Circular belief propagation as a model for optimal and suboptimal inference in the brain : extending the algorithm and proposing a neural implementation

Vincent Bouttier

Thèse de doctorat

Circular Belief Propagation as a model for optimal and suboptimal inference in the brain: extending the algorithm and proposing a neural implementation

par **Vincent BOUTTIER**

Thèse de doctorat - Université de Paris
Spécialité Neurosciences et Troubles Neuronaux
Laboratoire de Neurosciences Cognitives & Computationnelles,
École Normale Supérieure
École Doctorale n°474 Frontières de l'Innovation en Recherche et Education

Dirigée par **Renaud JARDRI** et **Sophie DENÈVE**

Présentée et soutenue publiquement à Paris, le 15 décembre 2021

Devant un jury composé de:

Peggy SERIÈS PhD, Senior Lecturer, University of Edinburgh	Rapporteur
Xaq PITKOW PhD, Assistant Professor, Rice University	Rapporteur
Fabien VINCKIER MD-PhD, MCU-PH, Université de Paris	Examinateur
Katharina SCHMACK PhD, Group Leader, Francis Crick Institute	Examinateur
Renaud JARDRI MD-PhD, PU-PH, Université de Lille	Directeur de thèse
Sophie DENÈVE PhD, Directeur de Recherche, École Normale Supérieure	Co-directeur de thèse

Preface

This thesis gathers the work I have accomplished as a doctoral student at École Normale Supérieure in Paris and University of Lille, from October 2017 to October 2021. First and foremost, I would like to thank all the people that made these last four years a period of my life which I will look back at with joy (definitely) and nostalgia (probably).

I was lucky to be supervised by two truly great people, Renaud Jardri and Sophie Denève. Thank you for believing in me by taking me as a Master's student, and later hiring and funding me as a doctoral student. Thank you for letting me space to grow as an independent researcher, and for working at strange times close to my deadlines. I hope that your long-lasting collaboration will keep thriving when I'm gone, together with the field of computational psychiatry.

Having two supervisors made the journey particularly interesting and stimulating, not only because it allowed me for some nice trips to Lille, but also because Renaud and Sophie have very complementary skills and expertise. The three of us definitely have very different styles as researchers, and I learnt a lot from it. Thank you, Renaud, for being so positive all along these 4 years, and for showing me that science was not only about the results but also the way to communicate it. I admire your multitasking abilities and your efficiency, and I'm still impressed by your ability to answer my messages any time of the day, week, or year. Thank you, Sophie, for your spontaneity and brightness, for always having an idea on everything and being able to cover paper with equations impressively fast. Thank you also for the nice moments spent at our relocated office during the pandemic with tea and cake, allowing us to connect in a different way than just work-wise.

Thanks to people who, for a short or long period of time, worked with me at the Group for Neural Theory within the team: Pantelis, Mirjana, Ivan, Lyudmila, Alireza, and Suhrit. Many thanks to the GNT as a whole, for being at the same time smart and dumb and as much into politics and work as into movie nights and breaks. I truly love the atmosphere inside the lab, with each member bringing his own personality to the group.

Thanks to all the researchers who I met along the way, from PhD students to professors, for the interactions and the interesting discussions. This particularly includes Xaq Pitkow and Peggy Seriès, who I am happy to have as reviewers of the work, Christian Machens and his lab, and Jean Daunizeau and Philippe Domenech for their feedback at the thesis committee meetings. Thanks to Valentin Wyart for your time dedicated to our collaboration; our common project does not even appear in this thesis, but sometimes what is important is the journey rather than getting the results you were hoping for.

Thanks for all the positive mindset that can be found in everyone doing science. That includes the other trainees at the wonderful CAJAL Course on Computational Neuroscience and the Barcelona Advanced Modeling of Behavior summer schools, and the participants to the conferences I attended to, in locations as diverse as Berlin, Lisbon, Lille and Helsinki.

Thanks to all of my friends and family for helping me keep a nice work-life balance. Special thanks to Eli for your excellent English proofreading service and your jokes. Parts of this work have been assessed by Marc and Adrian, two amazing friends who are always here to have fun, go hiking you-know-where, eat croziflette, share and apparently help when needed. I hope to keep you around. I also want to thank Lidiya for all the great moments spent together and for showing faith in me all along the way. Lastly and most importantly, thank you Mom and Dad for being so supportive and caring, that was extremely valuable to me.

I entered the world of research simply because I was curious about it and I do not regret this decision any minute. I wish you a good reading.

