Xi \leftarrow JC$
for $i \leftarrow 1$ **to** N
{
  for $j \leftarrow 1$ **to** N
    do $J_{ij} \leftarrow J^0_{ij}$
  $\hat{\Delta} \leftarrow \sum_j J^2_{ij}(1 - m^2_j)$
  while $\Delta \neq \hat{\Delta}$
    do {
      $\Delta \leftarrow \hat{\Delta}$
      Using m_i compute $u \equiv H_i + g$ by inverting eq.(13.1)
      for $j \leftarrow 1$ **to** N
        do {
          $J^{\mathrm{old}} \leftarrow J_{ij}$
          $J_{ij} \leftarrow J_{ij} - \varepsilon \left(D_{ij} - \Xi_{ij} \int Dx \left(1 - \tanh^2 \left[u + x\sqrt{\Delta} \right] \right) - \lambda \, \mathrm{sign} J_{ij} \right)$
          $\hat{\Delta} \leftarrow \hat{\Delta} + (J_{ij} - J^{\mathrm{old}})(1 - m^2_j)$
          for $k \leftarrow 1$ **to** N
            do $\Xi_{ik} \leftarrow \Xi_{ik} + (J_{ij} - J^{\mathrm{old}}) C_{jk}$
        }
    }
  $g \leftarrow \sum_j J_{ij} m_j$
  $H_i \leftarrow u - g$
}
return (J, H)

It may appear as if one must multiply J and C after every J_{ij} update but in fact only a linear number of terms is affected after one J_{ij} is changed. So with the proper treatment (see the use of matrix Ξ and its update in the code above) the algorithm is still $\mathcal{O}(N^3)$. The algorithm is, nonetheless, a bit slower than algorithm 13.2.1 because of the greater number of operations and the existence of the arbitrary parameter ϵ.

One very important feature of the above algorithm is that it doesn't need to invert the matrix C. This means that results can be obtained even if $p < N$. In the fully connected case this wouldn't make much sense since one would estimate $\mathcal{O}(1/\sqrt{N})$ couplings making $\mathcal{O}(1/\sqrt{p})$ errors, but when a system is sparse less information is needed to determine its structure. It has been shown in numerical simulations that the above algorithm is indeed capable of recovering much of the system structure using $p < N$. Even in the $p \gtrsim N$ regime, where algorithm 13.2.1 doesn't fail, algorithm 14.1.1 performs much better. What actually happens is that the l_1-norm regularization ensures, as in the algorithm of chapter 11, that a fraction of all couplings will be set to zero. The minimization of the log-likelihood, on the other hand, ensures that we find the best combination of non-zero and zero couplings to fit the data. Then, since the data are used for the estimation of the non-zero couplings only, we have a higher amount of information for the estimation of each coupling. Of course, in the $p \to \infty$ limit, both algorithms yield the same results.

The model used in the simulations for this section is an Ising model defined on a directed Erdős-Rényi graph. We define the *average in-degree* d_{in} as the average number of in-coming couplings per spin. Then the distribution of the couplings reads

$$P(J) = \frac{N - d_{\text{in}}}{N}\delta(J) + \frac{d_{\text{in}}}{N}\rho(J) \quad . \tag{14.13}$$

For the distribution of the non-zero couplings $\rho(J)$ we used a Gaussian with a disconnected support such that there is a minimum coupling strength $J_{\min} \equiv \min_{(i,j)}\{|J_{ij}| \neq 0\}$

$$\rho(J) = \theta(J - J_{\min})\sqrt{\frac{d_{\text{in}}}{2\pi}}e^{-d_{\text{in}}(J-J_{\min})^2/2} + \theta(-J - J_{\min})\sqrt{\frac{d_{\text{in}}}{2\pi}}e^{-d_{\text{in}}(J+J_{\min})^2/2} \quad . \tag{14.14}$$

In the simulations we used $d_{\text{in}} = 10$ and $J_{\min} = 0.2$. The motivation for the above choice is that, as it has been shown in [RavikumarWL 10] and [BentoM 09], if one lets $J_{\min} \equiv \min_{(i,j)}\{|J_{ij}| \neq 0\}$ come arbitrarily close to 0, then one must have $p \to \infty$ to correctly estimate the structure.

What we are interested in, in the simulations, is to correctly estimate the structure of the network, *i.e.* correctly identify the zero and non-zero couplings. We will therefore use again the error measure introduced in chapter 12. We rewrite the deffinition of these quantities here: The *True Positive Rate* (TPR)

$$\text{TPR} \equiv \frac{\text{TP}}{\text{TP} + \text{FN}} \quad , \tag{14.15}$$

and the *True Negative Rate* (TNR)

$$\text{TNR} \equiv \frac{\text{TN}}{\text{FP} + \text{TN}} \quad , \tag{14.16}$$

where

$$\begin{aligned} \text{TP} &\equiv \text{Number of non-zero couplings correctly identified} \\ \text{TN} &\equiv \text{Number of zero couplings correctly identified} \\ \text{FP} &\equiv \text{Number of non-zero couplings identified as zero} \\ \text{FN} &\equiv \text{Number of zero couplings identified as non-zero} \quad . \end{aligned}$$

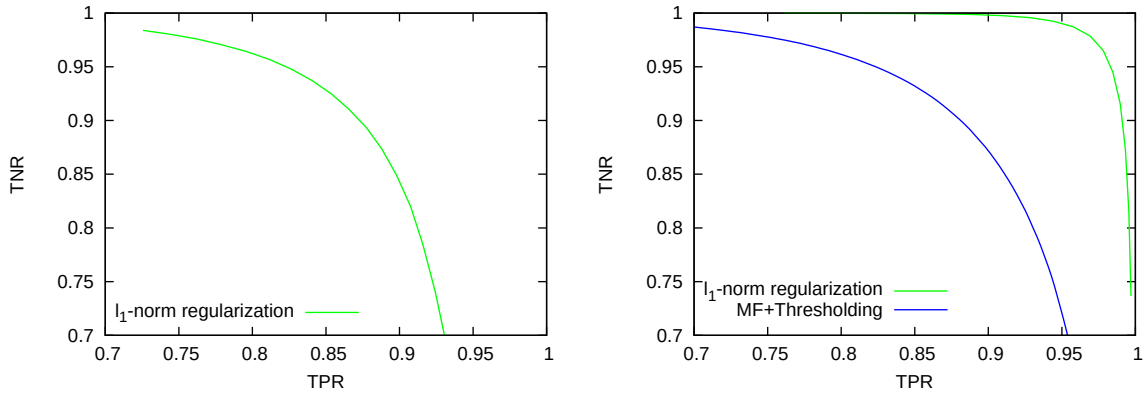


Figure 14.1: ROC curves for algorithm 14.1.1 and algorithm 13.2.1 with thresholding. LEFT: The system has size $N = 1000$ and average in-degree $d_{\rm in} = 10$ and is inferred from $p = 500$ samples obtained at inverse temperature $\beta = 0.5$. The results are averaged over 20 realizations of the system. Algorithm 13.2.1 is not shown as it fails to provide results for $p < N$. RIGHT: The same system but this time inferred using $p = 1500$.

Obviously, perfect inference corresponds to TPR=1 and TNR=1. In figure 14.1, we used for each system different values of the parameter λ and plotted the parametric ROC curve TPR(λ), TPR(λ). In the left frame we see the ROC curve for a system of size $N = 1000$ inferred using $p = 500$ samples with $\lambda \in [0.02, 0.1]$. We see that even for such a small number of samples compared to the system size, there is a value of λ for which almost 90% of the system structure is inferred correctly. In the right frame we used the same system size but this time we used $p = 1500$ samples for the inference. For comparison, we have included the results of algorithm 13.2.1 with additionally truncating the smaller, in absolute value, couplings under some threshold $J_{\rm thres}$. By varying $J_{\rm thres}$ we also obtain a ROC curve. We clearly see that there is no value of $J_{\rm thres}$ that achieves an error as small as the one that algorithm 14.1.1 achieves for the optimal value of λ.

14.2 Non stationary case

For the non stationary case we use the same model as in section 13.2.2. The log-likelihood is now

$$\mathscr{L} = \sum_{i=1}^{N} \sum_{t=1}^{t_{\max}} \sum_{r=1}^{p} [s_i^r(t+1) h_i^r(t) - \log 2\cosh h_i^r(t)] \quad , \tag{14.1}$$

where $h_i^r(t) = H_i(t) + \sum_j J_{ij} s_j^r(t)$. The corresponding learning rules are

$$\delta H_i(t) = \varepsilon \left(m_i(t+1) - \int Dx \tanh \left(H_i(t) + g_i(t) + x\sqrt{\Delta_i(t)} \right) \right) \quad , \tag{14.2}$$

$$\delta J_{ij} = \varepsilon \left(\langle D_{ij}(t) \rangle_t - \left\langle [JC(t)]_{ij} \int Dx \left(1 - \tanh^2 \left(H_i(t) + g_i(t) + x\sqrt{\Delta_i(t)} \right) \right) \right\rangle_t \right) \quad , \tag{14.3}$$

where the averages are taken as usual over all time steps $t = 1, \ldots, t_{\max}$.

The generalization of algorithm 14.1.1 for the non stationary case is straightforward and yields the following procedure

Algorithm 14.2.1: GAUSSIAN BOLTZMANN MACHINE,NON STATIONARY(m, C, D, λ)

$\overline{D} \leftarrow \langle D(t) \rangle_t$
for $t \leftarrow 1$ **to** $t_{\max}$
  do $\Xi(t) \leftarrow JC(t)$
for $i \leftarrow 1$ **to** N
{ **for** $j \leftarrow 1$ **to** N
    do $J_{ij} \leftarrow J_{ij}^0$
  for $t \leftarrow 1$ **to** $t_{\max}$
    do $\hat{\Delta}(t) \leftarrow \sum_j J_{ij}^2 (1 - m_j(t)^2)$
  while $\delta\Delta \neq 0$
  { **for** $t \leftarrow 1$ **to** $t_{\max}$
      do { $\Delta(t) \leftarrow \hat{\Delta}(t)$
             Using $m_i(t)$ compute $u(t) \equiv H_i(t) + g(t)$ by inverting eq.(13.1)
             Using $u(t)$ and $\Delta(t)$ compute $a(t)$ using eq.(13.4)
    for $j \leftarrow 1$ **to** N
      do { $J^{\text{old}} \leftarrow J_{ij}$
             $J_{ij} \leftarrow J_{ij} - \varepsilon \left(\overline{D}_{ij} - \langle a(t) \Xi_{ij}(t) \rangle_t - \lambda \operatorname{sign} J_{ij} \right)$
             for $t \leftarrow 1$ **to** $t_{\max}$
               do { $\hat{\Delta}(t) \leftarrow \hat{\Delta}(t) + (J_{ij} - J^{\text{old}})(1 - m_j(t)^2)$
                      for $k \leftarrow 1$ **to** N
                        do $\Xi_{ik}(t) \leftarrow \Xi_{ik}(t) + (J_{ij} - J^{\text{old}}) C_{jk}(t)$
  for $t \leftarrow 1$ **to** $t_{\max}$
    do { $g(t) \leftarrow \sum_j J_{ij} m_j(t)$
           $H_i(t) \leftarrow u(t) - g(t)$
return (J, H)

Chapter 15

Open questions

15.1 Application to neural data

It is clear from all numerical evidence presented in the previous chapters that the assumption of a Gaussian effective field allows to solve the inverse Ising problem both efficiently and exactly. An important question then arises: what happens when the data are not generated by an Ising model? We have argued in the first chapter that the Ising model is, from an information theoretical point of view, a good candidate for modeling biological systems while being at the same time simple enough to allow important theoretical advancements. It is clear, however, that it is a serious abstraction from the real world. It is natural to expect that data obtained from a real biological system will not behave as the synthetic data that we used in simulations throughout this thesis.