Vincent BOUTTIER
October 2021

Synthèse

Titre : La propagation circulaire de croyances comme modèle d'inférences optimales et sous-optimales dans le cerveau : extension de l'algorithme et proposition d'implémentation neurale

Résumé : Le modèle d'inférence circulaire est un modèle bayésien de troubles psychiatriques, initialement conçu pour rendre compte des manifestations cliniques de la schizophrénie et de la psychose. L'inférence circulaire repose sur l'algorithme de propagation circulaire de croyances, un algorithme d'inférence probabiliste approximative qui propose un paramètre additionnel comparé à l'algorithme de propagation de croyances ou Belief Propagation. Ce paramètre est appelé le facteur de correction des boucles. Il fixe la quantité de circularité dans l'inférence et est considéré comme représentant le niveau d'équilibre (local) entre les processus d'excitation et d'inhibition dans le réseau cérébral supposé réaliser les opérations d'inférence probabiliste. Dans ce cadre, les raisonnements circulaires et les symptômes psychotiques émaneraient d'une diminution du facteur de correction de boucles, c'est-à-dire d'un faible niveau d'inhibition comparé au niveau d'excitation.
Le travail présenté dans cette thèse permet d'appuyer le modèle d'inférence circulaire comme modèle d'inférences pathologiques (par exemple les hallucinations et les idées délirantes), d'inférences presque-optimales, et entre les deux d'inférences sous-optimales non cliniques, allant des biais usuels d'inférence (comme l'illustrent les phénomènes de perception bistable et de prise de décision hâtive, et la confiance excessive généralisée) aux comportements infracliniques comme le fait de croire en des théories du complot malgré des éléments contredisant ces théories.
De plus, cette thèse développe le modèle d'inférence circulaire de façons diverses. Premièrement, conceptuellement, en procurant à l'algorithme de propagation circulaire de croyances une fondation théorique, ce qui est réalisé en le reliant à des algorithmes existants comme la propagation fractionnaire de croyances. Deuxièmement, de façon plus pratique, en proposant des implémentations neurales (réseaux de neurones à rate ou à spikes, pour des variables binaires ou gaussiennes) et des mécanismes d'apprentissage biologiquement plausibles décrivant tous les deux comment les inférences probabilistes pourraient être réalisées dans le cerveau en utilisant cet algorithme. Enfin, le modèle est développé sur le plan théorique, en examinant les propriétés de convergence de l'algorithme de propagation circulaire, en formulant l'algorithme pour des distributions de probabilité plus complexes que précédemment, et en proposant une généralisation avec l'algorithme de propagation circulaire étendu.

Mots clés : psychiatrie computationnelle, neurosciences théoriques, inférence probabiliste, propagation des convictions, inférence circulaire, propagation circulaire des convictions, déséquilibre excitation-inhibition, schizophrénie, psychose

Synthesis

Title: Circular Belief Propagation as a model for optimal and suboptimal inference in the brain: extending the algorithm and proposing a neural implementation

Abstract: Circular Inference is a Bayesian model of psychiatric disorders, previously designed to account for clinical manifestations of schizophrenia and psychosis. Circular Inference relies on the Circular Belief Propagation algorithm, an approximate probabilistic inference algorithm that proposes an additional parameter compared to Belief Propagation, called the *loop correction factor*. This loop correction factor sets the amount of circularity in the inference and is seen as a proxy to the (local) level of excitation-inhibition balance in the brain network assumed to perform probabilistic inferences. According to this framework, circular reasoning and psychotic symptoms arise for lowered loop correction factor, which would mean, for low levels of inhibition compared to excitation.
The work presented in this thesis provides further evidence for Circular Inference as a model of pathological inferences (e.g., hallucinations and delusions), near-optimal inferences, and in between non-clinical suboptimal inferences, ranging from usual inference biases (exemplified by the bistable perception and the jumping to conclusions phenomena, and the general overconfidence) to sub-clinical behavior like believing in conspiracy theories despite contradicting evidence.
Additionally, this thesis develops the Circular Inference model in different ways. First, conceptually, by providing the Circular BP algorithm with a theoretical foundation, which is done by relating it to existing algorithms such as Fractional BP. Second, more practically, by proposing neural implementations (rate networks and spiking networks, for binary or Gaussian variables) and biologically-plausible learning mechanisms overall describing how probabilistic inferences could be carried out in the brain using this algorithm. Finally, the model is expanded theoretically, by investigating the convergence properties of the algorithm, by writing Circular BP for more complex probability distributions than previously, and by generalizing the initial Circular BP into extended Circular BP.

Keywords: computational psychiatry, theoretical neuroscience, probabilistic inference, belief propagation, circular inference, circular belief propagation, excitation-inhibition imbalance, schizophrenia, psychosis

Résumé en français

Le cerveau humain est extrêmement performant pour effectuer des tâches complexes, et en particulier, des tâches probabilistes. Ces tâches, par définition, impliquent des informations incertaines ou bruitées. Elles se retrouvent dans la vie de tous les jours – estimer quel âge a la personne en face de nous, décider quand partir de chez soi pour être à l'heure à 95 % – et pourtant, sont d'une grande complexité mathématique. L'hypothèse naturelle découlant de ces observations est que le cerveau humain représente d'une manière ou d'une autre les probabilités via ses neurones, et effectue des calculs probabilistes de façon (quasi-)optimale via l'activité de ces mêmes neurones.

Différentes hypothèses mathématiques ont été formulées pour proposer une implémentation (relativement) biologiquement plausible 1. de cette représentation des probabilités et 2. de ces calculs probabilistes, dans le cas des distributions de probabilité sur des variables binaires. Des candidats naturels sont, respectivement, des neurones se déchargeant proportionnellement aux quantités qu'ils encodent (liées directement aux probabilités), et l'algorithme de Propagation de Croyances (« PC », en anglais Belief Propagation) qui réalise une opération mathématique dite de marginalisation de probabilité. Cet algorithme agit sur des graphes et permet de propager l'information arrivant dans le réseau (dans l'exemple, la couleur des cheveux de la personne en face de nous) par un échange de messages entre ses nœuds au niveau de ses arêtes. Plus précisément, un message envoyé par un noeud A à un noeud B est composé de l'information totale reçue par le noeud A arrivant de l'extérieur du réseau, et des noeuds connectés à A sauf le noeud B (c'est-à-dire, le message allant de B à A). Cette exclusion permet à l'information de ne pas être réverbérée. En effet, dans le cas contraire, si l'information que le nœud B envoie à A était renvoyée à B (elle-même renvoyée à A, etc.), cela donnerait lieu à une amplification sans raison de l'information et des croyances, donnant lieu à un excès de confiance des noeuds du réseau en comparaison à la réalité. Ce modèle de représentation et de calcul probabiliste est un modèle purement abstrait (défini algorithmiquement) pouvant potiellement expliquer le comportement. Néanmoins, le modèle peut être rapproché de la biologie sur le plan intuitif. En effet, les noeuds du réseau peuvent représenter chacun une population de neurones représentant collectivement la probabilité d'une variable mathématique, et les arêtes du réseau peuvent représenter chacun un faisceau de connexions entre ces populations de neurones. L'échange de messages pourrait se faire entre des neurones excitateurs du cerveau, tandis que les neurones inhibiteurs pourraient empêcher que l'information ne se réverbère en effectuant l'opération d'exclusion (du message allant de B à A), contrôlant ainsi les échanges d'informations entre les neurones excitateurs.

C'est justement la perturbation, partielle ou totale, de ce mécanisme d'exclusion (de l'information provenant du noeud B) qui définit l'algorithme de Propagation Circulaire de Croyances (« PCC », en anglais Circular Belief Propagation). Une partie de l'information envoyée de B à A retourne à B à cause de la réciprocité des connexions et de la perturbation, ce qui constitue une boucle de longueur 2. La perturbation introduit des excès de confiance et des potentielles erreurs de raisonnements, ce qui a motivé le modèle dit « modèle d'inférence circulaire » basé sur l'algorithme de Propagation Circulaire de Croyances. La quantité de circularité de l'inférence est contrôlée par des paramètres associés au arêtes du graphe, dits taux de correction des boucles (égaux à 1 dans le cas de l'algorithme PC, et inférieurs à 1 dans le cas de l'algorithme PCC). Le modèle d'inférence circulaire vise à modéliser les croyances aberrantes et plus particulièrement l'état de psychose, définit comme un trouble mental de perte de contact avec la réalité et caractérisé par des pensées délirantes (croyances non partagées par la majorité et résistant à des informations qui les contredisent) ou des hallucinations (voir ou entendre des choses absentes). Biologiquement parlant, selon la vision simpliste adoptée ci-dessus, la perturbation du taux de correction des boucles pourrait venir d'un défaut d'inhibition : l'information envoyée entre populations excitatrices n'est pas suffisamment atténuée, et est donc réverbérée. Un défaut d'inhibition déplace l'équilibre de la balance excitation-inhibition dans le cerveau. Ce déplacement d'équilibre est une des

principales caractérisations de la schizophrénie, même s'il reste à clarifier biologiquement si c'est la véritable cause du trouble psychiatrique ou si cela serait une conséquence d'une autre cause – par exemple, de la perturbation du système dopaminergique ou encore d'une disconnection anatomique.