Recently our method has been applied to real spike trains obtained experimentally by J. Tyrcha et al [TyrchaRMH 12]. The data set, provided by Michael Berry of Princeton University, consist of the spike trains of 40 neurons recorded from the retina of a salamander under visual stimulation. The activity of each neuron was recorder during 120 repetitions of a 26.5 sec movie clip and was then divided in time-bins of length 20 ms. Then three algorithms were used to infer the couplings of an Ising model agreeing with the data: the non-stationary version of our algorithm [i] 13.2.2, a non-stationary version of the naive mean field algorithm (see [RoudiH 11a, RoudiH 11b] for details) and also a version of the Boltzmann machine learning algorithm found in chapter 5 also adapted to non-stationary systems (see [TyrchaRMH 12]). Following the authors we will refer to the latter algorithm simply as the exact algorithm.

The results confirm our hypothesis that the MF algorithm is better than the simpler NMF. However, the inferred couplings of MF and NMF are way closer to each other than to those of the exact algorithm. This is not what we expected based on results on synthetic data. As a measure of success of the three algorithms the authors compute the log-likelihoods on the data (with an Akaike penalty to discourage over-fitting). The values are -0.062748, -0.062872, and -0.062823 for the exact algorithm, NMF and MF respectively. It is important to stress that the exact algorithm takes many hours to yield results as opposed to a few minutes for the mean field ones. So, given the close values of the log-likelihood found by the three algorithms,

i. The external stimulus, in this case the movie clip, is evolving in time. To infer correctly the couplings one must account for time-dependent local fields which is why the non-stationary version is needed.

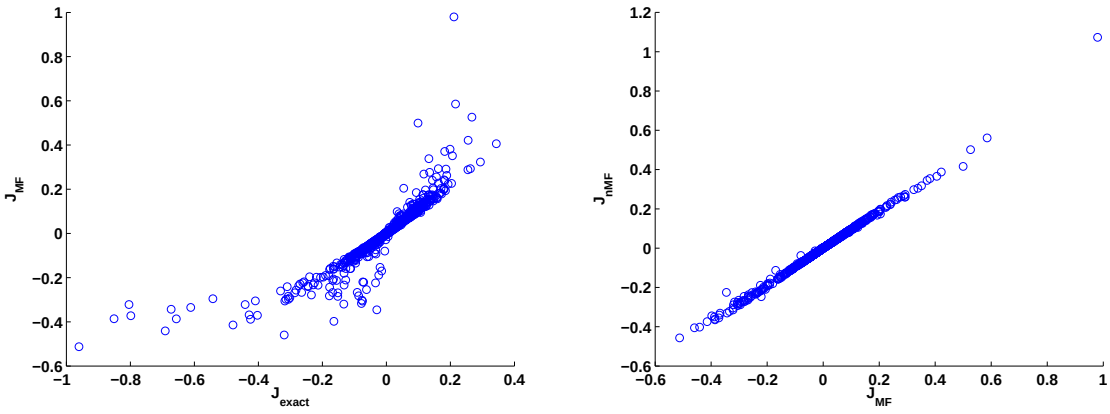


Figure 15.1: Taken from [TyrchaRMH 12]. LEFT: Scatter-plot of the inferred couplings from the retinal data using MF versus the ones obtained from the exact one. RIGHT: The same but this time comparing the MF and NMF algorithms.

the exact one doesn't offer an important advantage. The authors also provide scatter-plots of the inferred couplings, reprinted in fig.15.1. We see that for the majority of the couplings who are close to zero all algorithms agree more or less. For those with stronger synaptic strengths, which are the most important, the qualitative (inhibitory or excitatory) predictions agree but the values deviate from the exact algorithm to the others.

In the right frame we see that NMF and MF are very close, which is strange given the typical values of the couplings found. Indeed, looking at fig.13.11 we see that there are values of β where NMF and MF agree but they correspond to much weaker synaptic strengths than the ones found in the retinal data.

This leaves an important open question. What feature of the retinal data is causing the disagreement between the exact and MF algorithms. In the synthetic data generated from an Ising model there was no such disagreement. One obvious cause of the problem might be that the effective fields are not Gaussian distributed for some reason. Further investigation is needed in order to answer this and similar questions.

15.2 Systems with hidden variables

An other important question concerns systems with hidden variables. In the analysis of all inverse Ising algorithms studied in this thesis we always silently assumed that we have access to all the variables of our system. In practice, however, one often observes only a fraction of the system. For instance, in the previous example of the salamander retina, only 40 neurons where observed which is a tiny fraction of the total number of neurons found in the retina. Yet the algorithms where used as such, without accounting for the interaction of those 40 neurons with other, non observed, parts of the system. Let's see what could go wrong in such a case for our MF algorithm.

We rewrite the basic relation between model parameters and the measured quantities

$$D = A\,J\,C \tag{15.1}$$

$$A_{ij} = \delta_{ij}\beta \int Dx \left[1 - \tanh^2 \beta \left(H_i + g_i + x\sqrt{\Delta_i}\right)\right] , \tag{15.2}$$

where $g_i = \sum_j J_{ij} s_j$ and $\Delta_i = \sum_j J_{ij}^2 (1 - m_j^2)$. When the above equations are applied to a system with hidden variables there are two features that might lead to wrong results. Let's examine them.

The first concerns the equation for the matrix A. It is clear that having hidden variables will be reflected in the values of g_i and Δ_i. We introduce the notation

$$g_i = g_i^{\mathcal{O}} + g_i^{\mathcal{N}} \text{ and } \Delta_i = \Delta_i^{\mathcal{O}} + \Delta_i^{\mathcal{N}} , \tag{15.3}$$

where the superscripts $\mathcal{O}$ and $\mathcal{N}$ stand for the contributions of the observed and non-observed parts of the system. When performing inference the term $g_i^{\mathcal{N}}$ will be absorbed in the inferred local field H_i and thus we expect that there is no way of inferring the correct values of the local fields. However, there might be a possibility to say something about the couplings. In the thermodynamic limit, and given that the couplings of the whole system follow the same distribution, the parameter Δ_i becomes an extensive quantity and independent of the subscript i. We can thus estimate the correct value of the Δ's statistically if we have a prior knowledge on the size of the part that is not observed. For example, if we observe $N^{\mathcal{O}}$ variables and we known that we leave out of the observation $N^{\mathcal{N}}$ others, we can compute the correct value of Δ as

$$\Delta_i = \frac{N^{\mathcal{O}} + N^{\mathcal{N}}}{N^{\mathcal{O}}} \Delta_i^{\mathcal{O}} . \tag{15.4}$$

The algorithm 13.2.1 can be easily modified by replacing the line concerning the Δ update by $\hat{\Delta} \leftarrow \frac{N^{\mathcal{O}}+N^{\mathcal{N}}}{N^{\mathcal{O}}} \frac{\gamma}{a^2}$.

The second problem concerns the first equation (15.1). Again dividing the matrices in four blocks based on the division between observed and non-observed variables we have

$$D^{\mathcal{O}\mathcal{O}} = A^{\mathcal{O}\mathcal{O}} J^{\mathcal{O}\mathcal{O}} C^{\mathcal{O}\mathcal{O}} + A^{\mathcal{O}\mathcal{O}} J^{\mathcal{O}\mathcal{N}} C^{\mathcal{N}\mathcal{O}} . \tag{15.5}$$

It is clear that naively applying our algorithm in partially observed data will lead to errors because we completely omit the second term of the above equation since we don't have access to $C^{\mathcal{N}\mathcal{O}}$. It is worth noting that from numerical evidence it seems that at high temperatures the diagonal elements of C become so important that discarding the second term of the above equation doesn't lead to important errors. This shows, however, a regime where the spins are quite independent and where even simple algorithms, such as NMF, could provide satisfactory results.

There might be a way to compensate the lack of information due to the partial knowledge of the correlation matrix C. It is important to state that what follows is just a suggestion for future work and that we where not able to numerically verify our claims. Some of the missing information might come from looking at k-step time-delayed correlation functions. Consider

for example the matrix defined as $E_{ij} \equiv \langle \delta s_i(t+2) \delta s_j(t) \rangle$. It can be shown that a relation similar to eq.(15.1) holds

$$E = A\ J\ D\ . \tag{15.6}$$

Decomposing the above relation in terms of observed and non-observed parts we have

$$\begin{aligned} E^{\mathcal{OO}} &= A\ J^{\mathcal{OO}}\ D^{\mathcal{OO}} + A\ J^{\mathcal{ON}}\ D^{\mathcal{NO}} && (15.7)\\ &= \underbrace{A\ J^{\mathcal{OO}}\ A\ J^{\mathcal{OO}}\ C^{\mathcal{OO}} + A\ J^{\mathcal{OO}}\ A\ J^{\mathcal{ON}}\ C^{\mathcal{NO}}}_{A\ J^{\mathcal{OO}}\ D^{\mathcal{OO}}} && (15.8)\\ &\qquad + \underbrace{A\ J^{\mathcal{ON}}\ A\ J^{\mathcal{NO}}\ C^{\mathcal{OO}} + A\ J^{\mathcal{ON}}\ A\ J^{\mathcal{NN}}\ C^{\mathcal{NO}}}_{A\ J^{\mathcal{ON}}\ D^{\mathcal{NO}}} \end{aligned}$$

Interestingly we see that the second term, although containing the non-measurable correlations $C^{\mathcal{NO}}$, can be expressed, together with the first term, in terms of a quantity that is accessible namely $D^{\mathcal{OO}}$. The fourth term is still not accessible and must be discarded. However since, as we have supposed, all couplings follow the same distribution, it can be argued that we have gained information compared with discarding the second term of eq.(15.5). Consider for instance that we are observing half of the system, *i.e.* $N^{\mathcal{O}} = N^{\mathcal{N}}$, at low temperatures where all parts of the correlation matrix C are equally important. Then discarding the last term of eq.(15.8) amounts in discarding "one fourth" [ii] of the total information needed to correctly infer the model, as opposed to discarding "half" of the information when one naively uses eq.(15.1).

Let's rewrite eq.(15.8) using a lighter notation. We omit the superscript $\mathcal{OO}$ and use $J' = J^{\mathcal{ON}}\ A\ J^{\mathcal{NO}}$.

$$E = A\ J\ D + A\ J'\ C \tag{15.9}$$

The above relation cannot be solved for J and J' obviously so we need to find a similar relation. The answer is given by the correlations between spins of even greater temporal distance. The above reasoning can be repeated for the 3-step time-delayed correlations $F_{ij} \equiv \langle \delta s_i(t+3) \delta s_j(t) \rangle$ with the result

$$F = A\ J\ E + A\ J'\ D \tag{15.10}$$

The system of the two equations (15.9,15.10) can be solved and yields

$$J = A^{-1} \left[F - E\ C^{-1}\ D \right] \left[E - D\ C^{-1}\ D \right]^{-1}\ . \tag{15.11}$$

The above equation has the same structure with eq.(13.2) and so we can expect that it can be solved iteratively by a similar procedure in algorithm 13.2.1. Unfortunately it is not so simple. Problems arise because the last factor of eq.(15.11) is not invertible and alternative ways must be found.

One might expect that the above argument, using k-step correlations to obtain information about the hidden part of the system, might eventually lead to a method capable of overcoming the difficulties attached to those cases.

ii. The word "information" is of course used informally here.

15.3 Perspectives

The main idea of this thesis can be condensed in one phrase: whenever a system has asymmetric interactions thermal averages can be replaced with averages over Gaussian variables which leads to both exact and efficient algorithms for direct and inverse inference. This should have a positive impact on the ability of future researchers, especially those working on systems biology, to infer structural information about systems of their interest based on observed data. We have already seen an implementation of our method to a real world situation with limited success.