Suite à sa définition algorithmique en 2013, le modèle d'inférence circulaire été développé. Il a par exemple été utilisé pour modéliser le comportement lors d'une tâche probabiliste de combinaison d'information chez des personnes atteintes de schizophrénie et des personnes dans la population générale. De façon conforme à l'idée du spectre de la psychose, les personnes avec la plus grande perturbation comportementale dans la tâche (mesurée par le taux de correction des boucles) étaient celles qui avaient les symptômes les plus intenses. Néanmoins, le modèle d'inférence circulaire manque cruellement d'avancées qui permettraient de le considérer comme biologiquement plausible, c'est-à-dire comme une façon possible dont le cerveau pourrait effectuer des raisonnements probabilistes.

Cette thèse permet de développer le modèle d'inférence circulaire de plusieurs manières, à la fois sur le plan pratique et sur le plan théorique. Sur le plan pratique, une proposition d'implémentation neurale claire du modèle est précisée, permettant de relier directement la perturbation de l'algorithme (quantité d'exclusion du message) à des quantités biologiques. De plus, le modèle est utilisé dans son cadre initial – modéliser les troubles psychiatriques – cette fois-ci avec une vision plus concrète biologiquement, en l'occurrence, en prévoyant avec le modèle d'inférence circulaire des différences expérimentales particulières dans les données d'imagerie cérébrale entre la population schizophrène et non schizophrène. Ensuite, le modèle est utilisé hors de son cadre initial de modélisation de la pathologie, via l'utilisation de ce même modèle pour expliquer des phénomènes sous-optimaux et quasi-optimaux dans la population générale. Ensuite, sur le plan théorique, l'algorithme de Propagation Circulaire de Croyances est étudié, notamment ses propriétés de convergence et sa relation à des algorithmes existants, ce qui permet d'en formaliser une généralisation, plus puissante et tout autant réaliste biologiquement.

Tout d'abord, une proposition d'implémentation neurale claire du modèle est précisée dans le chapitre 4. Cette implémentation neurale comporte deux types d'unités. D'une part, les nœuds de représentation (population de neurones excitateurs et inhibiteurs) encodent la probabilité marginale de la variable mathématique associée au nœud. D'autre part, les nœuds de contrôle (population de neurones excitateurs et inhibiteurs) contrôlent le flux d'information entre les noeuds de représentation ; plus précisément, ces nœuds de contrôle encodent la part d'information redondante devant être soustraite des quantités échangées entre les nœuds de représentation. Cette proposition d'implémentation neurale permet de relier directement la perturbation de l'algorithme (taux d'exclusion du message provenant de B et arrivant en A) au gain synaptique du nœud de contrôle associé à la connexion A vers B, et localisé au niveau du nœud de représentation A. Si les neurones, dans l'implémentation naïve, se déchargent proportionnellement aux quantités qu'ils encodent, ce travail propose également une implémentation utilisant des neurones spikant, donc plus proches de la réalité, pour une représentation neurale plus plausible, même si les moindres détails biologiques n'apparaissent pas tous.

Le modèle de cerveau entier découlant de l'implémentation neurale de l'algorithme PCC (où un nœud du graphe représente une région cérébrale) permet d'expliquer des résultats expérimentaux liés à la schizophrénie ; voir section 2.3. Un premier résultat expérimental est celui des sur-activations relatives aux hallucinations chez les patients schizophrènes, dans des régions spécifiques du cerveau (les zones associatives) qui sont fortement connectées au reste du cerveau et sont donc cruciales pour une transmission efficace de l'information. Dans le modèle neural, les connecteurs du réseau (nœuds fortement connectés servant de relai entre les modules du réseau et au sein des modules), sont également suractivés lorsque la circularité de l'inférence augmente (ce qui correspond à une diminution des taux de correction des boucles). Un deuxième résultat expérimental est celui de la perturbation du réseau de connectivité fonctionnelle dans la schizophrénie, dans le sens d'une plus grande

ségrégation des modules anatomiques. Ce résultat est également prédit par le modèle via une modification du taux de correction des boucles.

Ensuite, le modèle d'inférence circulaire permet de modéliser le phénomène de perception bistable, une manifestation connue de la sous-optimalité du cerveau ; voir section 2.2. Ce phénomène, étudié principalement en laboratoire, correspond à l'alternance spontanée entre deux possibles interprétations d'une image ; par exemple, le cube de Necker (figure 2D représentant un cube 3D en transparence) peut être interprété de deux façons : cube vu d'en haut ou d'en bas. Les caractéristiques de l'alternance entre les possibles interprétations – temps moyen entre chaque alternance – font que le cerveau ne raisonne pas de façon optimale. Le modèle d'inférence circulaire, sous-optimal par nature puisque défini par la perturbation d'un algorithme, modélise des propriétés critiques de ce phénomène de perception bistable. Ces propriétés impliquent les lois de Levelt caractérisant les processus de rivalité binoculaire.

De façon intéressante, la perturbation de l'algorithme PC (lui-même sous-optimal si le graphe a des cycles) définissant l'algorithme PCC peut aller dans le sens d'une amélioration de l'algorithme PC. En effet, il est possible d'exclure plus que nécessaire via un taux de correction supérieur à 1, ce qui crée une rétroaction négative au niveau des cycles de longueur 2. Cela permet de se prémunir des rétroactions positives dues aux cycles du graphe (de longueur 3, 4, 5, …). Plus généralement, il est possible d'apprendre les taux de correction de boucle permettant de contrecarrer les réverbérations naturelles d'information retournant à l'envoyeur. Cette compensation n'est que partielle, mais fonctionne en pratique. Plus précisément, le bon taux de correction permet de réaliser des inférences de qualité remarquable, pour un graphe donné, pour n'importe quels signaux entrant dans le réseau. L'apprentissage peut se faire de manière supervisée, en fournissant des exemples d'entraînement d'inférences exactes afin de trouver les bons taux de correction de boucle ; voir chapitre 3. Par ailleurs, un apprentissage non-supervisé de type Hebbien et homéostatique est proposé pour apprendre les gains synaptiques, des types d'apprentissage qui ont été précédemment proposé pour modéliser la façon dont le cerveau pourrait apprendre les connexions neuronales ; voir section 4.5.

Cette amélioration de l'algorithme PC avec PCC conduit à considérer d'autres algorithmes d'inférence probabiliste approximée ; voir chapitre 3. En effet, l'algorithme PCC est une modification simple de l'algorithme de PC et permet de contrecarrer l'effet des boucles, rendant l'inférence moins sous-optimale. La même idée (retirer la contribution des cycles du graphe) pourrait être utilisée pour améliorer d'autres algorithmes d'inférence probabiliste approximée ayant une forme similaire à l'algorithme PC. L'algorithme de Propagation Circulaire de Croyances (« PCC ») à relié à celui existant de Propagation Fractionnaire de Croyances (« PFC »), très proche conceptuellement. Tous les deux généralisent l'algorithme PC avec des paramètres supplémentaires permettant de réaliser des inférences de meilleure qualité. La généralisation existante de l'algorithme PFC pousse naturellement à une généralisation de PCC. Cet algorithme PCC généralisé comporte des paramètres additionnels correspondant biologiquement aux poids des connexions entre les nœuds de représentation, à la force des signaux entrant dans le graphe, et au gain synaptique des nœuds de représentation. L'algorithme PCC généralisé a de meilleures garanties théoriques et de meilleurs résultats pratiques que PFC ou PCC, et est plus plausible biologiquement que l'algorithme PFC généralisé.

Enfin, l'algorithme PCC, défini initialement dans le cas particulier où de distributions de probabilités sur des variables binaires, est formulé dans le chapitre 5 dans le cas de variables quelconques, discrètes ou continues. Dans le cas particulier où les variables sont gaussiennes, l'algorithme peut être implémenté neuralement très facilement, de manière très similaire au cas où les variables sont binaires, avec simplement deux fois plus de neurones (une sous-population représentant la moyenne de la distribution, et une autre représentant sa variance). Il n'est pas clair dans quelle mesure il existerait une implémentation neurale directe de l'algorithme dans le cas général, c'est-à-dire dans le cas où les variables ne sont ni binaires ni gaussiènnes. Néanmoins, dans ce cas général, et sans se préoccuper d'une possi-

ble implémentation dans le cerveau, l'algorithme PCC (généralisé ou non) peut être utilisé en tant que tel dans la recherche en intelligence artificielle comme alternative à l'algorithme PC, améliorant sa performance dans les taches d'inférence probabiliste (utilisée dans de nombreuses applications réelles). L'apprentissage supervisé permet d'apprendre les paramètres du modèle pour effectivement améliorer la performance, mais ne prend un temps raisonnable que dans le cas d'un graphe de taille modeste. Pour des graphes avec un grand nombre de nœuds, l'apprentissage non supervisé peut être utilisé, permettant d'obtenir de bons résultats également, cette-fois sans avoir besoin de produire des exemples d'entraînement, ce qui rend l'apprentissage utilisable en pratique dans un temps raisonnable.