This work should also have an impact on the design of future algorithms aiming to solve similar inverse problems. We have mentioned in the introduction that the asymmetry of the interactions is widely found in the biological world. There is thus a great number of situations where the main idea of this thesis could be used. Different situations might need the design of new algorithms. One example is the gene regulation networks, described in the introduction, where interactions are not only found between pairs of genes but a more complex image of three or more genes interacting in a combinatorial way might be more accurate. In this case the simple pairwise Ising model must be extended to include higher order interactions. The fact is that there is no fundamental reason way the Gaussian assumption should only work for binary, pairwise models. Thus, we have reasons to hope that the main idea of this work will open new directions for future research and will lead to important results in the study of those systems.

Part IV

Reprints of publications

Exact mean field inference in asymmetric kinetic Ising systems

M. Mézard and J. Sakellariou
Laboratoire de Physique Théorique et Modèles Statistiques,
CNRS and Université Paris-Sud, Bât 100, 91405 Orsay Cedex, France
(Dated: May 19, 2011)

We develop an elementary mean field approach for fully asymmetric kinetic Ising models, which can be applied to a single instance of the problem. In the case of the asymmetric SK model this method gives the exact values of the local magnetizations and the exact relation between equal-time and time-delayed correlations. It can also be used to solve efficiently the inverse problem, i.e. determine the couplings and local fields from a set of patterns, also in cases where the fields and couplings are time-dependent. This approach generalizes some recent attempts to solve this dynamical inference problem, which were valid in the limit of weak coupling. It provides the exact solution to the problem also in strongly coupled problems. This mean field inference can also be used as an efficient approximate method to infer the couplings and fields in problems which are not infinite range, for instance in diluted asymmetric spin glasses.

Inference problems are as old as scientific modelling: given data, how can we reconstruct a model which accounts for it, and find the parameters of the model? This is particularly difficult when data is obtained from networks of many interacting components. The fast development of high-throughput technologies in various fields of biology, ranging from gene regulation to protein interaction and neural activity, is generating a lot of data, which is challenging our ability to infer the structure and parameters of the underlying networks.

This 'network reconstruction' problem is typically an inverse problem which has motivated a lot of activity in machine learning and in statistical physics[1–4, 9–16, 18–21, 23, 26, 28, 29]. Until recently the main efforts have been dedicated to reconstructing equilibrium Boltzmann-Gibbs distributions. In the so-called inverse Ising model, one typically assumes to have data in the form of some configurations, which we shall call 'patterns', of a N-spin Ising system drawn from the Boltzmann-Gibbs distribution with an energy function including one-body (local magnetic fields) and two-body (exchange couplings) terms. The problem is to reconstruct the local fields and the exchange couplings (collectively denoted below as 'couplings') from the data. This problem has been actively studied in recent years, in particular in the context of neural network inference based on multielectrode recordings in retinas [4, 22, 25]. The standard solution of this problem, known as the Boltzmann machine, computes, for some given couplings, the local magnetizations and the two-spin correlations, and compares them to the empirical estimates of magnetizations and correlations found from the patterns[1, 10]. The couplings are then iteratively adjusted in order to decrease the distance between the empirical magnetizations/correlations and the ones computed from the model. A Bayesian formulation shows that the problem of finding the couplings is actually convex, so that this iterative procedure is guaranteed to converge to the correct couplings, provided that the number of patterns is large enough to allow for a good estimate of correlations. The drawback of this method is that the reliable computation of the magnetizations/correlations, given the couplings, which is done using a Monte Carlo procedure, is extremely time consuming. Therefore this exact approach is limited to systems with a small number of spins. Most of the recent works on this issue have developed approximate methods to infer the couplings. Among the most studied approaches are the naive mean field method [9, 13, 26], the TAP approach [14, 21, 29], a method based on a small magnetization expansion [23], and a message-passing method called susceptibility propagation[11, 15, 16]. Another approach which has been developed is that of linear relaxation of the inference problem[18]. The inverse-Potts problem is a version of this same problem, with variables taking q possible states. The case $q = 20$ is relevant for inferring interaction in protein pairs from data on co-evolution of these pairs, and its solution by susceptibility propagation has given an accurate prediction of inter-protein residue contacts[28]. Another case which has received some attention is the problem of reconstruction in Boolean networks (see e.g.[3] and references therein).

However, in many applications to biological systems, in particular the ones concerning neural activity and gene expression network, the assumption that the patterns are generated by an underlying equilibrium Boltzmann-Gibbs measure is not well founded. Couplings are typically asymmetric, and they may vary in time, so there is no equilibrium measure. This has prompted the recent study of inference in purely kinetic models without an equilibrium measure [4, 9, 21, 29]. A benchmark on this dynamic inference problem is the inverse asymmetric kinetic Ising model. The framework is the same as the equilibrium one: one tries to infer the parameters of the dynamical evolution equation of an Ising spin systems, given a set of patterns generated by this evolution. The recent works [9, 20, 21, 29] have studied the performance of two mean-field methods on this problem, the naive mean field (nMF) and a weak-coupling expansion which they denote as TAP method. They have shown that, in the case of the fully asymmetric infinite range spin glass problem, the inference problem can be solved by these methods in the case where the spins are weakly coupled. In the present work we present a (non-naive!) mean field approach which solves the problem at all values of the couplings (and reduces to their TAP approach at weak coupling).

The kinetic Ising model which we shall study is the same as the one of [21]. N Ising spins s_i evolve in discrete time, with a synchronous parallel dynamics. Given the configuration of spins at time $t-1$, $s(t-1) = \{s_1(t-1), \dots, s_N(t-1)\}$, the spins $s_i(t)$ are independent random variables drawn from the distribution:

$$P(s(t)|s(t-1)) = \prod_{i=1}^{N} \frac{1}{2\cosh(\beta h_i(t))} e^{\beta s_i(t) h_i(t)} \tag{1}$$

where

$$h_i(t) = H_i(t-1) + \sum_j J_{ij}(t-1) s_j(t-1) \tag{2}$$

Note that both the local external fields $H_i(t)$ and the exchange couplings $J_{ij}(t)$ may depend on time. Here we are interested in a fully asymmetric model. We generate the J_{ij} by an asymmetric version [5, 8, 17] of the infinite - range Sherrington-Kirkpatrick spin glass model [24], in which for each directed edge (ij) the coupling is a gaussian random variable with variance $1/N$. Notice that J_{ij} and J_{ji} are independent random variables. We do not include self-interactions (we take $J_{ii} = 0$), although this could be done without changing the results. As initial conditions we

take $s_i(t=0) = \pm 1$ with probability 1/2. Our method also applies to the case of asynchronous dynamics, studied in [29] with the TAP approach, but to keep the presentation simple we shall study only the case of the synchronous parallel dynamics in this letter.

We first derive the mean-field equations for the magnetizations $m_i(t) = \langle s_i(t) \rangle$. Because the couplings are asymmetric, $\sum_j J_{ij} J_{ji} = O(1/\sqrt{N})$, therefore the Onsager reaction term is not present in this problem. This makes the derivation of our equations, which we shall denote just 'mean-field' equations, particularly easy. The approximate equations used in [21, 29], originally derived in [13], have been obtained by a second order expansion in the couplings. When this expansion is applied to the symmetric problem it gives back the TAP equations [27] with their Onsager reaction term. In the present case of asymmetric coupling, it keeps the correction of order $\sum_j J_{ij} J_{ij}$. We shall keep for these second-order-expanded equations the name 'TAP'-equations, as used by [13, 21, 29].

The local field on spin i due to the other spins, $\sum_j J_{ij}(t-1) s_j(t-1)$, is the sum of a large number of terms. Therefore it has a gaussian distribution with mean

$$g_i(t-1) = \sum_j J_{ij} m_j(t-1) \tag{3}$$

and variance

$$\Delta_i(t-1) = \sum_j J_{ij}^2 (1 - m_j(t-1)^2) \tag{4}$$

(in order to derive this last formula, one must use the fact that the typical connected correlation $\langle s_j s_k \rangle - m_j m_k$ is typically of order $1/\sqrt{N}$; this will be checked self-consistently below). Using this property and the definition of the dynamics (1), one obtains the magnetization of spin i at time t:

$$m_i(t) = \int Dx \, \tanh \left[\beta \left(H_i(t-1) + g_i(t-1) + x \sqrt{\Delta_i(t-1)} \right) \right] \ , \tag{5}$$

where $Dx = \frac{dx}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$ is the probability density for a Gaussian variable x with zero mean and variance unity.

Equations (3,4,5) are our mean field (MF) equations for this problem, valid on a given instance. Similar dynamical equations have been obtained in the study of the sample-averaged order parameter in asymmetric neural networks[6, 7] and spin glasses[5]. They can be iterated starting from some initial condition (in our case $m_i(0) = 0$) in order to get all the magnetizations $m_i(t)$ at any time. They rely only on the central limit theorem and they are exact in the large N limit, for any set of couplings and external fields, even if they are time-dependent. These differ from the 'TAP' equations of [13, 14, 21, 29] which can be written in our notation:

$$m_i(t) = \tanh \left[\beta H_i(t-1) + \beta g_i(t-1) - m_i(t) \beta^2 \Delta_i(t-1) \right] \ , \tag{6}$$

and from the naive mean field (nMF) equations:

$$m_i(t) = \tanh \left[\beta \left(H_i(t-1) + g_i(t-1) \right) \right] \ . \tag{7}$$

The nMF equations and the 'TAP' equations actually give the same result as our exact MF equations, when expanded in powers of Δ_i, respectively to order Δ_i^0 and Δ_i^1, but they differ at order Δ_i^2. The fact that 'TAP' equations agree with the exact MF to second order in a weak coupling expansion is consistent with their derivation through second order Plefka-type expansion[14]. The correctness of the MF equations (5,3,4) can be easily checked numerically as shown in the left panels of Fig.1.

We now turn to the computation of correlations. We shall establish the mean field relation between the time-delayed and the equal-time correlation matrices:

$$D_{ij}(t) \equiv \langle \delta s_i(t+1) \delta s_j(t) \rangle \tag{8}$$

$$C_{ij}(t) \equiv \langle \delta s_i(t) \delta s_j(t) \rangle \ , \tag{9}$$

where we define $\delta s_i(t)$ as the fluctuation of the magnetization: $\delta s_i(t) = s_i(t) - \langle s_i(t) \rangle$.

We start by writing $\sum_j J_{ij}(t) s_j(t) = g_i(t) + \delta g_i(t)$, where $\delta g_i(t)$ is gaussian distributed with mean 0 and variance $\Delta_i(t)$. Now, by definition of D_{ij} we have

$$D_{ij}(t) = \langle s_j(t) \tanh \left[\beta \left(H_i(t) + g_i(t) + \delta g_i(t) \right) \right] \rangle - \langle s_j(t) \rangle \langle \tanh \left[\beta \left(H_i(t) + g_i(t) + \delta g_i(t) \right) \right] \rangle \tag{10}$$

Hereafter in order to keep notations simple in the derivation of the relation between $D(t)$ and $C(t)$ we work at a fixed time t and we thus drop the explicit time indices: all time indices in this derivation are equal to t (e.g. $J_{ij} = J_{ij}(t)$, $\delta s_i = \delta s_i(t)$, $g_i = g_i(t)$ etc.) We get:

$$\begin{aligned} \sum_k J_{jk} D_{ik} &= \langle (g_j + \delta g_j) \tanh\left[\beta\left(H_i + g_i + \delta g_i\right)\right]\rangle - g_j \langle \tanh\left[\beta\left(H_i + g_i + \delta g_i\right)\right]\rangle \\ &= \langle \delta g_j \tanh\left[\beta\left(H_i + g_i + \delta g_i\right)\right]\rangle \end{aligned} \tag{11}$$