L'algorithme PCC propose une vision simple du spectre de la psychose, où la quantité de symptômes serait proportionnelle à la perturbation des taux de correction de boucles par rapport à leur valeur optimale. Sur le moyen terme, il serait souhaitable de trouver ces quantités en utilisant uniquement des données d'imagerie cérébrales (via le modèle neural dévelopés dans cette thèse) et par ailleurs uniquement des données comportementales, afin de vérifer les corrélations attendues entre ces paramètres. Des études sont en cours pour caractériser des population infra-cliniques situées sur le spectre de la psychose entre l'optimalité et la pathologie, adhérant par exemple aux théories complotistes ou ayant des croyances anormales inchangeables mais sans hallucinations.

Contents

Chapter 1

Introduction

The confidence that individuals have in their beliefs depends mostly on the quality of the story they can tell about what they see, even if they see little.

— Daniel Kahneman, *Thinking, Fast and Slow*

1.1 General context

The research field of computational neuroscience aims at designing models of the brain and testing them on behavioral and/or brain data in order to understand better our own functioning and imperfections. The brain being an incredibly complex organ (the cerebral cortex contains more than 10 billion neurons) and our experimental techniques being limited, it is simply impossible to model all its details. Research in computational neuroscience is guided by two different classes of approaches to building neural models: mechanistic (bottom-up) versus normative (top-down). Mechanistics approaches attempt to simulate a biological process using a descriptive model of neural circuits with the highest amount of detail possible (anatomical-functional data coming from brain imaging, neuromodulators, etc.) and study how neural computations can give rise to behavior. On the contrary, normative approaches start from the level of behavior and formulate hypotheses concerning the function of a brain area, before potentially proposing how the cortical circuitry might implement the computations necessary to fulfill this function. Altogether, mechanistic and normative approaches each have their pros and cons (Eliasmith and Trujillo, 2014), and together offer complementary views to understand the brain.

More recently, a sub-branch of computational neuroscience has been growing: computational psychiatry (Wang and Krystal, 2014; Stephan and Mathys, 2014; Seriès, 2020). Its goal is to explain the origins and mechanisms of psychiatric symptoms , most of the time by considering first a model of normal functioning, and by altering something in it to account for clinical differences between a psychiatric population and a healthy (that is, non-psychiatric) population. Both mechanistic and normative approaches are used.

In this thesis, I question how probabilistic reasoning is made possible by the brain. More precisely, I consider the Circular Inference model, a normative model of how the brain performs suboptimal inference, previously proposed by Jardri and Denève as a model of schizophrenia and more broadly, of psychosis (Jardri and Denève, 2013a). I develop this model mathematically and question its biological plausibility. I also suggest that this model is not only a good one for psychiatric disorders but also for normal functioning.

1.2 Inference in the brain

1.2.1 Probabilistic reasoning in everyday life

Humans face situations of uncertainty on a daily basis. This includes sensory processing: for instance, we are remarkably good at determining the age or origin of a person based on a simple photograph. More precisely, we are able to make an educated guess, and estimating the uncertainty associated with that guess. Common situations of uncertainty also include motor control: we are able to play the piano or reach the light switch at night, not 100% accurately because making the exact same movement twice is impossible, but approximately, and we are able to estimate our precision very well. Finally, another common example of dealing with uncertainty on a daily life is cognitive reasoning. Should I get vaccinated from COVID-19 given the amount of evidence at hand? How early should I leave my apartment to be on time at my friends' for dinner? Should I invest money in the stock market now or wait?[1]

The fact that people are good at carrying out *probabilistic inferences* - that is, form conclusions or opinions based on known facts or evidence, here probabilistic - has been confirmed in the laboratory. Psychophysical and behavioral experiments suggest that humans perform probabilistic reasoning when perceiving objects (Knill and Richards, 1996; Kersten et al., 2004), moving (Wolpert et al., 1995; Körding and Wolpert, 2004), or reasoning (Tenenbaum et al., 2006; Chater et al., 2006).