In order to evaluate the average we need the joint distribution of δg_i and δg_j. The crucial point to keep in mind is that, as the couplings are of order $1/\sqrt{N}$, each matrix element of C and D is also small, of order $1/\sqrt{N}$. Their covariance is therefore small:

$$\langle \delta g_i \delta g_j \rangle = \langle \sum_k J_{ik}\left(s_k - \langle s_k\rangle\right) \sum_l J_{jl}\left(s_l - \langle s_l\rangle\right)\rangle \tag{12}$$

$$= \sum_{k,l} J_{ik} J_{jl} C_{kl} = \left(JCJ^T\right)_{ij} \equiv \varepsilon \ , \tag{13}$$

where ε is typically of order $1/\sqrt{N}$. So the joint distribution of $x = \delta g_i$ and $y = \delta g_j$ takes the form, in the large N limit (omitting terms of order ε^2):

$$P(x,y) = \frac{1}{2\pi\sqrt{\Delta_i \Delta_j}} \exp\left(-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j} + \varepsilon \frac{xy}{\Delta_i \Delta_j}\right) \tag{14}$$

Using the small ε expansion of eq. (14) we can rewrite eq. (11) as

$$\sum_k J_{jk} D_{ik} = \frac{\varepsilon}{\Delta_i \Delta_j} \int \frac{dx}{\sqrt{2\pi\Delta_i}} \frac{dy}{\sqrt{2\pi\Delta_j}} \mathrm{e}^{-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j}} x y^2 \tanh\left[\beta\left(H_i + g_i + x\right)\right] \tag{15}$$

$$= \varepsilon\beta \int \frac{dx}{\sqrt{2\pi\Delta_i}} \exp^{-\frac{x^2}{2\Delta_i}} \left(1 - \tanh^2\left[\beta\left(H_i + g_i + x\right)\right]\right) \ . \tag{16}$$

Combining eq. (12) and eq. (15) we get:

$$\left(DJ^T\right)_{ij} = \left(JCJ^T\right)_{ij} \beta \int \frac{dx}{\sqrt{2\pi\Delta_i}} \mathrm{e}^{-\frac{x^2}{2\Delta_i}} \left(1 - \tanh^2 \beta\left(H_i + g_i + x\right)\right) \ , \tag{17}$$

which gives the explicit mean-field relation between C and D. Putting back the time indices, we obtain the final result in matrix form:

$$D(t) = A(t)\ J(t)\ C(t) \ , \tag{18}$$

where $A(t)$ is a diagonal matrix: $A_{ij}(t) = a_i(t)\delta_{ij}$, with:

$$a_i(t) = \beta \int Dx \ \left[1 - \tanh^2 \beta\left(H_i(t) + g_i(t) + x\sqrt{\Delta_i(t)}\right)\right] \ . \tag{19}$$

The final result (18) takes exactly the same form as the one found with the naive mean field equation and with the 'TAP' approach. The predictions of all three methods, nMF, 'TAP' and our MF method is always $D(t) = A(t)\ J(t)\ C(t)$, with a diagonal matrix $A(t)$ which differs in each case. As shown in [21], the nMF approximation gives:

$$a_i^{nMF}(t) = \beta\left[1 - m_i(t+1)^2\right] \ , \tag{20}$$

the 'TAP' approximation gives:

$$a_i^{TAP}(t+1) = \beta\left[1 - m_i(t+1)^2\right] \left[1 - (1 - m_i(t+1)^2)\beta^2 \sum_k J_{ik}^2 (1 - m_k(t)^2)\right] \tag{21}$$

and our mean field prediction is the one given in (19).

We claim that, as in the case of the magnetizations, our mean field equations connecting D to C are exact in the asymmetric SK model, in the large N limit. This statement can be checked numerically by comparing $(AJC)_{ij}$ with the experimental values of D_{ij} found by monte carlo simulations, as shown in Fig.1.

These results on the mean field relation between C and D can be used for the inverse problem. Given P 'patterns', which are time sequences of length t generated from the distribution (1), one can estimate for each $\tau = 1, \ldots, t$, the magnetizations $m_i(\tau)$, the equal time correlations $C_{ij}(\tau)$ and the time-delayed correlations $D_{ij}(\tau)$. The problem is to infer from these data the values of the couplings $J_{ij}(\tau)$ and of the local fields $H_i(\tau)$. Without loss of generality, we can use $\beta = 1$ as it is absorbed in the strength of couplings and fields that we want to infer. We shall solve this problem using the mean field equations.

The problems corresponding to different times and sites decouple. So let us consider a fixed value of i and τ, and infer the $J_{ij}(\tau)$ for $j = 1, \ldots, N$, and $H_i(\tau)$. To lighten notation we drop the explicit indices τ and i, and we denote $H = H_i(\tau)$, $m_j = m_j(\tau)$, $m = m_i(\tau + 1)$, $g = g_i(\tau)$, $\Delta = \Delta_i(\tau)$, $a = a_i(\tau)$. Following [21], one can obtain J by inverting the relation (18). The first step is to invert the empirical C matrix and compute:

$$b_j = \sum_k D_{ik}(\tau) C^{-1}_{kj}(\tau) \ . \tag{22}$$

If one knows the number $a = a_i(\tau)$ one can then infer the couplings from (18):

$$J_{ij}(\tau) = b_j / a \ . \tag{23}$$

Let us now see how a can be computed. The mean field equation (5) for the magnetization reads:

$$m = \int Dx \ \tanh\left[H + g + x\sqrt{\Delta}\right] \ . \tag{24}$$

The equation (19) for a is

$$a = \int Dx \left(1 - \tanh^2\left[H + g + x\sqrt{\Delta}\right]\right) \ . \tag{25}$$

The link between a and Δ is obtained from (4), which reads:

$$\Delta = \frac{1}{a^2} \sum_j b_j^2 (1 - m_j^2) = \frac{\gamma}{a^2} \ . \tag{26}$$

To solve this system of equations, we propose the following iterative procedure. Using the empirical correlations and magnetizations estimated from the patterns, we first compute from (22) the $\{b_j\}$, $j \in \{1, \ldots, N\}$, and $\gamma = \sum_j b_j^2 (1 - m_j^2)$.

Then we use the following mapping to find Δ.

- Start from a given value of Δ.
- Using the empirical value of m and the value of Δ, compute $H + g$ by inverting (24). The right-hand side of this equation is an increasing function of $H + g$ so this inversion is easy.
- Using $H + g$ and Δ, compute a using (25)
- Compute the new value of Δ, called $\hat{\Delta}$, using (26).

It is worth pointing out that in the thermodynamic limit, $N \to \infty$, the value of Δ becomes independent of i. So, if the system under consideration is large enough, the above iteration could be perfomed only once in order to reduce computation time.

This procedure defines a mapping from Δ to $\hat{\Delta} = f(\Delta)$, and we want to find a fixed point of this mapping. It turns out that a simple iterative procedure, starting from an arbitrary Δ_0 (for instance $\Delta_0 = 1$) and using $\Delta_{n+1} = f(\Delta_n)$, usuallly converges. More precisely, it can be shown that $f(0) = \frac{\gamma}{(1-m^2)^2}$ and that the asymptotic form for the slope of f for $\Delta \gg 1$ is $f' \sim \frac{\pi}{2} \gamma \exp(\hat{u}^2) \Delta_n$, where $\hat{u}$ is such that $m = \mathrm{erf}(\hat{u}/\sqrt{2})$. We have found numerically that when the number of patterns is large enough the slope verifies: $df/d\Delta \in]0, 1[$. Therefore the mapping converges exponentially fast to the unique fixed point. This method therefore works when the number of patterns per spin P/N is large enough. In the double limit $P, N \to \infty$ and P/N large enough the above procedure thus allows to get the exact result for Δ; and therefore to find the couplings $J_{ij}(\tau) = b_j / a$. Once the couplings have been found, one can easily compute

$g = \sum_j J_{ij}(\tau) m_j(\tau)$, and therefore get the local field $H(\tau)$. The number of operations needed for the full inference of the couplings and fields is dominated by the inversion of the correlation matrix C, a time which is typically at most of order N^3. If the number of patterns is too small, it may happen that there is no solution to the fixed point equation $f(\Delta) = \Delta$. Then one can decide to use $\Delta = f(0)$, which is nothing but the nMF estimate for $a_i(\tau)$.

We have tested our mean field inference method on the asymmetric SK problem, where the couplings J_{ij} are time-independent, gaussian distributed with variance β^2 and the fields are time independent, uniformly distributed on $[-\beta, \beta]$. Fig.(2) shows a scatter plot of the result on one given instance at $\beta = .4$ and $\beta = 1.4$, and compares it to the inference method of [21] using nMF and 'TAP' (the 'TAP' inference is limited to small values of β: at large β it fails). Figs. 3 and 4 show a statistical analysis of the performance of MF inference. It accurately infers the couplings and fields even in the strong coupling regime.

The method that we propose is exact and allows for a very precise inference of the couplings when applied to the fully asymmetric SK spin glass, at any temperature, if the number of patterns is large enough. At the same time, it is an easy and versatile method which can be used as an approximate inference method when the number of patterns is not very large (although one should at least have $P > N$ in order for C to be invertible), or when the underlying model is not of the SK type. As an example showing the possible use of the method, we have applied it to a sample where the J_{ij} matrix is sparse, generated as follows. We first generate a regular graph with 200 vertices and degree 6 on each vertex. For each edge ij of this graph we choose randomly with probability $1/2$ an orientation, say $i \to j$. Then we take $J_{ji} = 0$ and J_{ij} is drawn randomly from the probability density $(|x|/2)e^{-x^2/2}$. All the other couplings corresponding to pairs of sites kl which are not in the graph are set to 0. One then iterates the dynamics (1) 100000 times at $\beta = .6$, and uses this data to reconstruct the couplings. Fig. 5 shows the resulting couplings as a scatter plot. The topology of the underlying interaction graph can be reconstructed basically exactly, both by nMF and MF by using a threshold, deciding that all reconstructed couplings with $|J_{ij}| < .04$ vanish. The non-zero couplings are found accurately by the MF inference method.

To summarize, we have introduced a simple mean field method which can be applied on a single instance of a dynamical fully asymmetric Ising model. In the case of the asymmetric SK model this MF method gives the exact values of the local magnetizations and the exact relation between equal-time and time-delayed correlations. This method can be used to solve efficiently the inverse problem, i.e. determine the couplings and local fields from a set of patterns. Again this inference method is exact in the limit of large sizes and large number of patterns, in the asymmetric SK case. It can also be used in cases where the underlying model is different, for instance for diluted models. This could be quite useful for many applications.

We thank Lenka Zdeborová for useful discussions.This work has been supported in part by the EC grant 'STAMINA', No 265496.

[1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[2] E. Aurell, C. Ollion, and Y. Roudi. Dynamics and performance of susceptibility propagation on synthetic data. *The European Physical Journal B - Condensed Matter and Complex Systems*, 77:587–595, 2010. 10.1140/epjb/e2010-00277-0.

[3] A Braunstein, A Pagnani, M Weigt, and R Zecchina. Gene-network inference by message passing. *Journal of Physics: Conference Series*, 95(1):012016, 2008.

[4] Simona Cocco, Stanislas Leibler, and Rémi Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

[5] B. Derrida. Dynamical phase transition in non-symmetric spin glasses. *J. Phys. A*, 20:L721–L725, 1987.

[6] B. Derrida, E. Garner, and A. Zippelius. An exactly solvable asymmetric neural network model. *Europhys. Lett.*, 4:167–173, 1987.

[7] H. Gutfreund and M. Mézard. Processing of temporal sequences in neural networks. *Phys. Rev. Lett.*, 61:235–238, 1988.

[8] J. Hertz, G. Grinstein, and S. Solla. Irreversible spin glasses and neural networks. In J. van Hemmen and I. Morgenstern, editors, *Heidelberg Colloquium on Glassy Dynamics*, volume 275 of *Lecture Notes in Physics*, pages 538–546. Springer Berlin / Heidelberg, 1987. 10.1007/BFb0057533.

[9] John Hertz, Yasser Roudi, Andreas Thorning, Joanna Tyrcha, Erik Aurell, and Hong-Li Zeng. Inferring network connectivity using kinetic ising models. *BMC Neuroscience*, 11(Suppl 1):P51, 2010.

[10] Geoffrey E. Hinton. Deterministic boltzmann learning performs steepest descent in weight-space. *Neural Computation*, 1(1):143–150, 1989.