This idea that the human brain performs probabilistic reasoning is commonly referred to as the *Bayesian brain* hypothesis, as it relies on Bayes' rule (Bayes, 1763) named after Thomas Bayes. Bayes' rule which states how to determine a conditional probability or posterior probability (e.g., probability of being close to a tree given our senses) given some prior knowledge (e.g., probability of being in a forest) and sensory evidence (e.g., visual or auditory clues). Models originating from Bayes' rule are referred to as *Bayesian models*. Similarly, probabilistic inference is also referred to as *Bayesian inference*.

In fact, not only behavior (meaning in practice, response of subjects to a task) but also neural responses themselves can be analysed in terms of the posterior distribution. The underlying intuition is that the probability of events is represented somehow by neurons, basic components of the brain. Bayesian theories of the brain investigate how probabilistic inference could be carried out in practice (Knill and Pouget, 2004; Doya et al., 2007; Lochmann and Denève, 2011; Pouget et al., 2013); see also section 1.3 for more detail. In order to understand how the brain can perform Bayesian inference, we provide in the following section necessary mathematical definitions which will help build our model.

1.2.2 The problem: Inference in probabilistic graphical models

The object of study is a probability distribution $p(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The distribution can be decomposed into a product of conditionally independent factors that each describes interactions between distinct subsets of variables (Koller and Friedman, 2009; Wainwright and Jordan, 2008):

$$p(\mathbf{x}) \propto \prod_{(i,j)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i) \tag{1.1}$$

where we consider only unitary and pairwise interactions for simplicity. Note, however, that everything can be extended to any factorization of $p(\mathbf{x})$, that is, any Markov random field (MRF) (Kschischang et al., 2001) involving higher-order potentials (e.g. $\psi_{ijk}(x_i, x_j, x_k)$), as explained in section 5.2.

[1] Hopefully, PhD students do not have to face such a dilemma.

As Figure 1.1A shows, the probability distribution can be represented graphically as a graphical model called *factor graph*, composed of variable nodes x_i and factor nodes ψ_{ij} and ψ_i. ψ_i represents prior knowledge about variable x_i, while ψ_{ij} describes the interactions between x_i and x_j. Importantly, ψ_i also represents the potential sensory observations (e.g., noisy measurement of x_i). Indeed, adding local measurements of the form $p(\mathbf{y}|\mathbf{x}) = \prod p(y_i|x_i)$ to the prior model $p(\mathbf{x})$, the posterior distribution becomes $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$. Consequently, introducing sensory observations is equivalent to modifying the potentials $\{\psi_i(x_i)\}$ into $\{\psi_i(x_i)p(y_i|x_i)\}$. In the example of vision, if $x_i \in \{-1, +1\}$ represents the presence (or absence) of tree in the environment of the person, $\psi_i(x_i)$ provides information about whether we should expect trees (low value if the person is a sailor or lives in a desert), and $\psi_{ij}(x_i, x_j)$ represents the relation between x_i and x_j, which is necessary to use knowledge of other variables (x_j) to estimate x_i; for instance, if x_j represents the presence (or absence) of a leaf, then $\psi_{ij}(x_i, x_j)$ describes the fact that the presence of a leaf is often correlated to the presence of a tree and the absence of a leaf is often correlated to the absence of a tree (in case there is are leaves, knowledge about the presence of snow would be helpful as well, for instance).