[11] Haiping Huang. Message passing algorithms for the hopfield network reconstruction: Threshold behavior and limitation. *Phys. Rev. E*, 82(5):056111, Nov 2010.

[12] Haiping Huang. Reconstructing the hopfield network as an inverse ising problem. *Phys. Rev. E*, 81(3):036104, Mar 2010.

[13] H. J. Kappen and F. B. Rodríguez. Efficient learning in boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156, 1998.

[14] H. J. Kappen and J. J. Spanjers. Mean field theory for asymmetric neural networks. *Phys. Rev. E*, 61(5):5658–5663, May 2000.
[15] Enzo Marinari and Valery Van Kerrebroeck. Intrinsic limitations of the susceptibility propagation inverse inference for the mean field ising spin glass. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(02):P02008, 2010.
[16] Marc Mézard and Thierry Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1-2):107 – 113, 2009. Neuromathematics of Vision.
[17] G. Parisi. Asymmetric neural networks and the process of learning. *Journal of Physics A Mathematical General*, 19:L675–L680, August 1986.
[18] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ${\ell_1}$-regularized logistic regression. *ArXiv e-prints*, October 2010.
[19] Yasser Roudi, Erik Aurell, and John Hertz. Statistical physics of pairwise probability models. *ArXiv e-prints*, 2009.
[20] Yasser Roudi and John Hertz. Dynamical tap equations for non-equilibrium ising spin glasses. *ArXiv e-prints*, 2011.
[21] Yasser Roudi and John Hertz. Mean field theory for nonequilibrium network reconstruction. *Phys. Rev. Lett.*, 106(4):048702, Jan 2011.
[22] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, April 2006.
[23] Vitor Sessak and Rémi Monasson. Small-correlation expansions for the inverse ising problem. *ArXiv e-prints*, 2008.
[24] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35(26):1792–1796, Dec 1975.
[25] Jonathon Shlens, Greg D. Field, Jeffrey L. Gauthier, Matthew I. Grivich, Dumitru Petrusca, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *The Journal of Neuroscience*, 26(32):8254–8266, 2006.
[26] Toshiyuki Tanaka. Mean-field theory of boltzmann machine learning. *Phys. Rev. E*, 58(2):2302–2310, Aug 1998.
[27] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35:593–601, 1977.
[28] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in proteinprotein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.
[29] H.-L. Zeng, E. Aurell, M. Alava, and H. Mahmoudi. Network inference using asynchronously updated kinetic Ising Model. *ArXiv e-prints*, November 2010.

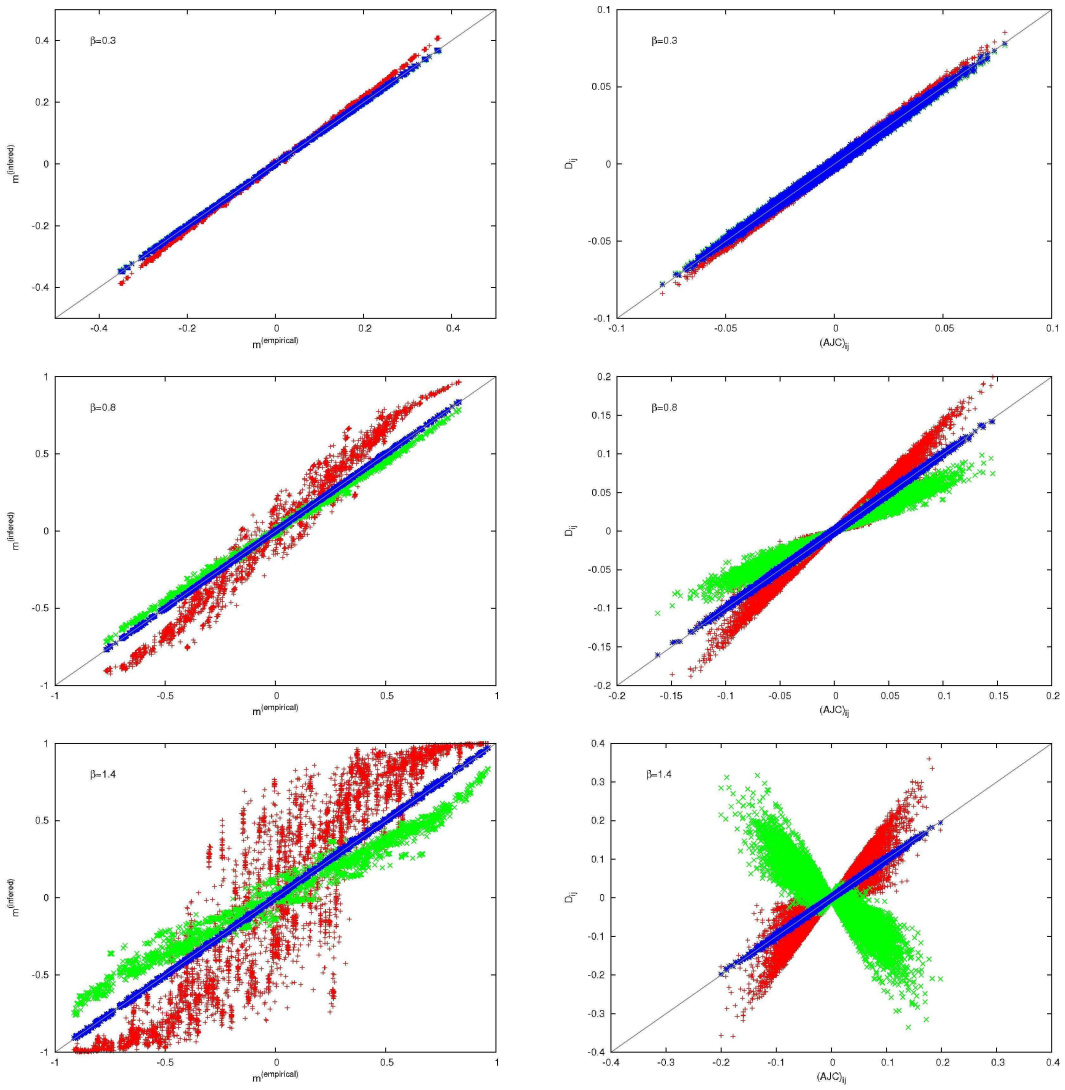


FIG. 1. Magnetizations (left column) and correlations (right column) obtained by MF (blue), 'TAP' (green) and nMF (red). One $N = 200$ spin model is simulated 500000 times for 31 time steps. The three plots in each column correspond to inverse temperature $\beta = .3$, .8 and 1.4 (from top to bottom). In the left column, the magnetizations predicted by each method for all time steps are plotted versus the experimental ones found by monte carlo simulation. For the plots of the right column, the correlation matrices C and D are obtained at $t = 30$. The scatter plot shows for each pair ij, the value of D_{ij} in ordinate, and the value of $(AJC)_{ij}$ in abscissa. The three methods differ in their predictions for A. At high temperature, $\beta = .3$, all methods are good for both the magnetizations and correlations; the MF and 'TAP' methods nearly coincide and are slightly better than nMF. At larger and larger β, the 'TAP' correction to naive mean field overshoots, and only the MF results is correct. The data supports the statement that MF is exact at all temperatures, while nMF and 'TAP' are only high temperature approximations.

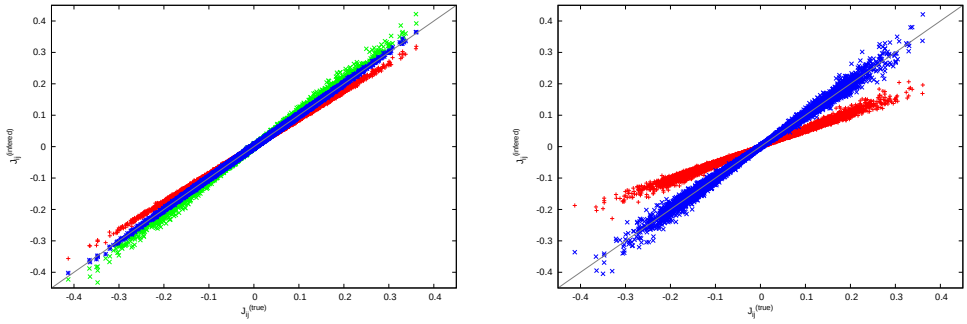


FIG. 2. Left: The infered couplings found by MF (blue), 'TAP' (green) and nMF (red) plotted versus the real ones for a $N = 100$ model, given $P = 1000000$ patterns generated at inverse temperature $\beta = 0.4$. Right: The same for $\beta = 1.4$ (MF (blue) and nMF (red), 'TAP' is not shown as it fails at this high β)

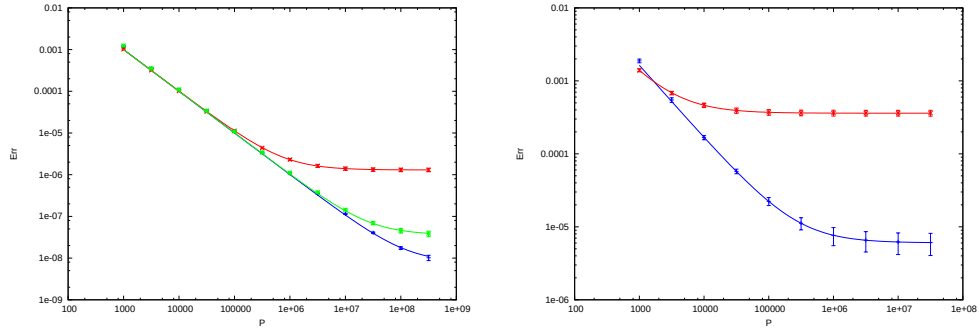


FIG. 3. Mean square error of the infered couplings $\overline{(J_{ij}^{\rm infered} - J_{ij}^{\rm real})^2}$ obtained by MF inference (blue), 'TAP' (green) and nMF (red) versus the number of patterns used to estimate the correlations, for a system of size $N = 40$, where the patterns were generated from a gaussian distribution with a root-mean-square $\beta = 0.2$ (left) and $\beta = 0.6$ (right). The curves are averages performed over 20 realizations of the couplings and fields. Notice that the 'TAP' method is absent in the right figure because it fails to provide results at strong coupling.

10

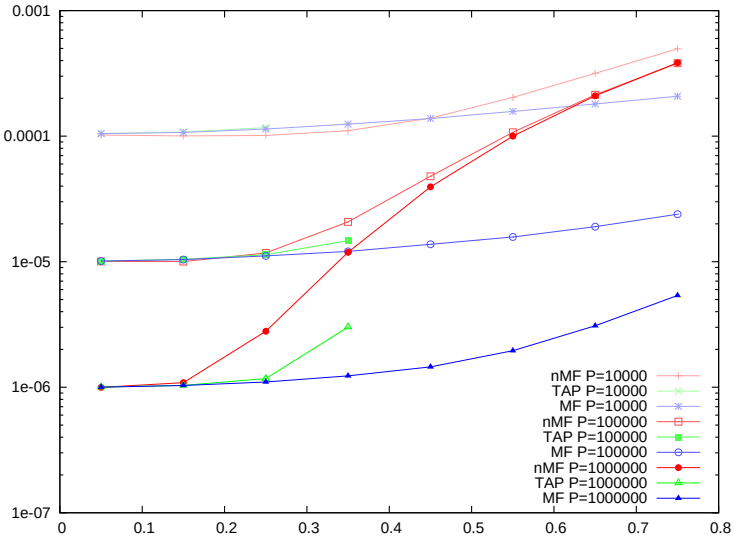


FIG. 4. Mean square error of the reconstructed couplings versus β, averaged over 10 systems with 100 spins, using the three inference methods nMF, 'TAP' and MF, with a number of patterns $P = 10000$, $P = 100000$ and $P = 1000000$. All three methods agree at small β. The nMF error can increase by several orders of magnitude at large β. The 'TAP' method fails to provide results above $\beta \approx 0.4$. The MF inference method gives good results in the whole range of β.

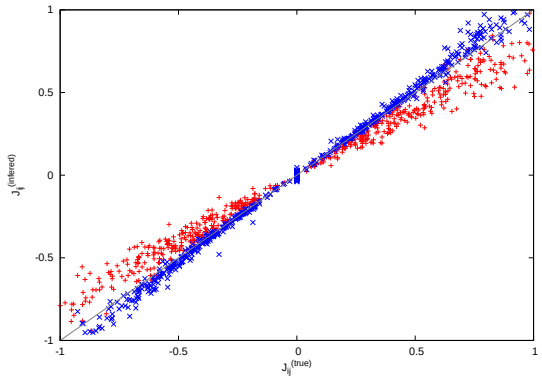


FIG. 5. Mean field inference of a finite connectivity model. The couplings found by MF (blue) and nMF (red) are plotted versus the real couplings used to generate the data for a $N = 200$ model on an asymmetric random regular graph with connectivity $c = 6$ (average in-degree=3) , given $P = 100000$ patterns generated at inverse temperature $\beta = .6$.

June 3, 2011 0:14 Philosophical Magazine v5

Philosophical Magazine
Vol. 00, No. 00, 00 Month 200x, 1–6

RESEARCH ARTICLE

Effect of coupling asymmetry on mean-field solutions of direct and inverse Sherrington-Kirkpatrick model

Jason Sakellariou[a], Yasser Roudi[bc],Marc Mezard [a],John Hertz[cd]

[a]*LPTMS, CNRS and Université Paris-Sud, 91405 Orsay Cedex, France*; [b]*Kavli Institute for Systems Neuroscience, NTNU, 7014 Trondheim, Norway*; [c]*NORDITA, 10691 Stockholm, Sweden*; [d]*The Niels Bohr Institute, , 2100 Copenhagen, Denmark*

(*Received 00 Month 200x; final version received 00 Month 200x*)

We study how the degree of symmetry in the couplings influences the performance of three mean field methods used for solving the direct and inverse problems for generalized Sherrington-Kirkpatrick models. In this context, the direct problem is predicting the potentially time-varying magnetizations. The three theories include the first and second order Plefka expansions, referred to as naive mean field (nMF) and TAP, respectively, and a mean field theory which is exact for fully asymmetric couplings. We call the last of these simply MF theory. We show that for the direct problem, nMF performs worse than the other two approximations, TAP outperforms MF when the coupling matrix is nearly symmetric, while MF works better when it is strongly asymmetric. For the inverse problem, MF performs better than both TAP and nMF, although an ad hoc adjustment of TAP can make it comparable to MF. For high temperatures the performance of TAP and MF approach each other.

Keywords: spin glass, mean field theory, inverse problems

1. Introduction

Predicting the dynamical properties of a disordered system given a specific realisation of its parameters is an old and important problem in statistical mechanics. This is what one can call a direct problem. Apart from being important on its own, solving the direct problem is also a crucial step in solving the inverse problem: inferring the parameters of a system from measurements of its dynamics. With the rapid advance of methods for observing the dynamics of biological systems composed of many elements, the inverse problem has received a lot of recent attention. This line of research has allowed inferring functional and physical connections in neuronal networks [1–5], gene regulatory networks [6] and protein residue contacts [7].

A useful platform for studying the inverse problem is a dynamical version of the Sherrington-Kirkpatrick (SK) model: a set of N classical spins, $s_i = \pm 1$ subject to a potentially time-varying external field $h_i(t)$ with couplings J_{ij} between them and a stochastic update rule. In the direct problem one tries to predict the magnetizations $m_i(t)$ given the coupling and fields. In the inverse problem one does the opposite, i.e. one infers the couplings and the fields from measured magnetizations and correlations.

When the system is in equilibrium and the distribution of states follows the Boltzmann distribution, several approaches for both direct and inverse problems have been developed. These include both exact and approximate iterative algorithms, such as Boltzmann learning and Susceptibility propagation [8, 9] relating

ISSN: 1478-6435 print/ISSN 1478-6443 online

DOI: 10.1080/14786435.20xx.xxxxxx
http://www.informaworld.com

the magnetizations to model parameters, as well as closed-form equations based on naive mean field (nMF) and TAP [10, 11] equations for the SK model. Motivated by the fact that biological systems are usually out of equilibrium, some recent work has focused on reconstructing the parameters of a dynamical Ising spin glass model obeying either synchronous or asynchronous updating from observing its out-of-equilibrium dynamics [5, 12, 13].

In this paper, we investigate how three recently proposed mean field methods for the direct and inverse problems perform on models with different degrees of symmetry in their coupling matrices. The three methods are the nMF and TAP equations, derived using the high-temperature Plefka expansions of the generating functional to first order and second order [14], and a mean field theory (denoted simply MF) [13] that is exact for the SK model with fully asymmetric couplings.

2. Solutions to the direct and inverse problems

We consider a model in which the probability of being in state $\mathbf{s}$ at time step t, $p_t(\mathbf{s})$, is given by

$$p_{t+1}(\mathbf{s}) = \sum_{\mathbf{s}'} W_t[\mathbf{s};\mathbf{s}']p_t(\mathbf{s}') \tag{1a}$$

$$W_t[\mathbf{s};\mathbf{s}'] = \prod_i \frac{\exp(s_i\theta_i)}{2\cosh\theta_i} \tag{1b}$$

$$\theta_i(t) = h_i(t) + \sum_j J_{ij}\, s'_j(t). \tag{1c}$$

For the choice of couplings J_{ij}, we follow [15], taking

$$J_{ij} = J_{ij}^{sym} + kJ_{ij}^{asym} \tag{2}$$

where $J_{ij}^{sym} = J_{ji}^{sym}$ is the symmetric part of the couplings while $J_{ij}^{asym} = -J_{ji}^{asym}$ is the antiymmetric part. All the couplings J_{ij}^{sym} and J_{ij}^{asym} are drawn independently from a zero-mean Gaussian distribution with variance

$$\overline{[J_{ij}^{symm}]^2} = \overline{[J_{ij}^{asym}]^2} = \frac{g^2}{(1+k^2)N}. \tag{3}$$

With Eqs. 2 and 3, the couplings J_{ij} have variance of g^2/N and the degree of symmetry is controlled by k: for $k = 0$ the model is fully symmetric ($J_{ij} = J_{ji}$) while for $k = 1$, it is fully asymmetric (J_{ij} independent of J_{ji}).

The direct problem consists in estimating the instantaneous magnetization of spin i at time t, $m_i(t)$. The estimation obtained from the nMF, TAP and MF are respectively:

$$m_i(t+1) = \tanh\Big[h_i(t) + \sum_j J_{ij} m_j(t)\Big] \tag{4a}$$

$$m_i(t+1) = \tanh\Big[h_i(t) + \sum_j J_{ij} m_j(t) - m_i(t+1)\sum_j J_{ij}^2(1-m_j^2(t))\Big] \qquad (4b)$$

$$m_i(t+1) = \int \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} \tanh\Big[h_i(t) + \sum_j J_{ij} m_j(t) + x\sqrt{\Delta_i(t)}\Big] \qquad (4c)$$

where in the last equation

$$\Delta_i(t) = \sum_j J_{ij}^2(1-m_i^2(t)) \ . \qquad (5)$$

For deriving Eqs. 4a and 4b, i.e. nMF and TAP, one first writes down the generating functional for the process defined by Eq. 1, performs a Legendre transform to fix the magnetizations and expands the results for small g (i.e. high temperature). To the first order, this expansion gives the nMF equations, Eq. 4a. Keeping terms up to the second order yields a correction to the nMF equations resulting in the the TAP equations, Eq. 4b, for this dynamical model. nMF and TAP are, therefore, high temperature expansions for an arbitrary set of couplings, with no assumption about their distribution or its degree of symmetry. The third equation is derived for arbitrary g, but under the mean-field assumption that at each time step the fields acting on the spins are independent Gaussian variables. This is exact for this SK model when the coupling matrix is fully asymmetric i.e. when $k = 1$.

These direct equations can also be used for solving the inverse problem. The idea is to use the data in order to measure the magnetizations $m_i(t)$, the equal time correlations $C_{ij} = \langle \delta s_i(t) \delta s_j(t) \rangle$, and the time-delayed correlations $D_{ij} = \langle \delta s_i(t+1) \delta s_j(t) \rangle$, where $\delta s_i(t) = s_i(t) - m_i(t)$. For the process in Eq. 1, one can write the time-delayed correlations as

$$D_{ij} = \langle \tanh\big[\theta_i(t)\big] s_j(t) \rangle - \langle \tanh\big[\theta_i(t)\big] \rangle \langle s_j(t) \rangle. \qquad (6)$$

To derive the inverse TAP and nMF, one then uses Eq. 6, expands the tanh around m_i that satisfies one of the direct equations 4a and 4b. In the case of MF, one writes an expression for the joint distribution of $\theta_i(t)$ and $\theta_j(t)$ that is exact for a fully asymmetric SK model. This joint distribution can then be used to relate $\mathbf{JD}$ to $\mathbf{C}$ in the limit of small C_{ij}; for details see [5, 13]. Within all three approximations, nMF, TAP, and MF, the resulting expression takes the form

$$\mathbf{D} = \mathbf{AJC} \ , \qquad (7)$$

where the matrix A is a diagonal matrix that depends on the approximation:

$$A_{ij}^{\rm nMF} = \delta_{ij}(1-m_i^2) \ , \qquad (8a)$$

$$A_{ij}^{\rm TAP} = \delta_{ij}(1-m_i^2)(1-F_i) \ , \qquad (8b)$$

$$A_{ij}^{\rm MF} = \delta_{ij} \int \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} \Big[1 - \tanh^2(h_i(t) + \sum_j J_{ij} m_j + x\sqrt{\Delta_i})\Big] \ . \qquad (8c)$$

In Eq. 8b F_i satisfies a cubic equation. For details see [5] and [13]. Not surprisingly, expanding Eq. 8c to linear or second order in J_{ij} yields $A^{\rm nMF}$ and $A^{\rm TAP}$ in Eqs. 8a and 8b, respectively.

Eq. 7 can be solved for $\mathbf{J} = \mathbf{A}^{-1}\mathbf{D}\mathbf{C}^{-1}$, provided one has enough data so that the estimation of C is good, allowing its numerical inversion.

3. Effect of Symmetry

As mentioned before, for the direct problem, we expect that the MF becomes exact for $k = 1$ for any coupling strength g. TAP equations should also become exact for $k = 0$ in the limit of weak couplings. This is shown in Fig. 1, where we plot the mean squared error in predicting the magnetizations at time $t + 1$ given the magnetizations at time t. This is done both for a constant field and for an external field that varies sinusoidally with time. As can be seen in this figure, for both types of external fields, TAP equations outperform the other two methods for small k. As temperature is increased, all three approximations perform better and become almost equally good. As k increases, MF wins over TAP while nMF performs worse than both of them.

The situation for the inverse problem is slightly more complicated. This is because, for strong couplings, the cubic equation that F_i solves develops complex roots. In this case one can take three approaches: (i) take the nMF result, (ii) take the real part of the solution, (iii) take the solution for the largest g for which the solutions are real. This value can be shown to be $F_i = 1/3$. The results for the last two strategies almost coincide, with strategy (iii) performing slightly better in lower temperatures, so we chose this one. In strategy (i) the results just coincide with the nMF approach after the temperature at which the cubic equation for F_i develops complex roots. The results from strategy (iii) are shown in Fig. 2. It is clear from this figure that nMF always performs worse than the other two and that the difference between the three methods vanishes in the high-temperature limit. On the other hand, MF is superior, as expected, when one gets closer to the asymmetric case i.e. for k is close to 1. The TAP result has a more complicated behavior, due to the intrinsic limitations imposed by the lack of real solutions of the cubic equation at strong couplings. However, one can notice that, when k is close to zero, there is a range of couplings g where TAP becomes better than MF as it is expected.

As can be seen in the right column of Fig. 2, the mean squared error $\overline{(J_{ij}^{\rm infered} - J_{ij}^{\rm real})^2}$ becomes larger for non-zero external fields. This is a general feature of all three methods. Large fields and/or couplings are estimated with larger errors than small ones. This is because errors in the estimation of the empirical magnetizations/correlations, when the later are close to ± 1, produce large errors in the estimation of the fields/couplings (consider for example, in zeroth order approximation, a sigmoid map between m_i and h_i , and c_{ij} and J_{ij}). Numerical simulations show that, for large external field amplitude, these errors become so important that the differences between the three methods are insignificant.

4. Conclusions

Within the mean field approaches that we have studied, the solution of the inverse problem derives from the solution of the direct problem. We have studied here three methods that provide an approximate solution to the direct problem in the case of systems with infinite range interactions. We have explored their behaviors

on both the direct and the inverse problem in the case of SK models with different degrees of symmetry of the interactions. As expected, the MF approach is the best one when the degree of asymmetry is large enough, but the TAP approach turns out to be slightly better in some range of coupling strength when the couplings are more symmetric. The nMF approach is just a first order approximation to both MF and TAP and is systematically worse than the other two methods.

As noted before, the derivation of inverse nMF and TAP rely on expanding the tanh in the around the solutions of the nMF and TAP. This expansion is not required for the MF solution: in the case with the assumption of full asymmetry, the joint distribution of the local field to each pair of spins will be Gaussian and can be easily calculated. It is therefore possible to write an exact equation relating D_{ij} to C_{ij} and the couplings which in the limit of small C_{ij} can be linearized and takes the form of Eq. 7. It would be interesting to see if a similar approach can be done within the TAP framework: calculate the joint distribution of the local fields in a systematic small coupling expansion, and use the same procedure done in MF to relate D_{ij} to C_{ij}.

In real applications, for instance in neural data analysis or gene regulation network reconstruction, one does not deal with data generated from a model with the particular size dependence of the couplings of the SK model. Our previous work shows that TAP and nMF perform at the same level in identifying the connections of a simulated neural network, and they both perform worse than the exact iterative Boltzmann like learning rule that one can write down for the dynamical SK model [5, 16]. We will leave the comparison of TAP, MF and the exact learning on biological data to future work.

Acknowledgement

The work of MM and JS has been supported in part by the EC grant 'STAMINA', No 265496.

References

[1] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.
[2] J. Shlens, G.D. Field, J.L. Gauthier, M.I. Grivich, D. Petrusca, A. Sher, A.M. Litke, and E.J. Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.*, 26:8254–8266, 2006.
[3] S. Cocco, S. Leibler, and R. Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc Natl Acad Sci U S A*, 106:14058–62, 2009.
[4] Y. Roudi, S. Nirenberg, and P. E. Latham. Pairwise maximum entropy models for studying large biological systems: when they can work and when they cant. *PLoS Comput Biol*, 5:e1000380, 2009.
[5] Y. Roudi and J. Hertz. Mean field theory for nonequilibrium network reconstruction. *Phys. Rev. Lett.*, 106:048702, 2011.
[6] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. USA*, 103, 2006.
[7] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *PNAS*, 106:67–72, 2009.
[8] Erik Aurell Charles Ollion and Yasser Roudi. Dynamics and performance of susceptibility propagation on synthetic data. *Eur. Phys. J. B*, 77:587–595, 2010.
[9] Marc Mézard and Thierry Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103:107–113, 2009.
[10] T. Tanaka. Mean-field theory of boltzmann machine learning. *Phys. Rev. E*, 58:2302–2310, 1998.
[11] H. J. Kappen and F. B. Rodriguez. Efficient learning in boltzmann machines using linear response theory. *Neur. Comp.*, 10:1137–1156, 1998.
[12] H.-L. Zeng, M. Alava, H. Mahmoudi, and E. Aurell. Network inference using asynchronously updated kinetic ising model. *Phys. Rev. E*, 83:041135, 2011.
[13] M. Mezard and J. Sakellariou. Exact mean field inference in asymmetric kinetic ising systems. *arXiv:1103.3433v2*, 2011.

[14] Y. Roudi and J. Hertz. Dynamical tap equations for non-equilibrium ising spin glasses. *J. Stat. Mech.*, page P03031, 2011.
[15] A. Crisanti and H. Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Phys. Rev. A*, 36:4922–4939, 1987.
[16] J. A. Hertz, Y. Roudi, A. Thorning, J. Tyrcha, E. Aurell, and H-L. Zeng. Inferring network connectivity using kinetic ising models. *BMC Neuroscience*, 10, 2010.

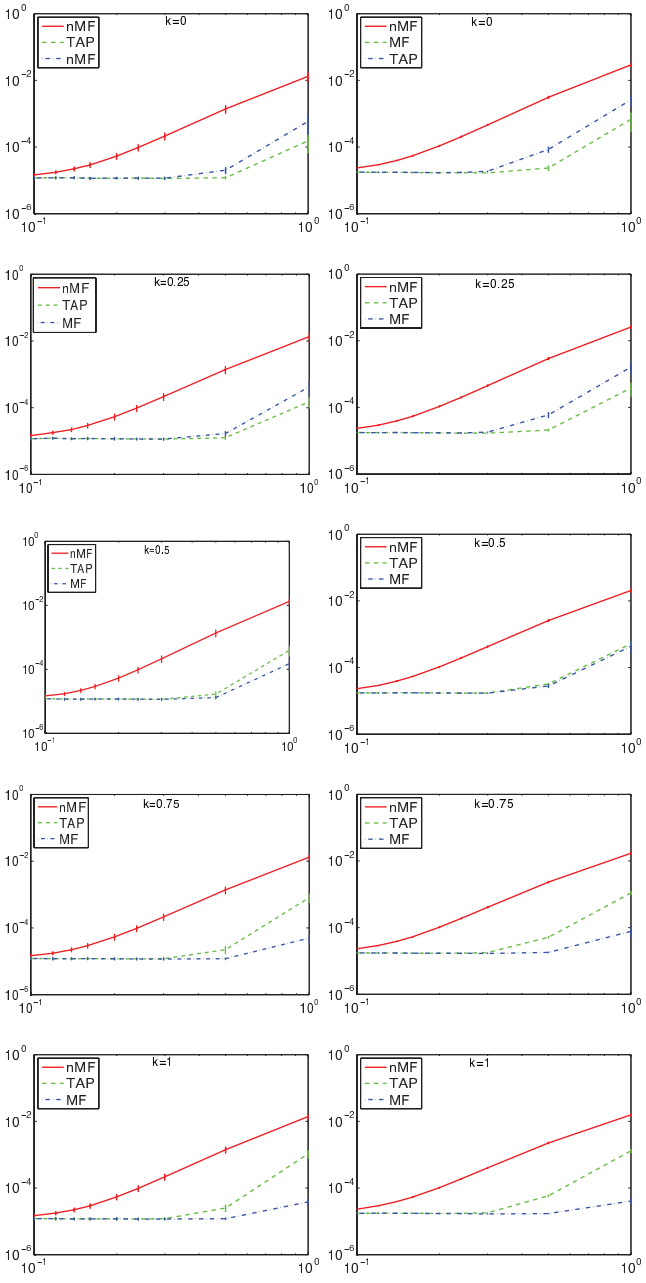


Figure 1.
LEFT PANELS: Mean squared error of the three methods for predicting the magnetizations at time t given at time $t-1$, averaged over spins and times, $\overline{(m_i^{\text{predicted}}(t) - m_i^{\text{measured}}(t))^2}$. This mean squared error is plotted as a function of g for a system of size $N = 50$ with a temporally constant field drawn independently for each spin from a normal distribution. We have used 100 time steps and 50000 repeats to calculate the experimental magnetizations and have averaged the errors over 10 realizations of the couplings. The different figures correspond to different values of k. From top to bottom $k = 0, 0.25, 0.5, 0.75, 1$. RIGHT PANELS: The same but with the addition of a sinusoidal external field of period 10 time steps and amplitude 0.5.

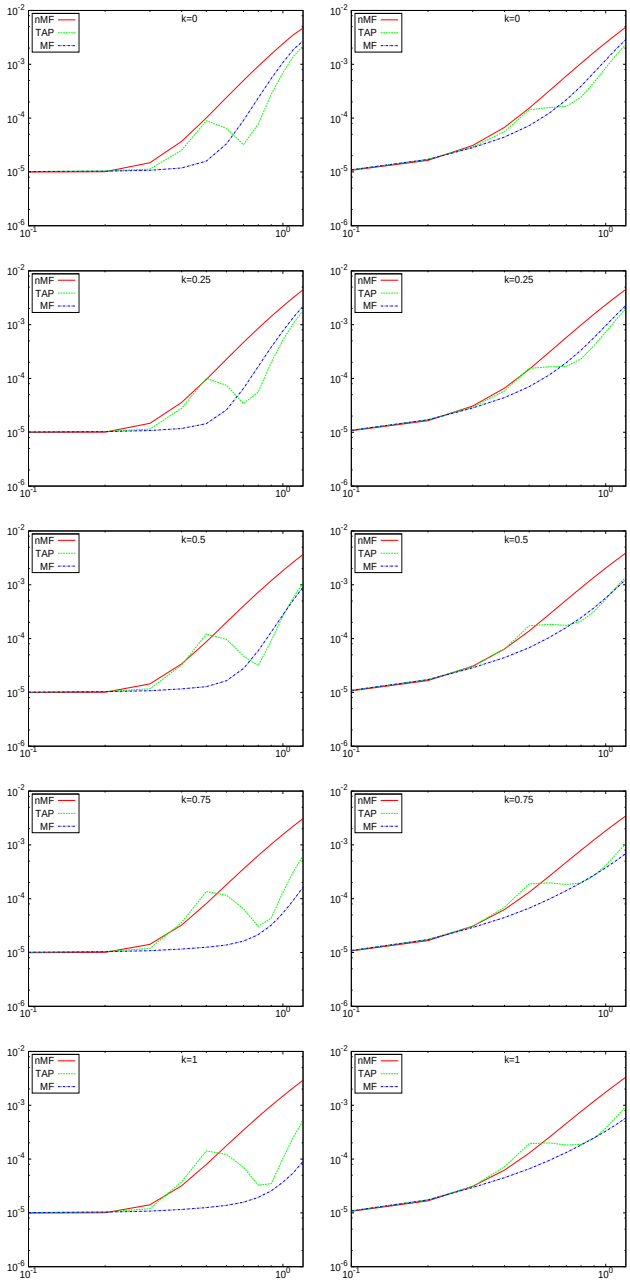


Figure 2.
LEFT PANELS: Mean squarre error of the three methods on the infered couplings $\overline{(J_{ij}^{\text{infered}} - J_{ij}^{\text{real}})^2}$ as a function of g for systems of size $N = 100$ with zero external field, given $P = 100000$ patterns, averaged over 10 realizations of the couplings. The different figures correspond to different values of k. From top to bottom $k = 0, 0.25, 0.5, 0.75, 1$. RIGHT PANELS: The same but with the addition of a sinusoidal external field of period 10 time steps and amplitude 0.5.

Bibliography

[AckleyHS 85] Ackley D. H., Hinton G. E. & Sejnowski T. J. (1985) *A learning algorithm for boltzmann machines.* Cognitive Science, **9**(1):147-169.

[AlbertsJL+ 02] Alberts B., Johnson A., Lewis J., Raff M., Roberts K. & Walter P. (2002) *Molecular biology of the cell.* New York: Garland Science.

[AmariKN 92] Amari S., Kurata K. & Nagaoka H. (1992) *Information geometry of Boltzmann machines.* IEEE Transactions on Neural Networks, Volume: **3** , Issue: 2, Pages:260-271.

[AurellOR 10] Aurell E., Ollion C. & Roudi Y. (2010) *Dynamics and performance of susceptibility propagation on synthetic data.* The European Physical Journal B, Volume **77**, Issue 4, pp 587-595.

[BentoM 09] Bento J. & Montanari A. (2009) *Which graphical models are difficult to learn?* ArXiv e-prints arXiv:0910.5761.

[BerrouG 96] Berrou C. & Glavieux A. (1996) *Near optimum error correcting coding and decoding: turbo-codes.* IEEE Trans. Commun., **44**, 1261-1271.

[BraunsteinMZ 05] Braunstein A., Mézard M. & Zecchina R. (2005) *Survey propagation: An algorithm for satisfiability.* Random Struct. Alg., **27**: 201-226.

[BraunsteinPWZ 08] Braunstein A., Pagnani A., Weigt M. & Zecchina, R. (2008) *Gene-network inference by message passing.* J. Phys.: Conf. Ser. **95** 012016.

[ChowL 68] Chow C. & Liu C. (1968) *Approximating discrete probability distributions with dependence trees.* IEEE Transactions on Information Theory, Volume **14** Issue 3 page 462-467.

[CoccoLM 09] Cocco S., Leibler S. & Monasson R. (2009) *Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods.* Proceedings of the National Academy of Sciences, **106**(33):14058-14062

[CoccoM 11] Cocco S. & Monasson R. (2011) *Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data.* Phys. Rev. Lett. **106**, 090601.

[CoccoM 12] Cocco S. & Monasson R. (2012) *Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests.* Journal of Statistical Physics, Volume **147**, Issue 2, pp 252-314.

[Cooper 90] Cooper G. F. (1990) *The computational complexity of probabilistic inference using bayesian belief networks.* Artificial Intelligence, Volume **42**, Issues 2-3, pages 393-405.

[CoverT 91] Cover T. M. & Thomas J. A. (1991) *Elements of Information Theory.* John Wiley & Sons, ISBN 0471062596.

[Derrida 87] Derrida B. (1987) *Dynamical phase transition in non-symmetric spin glasses.* J. Phys. A, **20**:L721-L725.

[DerridaGZ 87] Derrida B., Garner E. & Zippelius A. (1987) *An exactly solvable asymmetric neural network model.* Europhys. Lett., **4**:167-173.

[EdwardsA 75] Edwards S. F. & Anderson P. W. (1975) *Theory of spin glasses.* J. Phys. F: Met. Phys. **5** 965.

[Freyre-GonzalezT 10] Freyre-Gonzalez, J. A. & Trevino-Quintanilla L. G. (2010) *Analyzing Regulatory Networks in Bacteria.* Nature Education **3**(9):24.

[Gallager 62] Gallager R. (1962) *Low-density parity-check codes.* IRE Trans. Inform. Theory IT-**8**, 21-28.

[GeorgesY 91] Georges A. & Yedidia J. A. (1991) *How to expand around mean-field theory using high-temperature expansions.* J. Phys. A: Math. Gen. **24** 2173.

[GutfreundM 88] Gutfreund H. & Mézard M. (1988) *Processing of temporal sequences in neural networks.* Phys. Rev. Lett., **61**:235-238.

[HertzGS 87] Hertz J., Grinstein G. & Solla S. (1987) *Irreversible spin glasses and neural networks.* Heidelberg Colloquium on Glassy Dynamics, volume **275** of Lecture Notes in Physics, pages 538-546. Springer Berlin / Heidelberg, 1987. 10.1007/BFb0057533.

[HertzRTT$^+$ 10] Hertz J., Roudi Y., Thorning A., Tyrcha J., Aurell E. & Zeng H.-L. (2010) *Inferring network connectivity using kinetic ising models.* BMC Neuroscience, **11**(Suppl 1):P51.

[Hinton 89] Hinton G. E. (1989) *Deterministic boltzmann learning performs steepest descent in weight-space.* Neural Computation, **1**(1):143-150.

[Hopfield 82] Hopfield J. J. (1982) *Neural networks and physical systems with emergent collective computational abilities.* PNAS vol. **79** no. 8 2554-2558.

[Huang 10a] Huang H. (2010) *Reconstructing the hopfield network as an inverse ising problem.* Phys. Rev. E, **81**(3):036104.

[Huang 10b] Huang H. (2010) *Message passing algorithms for the hopfield network reconstruction: Threshold behavior and limitation.* Phys. Rev. E, 82(5):056111.

[Jaynes 57a] Jaynes E. T. (1957) *Information theory and statistical mechanics.* Phys. Rev. **106**, 620-630.

[Jaynes 57b] Jaynes E. T. (1957) *Information theory and statistical mechanics II.* Phys. Rev. **108**, 171-190.

[KappenR 97] Kappen H. J. & Rodríguez F. B. (1997) *Mean field approach to learning in Boltzmann Machines.* Pattern Recognition Letters, Volume **18**, Issues 11-13, Pages 1317-1322.

[KappenR 98] Kappen H. J. & Rodríguez F. B. (1998) *Efficient learning in boltzmann machines using linear response theory.* Neural Computation, **10**(5):1137–1156.

[KappenS 00] Kappen H. J. & Spanjers J. J. (2000) *Mean field theory for asymmetric neural networks.* Phys. Rev. E, **61**(5):5658-5663.

[Kikuchi 51] Kikuchi R. (1951) *A Theory of Cooperative Phenomena.* Phys. Rev. **81**, 988-1003.

[KullbackL 51] Kullback S. & Leibler R. A. (1951) *On Information and Sufficiency.* The Annals of Mathematical Statistics Vol. **22**, No. 1, pp. 79-86.

[MarinariK 10] Marinari E. & Van Kerrebroeck V. (2010) *Intrinsic limitations of the susceptibility propagation inverse inference for the mean field Ising spin glass.* Journal of Statistical Mechanics: Theory and Experiment, **2010**(02):P02008.

[MeisterPB 94] Meister M., Pine J. & Baylor D. A. (1994) *Multi-neuronal signals from the retina: acquisition andanalysis.* Journal of neuroscience methods, vol. **51**, no. 1, page 95.

[MetropolisRRT$^+$ 53] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H. & Teller E. (1953) *Equation of State Calculations by Fast Computing Machines.* J. Chem. Phys. **21**, 1087.

[MezardM 09] Mézard, M. and Montanari, A. (2009) *Information, Physics, and Computation.* Oxford Graduate Texts, ISBN-10: 019857083X.

[MezardPV 87] Mézard M., Parisi G. & Virasoro M. A. (1987) *Spin glass theory and beyond.* Singapore: World Scientific, ISBN 9971-5-0115-5.

[MezardS 11] Mézard & Sakellariou J. (2011) *Exact mean-field inference in asymmetric kinetic Ising systems.* J. Stat. Mech. **2011** L07001.

[MoraM 09] Mora T. & Mézard M. (2009) *Constraint satisfaction problems and neural networks: A statistical physics perspective.* Journal of Physiology-Paris, Volume **103**, Issues 1-2, Pages 107-113.

[Nishimori 01] Nishimori H. (2001) *Statistical Physics of Spin Glasses and Information Processing: An Introduction.* Oxford University Press, ISBN 0198509405.

[Parisi 86] Parisi G. (1986) *Asymmetric neural networks and the process of learning.* Journal of Physics A Mathematical General, **19**:L675-L680.

[Pearl 88] Pearl J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann, ISBN 1558604790.

[PeyracheBK$^+$ 09] Peyrache A., Benchenane K., Khamassi M., Wiener S.I. & Battaglia F.P. (2009) *Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution.* Journal of Computational Neuroscience, pages 1-17.

[Plefka 82] Plefka T. (1982) *Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model.* J. Phys. A: Math. Gen. **15** 1971.

[Prim 57] Prim R. C. (1957) *Shortest connection matrix network and some generalizations.* Bell System Tech. J., **36**, 1389-1401.

[RavikumarWL 10] Ravikumar P., Wainwright M. J. & Lafferty J. D. (2010) *High-dimensional Ising model selection using ℓ_1-regularized logistic regression.* Ann. Statist. Volume **38**, Number 3, 1287-1319.

[Ricci-Tersenghi 12] Ricci-Tersenghi F. (2012) *The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods.* Journal of Statistical Mechanics: Theory and Experiment, **2012**, Issue 08, pp. 08015.

[RoudiAH 09] Roudi Y., Aurell E. & Hertz J. (2009) *Statistical Physics of Pairwise Probability Models.* Front Comput Neurosci., **3**: 22.

[RoudiH 11a] Roudi Y. & Hertz J. (2011) *Mean Field Theory for Nonequilibrium Network Reconstruction.* Phys. Rev. Lett. **106**, 048702.

[RoudiH 11b] Roudi Y. & Hertz J. (2011) *Dynamical TAP equations for non-equilibrium Ising spin glasses.* J. Stat. Mech. **2011** P03031.

[RoudiTH 09] Roudi Y., Tyrcha J. & Hertz J. (2009) *Ising model for neural data: Model quality and approximate methods for extracting functional connectivity.* Phys. Rev. E **79**, 051915.

[SakellariouRMH 12] Sakellariou J., Roudi Y., Mezard M., & Hertz J. (2012) *Effect of coupling asymmetry on mean-field solutions of direct and inverse Sherrington-Kirkpatrick model.* Philosophical Magazine Volume **92**, Issue 1-3, Special Issue: Many-body Theory, Magnetism, Spin Glasses and Related Phenomenon.

[SchneidmanBSB 06] Schneidman E., Berry M. J., Segev R. & Bialek W. (2006) *Weak pairwise correlations imply strongly correlated network states in a neural population.* Nature, **440**(7087):1007–1012.

[SessakM 09] Sessak V. & Monasson R. (2009) *Small-correlation expansions for the inverse Ising problem.* J. Phys. A: Math. Theor. **42** 055001.

[Shannon 48] Shannon C. E. (1948) *A Mathematical Theory of Communication.* Bell System Technical Journal, Vol. **27**, pp. 379-423, 623-656.

[SherringtonK 75] Sherrington D. & Kirkpatrick S. (1975) *Solvable model of a spin-glass.* Physics Review Letters **35** (26): 1792-1796.

[ShlensFGG$^+$ 06] Shlens J., Field G. D., Gauthier J. L., Grivich M. I., Petrusca D., Sher A., Litke A. M. & Chichilnisky E. J. (2006) *The structure of multi-neuron firing patterns in primate retina.* The Journal of Neuroscience, **26**(32):8254-8266.

[Tanaka 98] Tanaka T. (1998) *Mean-field theory of boltzmann machine learning.* Phys. Rev. E, **58**(2):2302-2310

[ThoulessAP 77] Thouless D. J., Anderson P. W. & Palmer R. G. (1977) *Solution of 'Solvable model of a spin glass'.* Philosophical Magazine, Volume **35**, Issue 3.

[Tibshirani 96] Tibshirani R. (1996) *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological) Vol. **58**, No. 1, pp. 267-288.

[TyrchaRMH 12] Tyrcha J., Roudi Y., Marsili M. & Hertz J. (2012) *Effect of Nonstationarity on Models Inferred from Neural Data.* ArXiv e-prints arXiv:1203.5673.

[WeigtWS$^+$ 09] Weigt M., White R. A., Szurmant H., Hoch J. A. & Hwa T.(2009) *Identification of direct residue contacts in proteinprotein interaction by message passing.* Proceedings of the National Academy of Sciences, **106**(1):67-72.

[Wentian 90] Wentian L. (1990) *Mutual information functions versus correlation functions.* Journal of Statistical Physics, Volume **60**, Issue 5-6, pp 823-837.

[YedidiaFW 03] Yedidia J. S., Freeman W. T. & Weiss Y. (2003) *Understanding Belief Propagation and its Generalizations.* Exploring artificial intelligence in the new millennium, ISBN:1-55860-811-7, Pages 239-269.

[ZengAAM 10] Zeng H.-L., Aurell E., Alava M. & Mahmoudi H. (2010) *Network inference using asynchronously updated kinetic Ising Model.* ArXiv e-prints, arXiv:1011.6216.