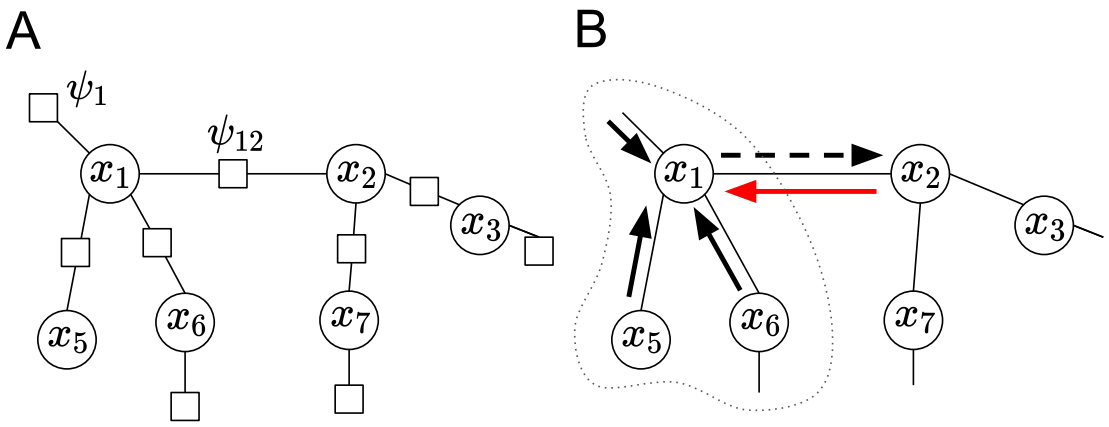

FIGURE 1.1: **Understanding the Belief Propagation (BP) algorithm update equation and the Circular BP impairment**. **(A)** The probability distribution $p(\mathbf{x})$ is represented by a so-called *factor graph* with pairwise potentials ψ_{ij} and unitary potentials ψ_i. Here $p(\mathbf{x}) = \psi_{12}(x_1, x_2)\psi_{15}(x_1, x_5)\psi_{16}(x_1, x_6)\psi_{23}(x_2, x_3)\psi_{27}(x_2, x_7)\psi_1(x_1)\psi_3(x_3)\psi_6(x_6)\psi_7(x_7)$. **(B)** Belief Propagation aims at estimating marginals $p_i(x_i)$ by exchanging messages in the probabilistic graph. The message $m_{1\to2}$ (dotted black line) sent by node x_1 to node x_2 represents the information brought by the subgraph composed of nodes x_1, x_5 and x_6, about x_2 (but not information brought by nodes x_2, x_3 or x_7). This message $m_{1\to2}$ depends on three components (see full black lines). First, all the messages received by node x_1 from its neighbors except x_2. Second, the unitary potential ψ_1 at node x_1. Third, the interaction ψ_{12} between nodes x_1 and x_2. The estimated marginal or belief $b_i(x_i)$ is formed based on all messages sent to node x_i. See Equations (1.3) and (1.4) for more details. Note that in the example taken, the probabilistic graph is an acyclic graph and BP is thus exact. On the contrary, with Circular BP, $m_{1\to2}$ depends on a fourth component: the message $m_{2\to1}$ sent in the other direction (full red line) with weight $1 - \alpha_{12}$; see Equations (1.12) and (1.13) for more details. Because it reverberates the same piece of information (for instance through $x_2 \to x_1 \to x_2$), Circular BP is not an exact algorithm.

Given partial and noisy information $\{\psi_i(x_i)\}$ about $\mathbf{x}$, one might want to compute the marginal probabilities of the distribution p:

$$p_i(x_i) \equiv \sum_{\mathbf{x}\setminus x_i} p(\mathbf{x}) = \sum_{\mathbf{x}\setminus x_i} p(x_1, \ldots, x_n) \tag{1.2}$$

This is the inference problem which is considered in this thesis. Interestingly, the same methods used to obtain marginals $p_i(x_i)$ can be used to compute marginals of groups of variables, e.g., pairwise marginals $p_{ij}(x_i, x_j) \equiv \sum_{\mathbf{x}\setminus\{x_i,x_j\}} p(\mathbf{x})$; see Appendix A.

Unfortunately, direct calculation of the marginals $p_i(x_i)$ can take exponential time in the number of variables n, as the formula above simply suggests to evaluate $p(x_1, \ldots, x_n)$ where for all possible $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for a given x_i (overall 2^{n-1} terms) and sum all these terms. That is why algorithms have been developed, where an *algorithm* is by definition "a procedure for solving a mathematical problem in a finite number of steps, that frequently involves repetition of an operation", to solve this inference problem. Such approximate inference algorithms includes the *Belief Propagation algorithm* and sampling methods. The advantage of these algorithms is to be orders of magnitude faster than the simple summation or terms described above, but their drawback is that they produce only approximate marginal probabilities $b_i(x_i) \approx p_i(x_i)$.

Other inference problems exist besides the marginalization problem. This includes finding the most probable configuration of $\mathbf{x}$ given the observations (known as the maximum a posteriori or MAP problem) $\text{argmax}_{\mathbf{x}}\, p(\mathbf{x})$, or computing the partition function (proportionality constant of Equation (1.1), ensuring that the probability distribution sums to 1). However, we focus in this thesis on the particular inference problem of finding the marginals of $p(\mathbf{x})$.

1.3 Neural implementations of probabilistic inference

Although the idea that humans perform Bayesian inference is now well established, it is much less clear how inference is implemented in practice in human brains at the level of neural circuits. The two important points are how to represent probabilities using neurons, and how to implement probabilistic computations in neural circuits.

The representation problem The first important question is how neural activities possibly encode for probabilities (*representation problem*). For reviews on the topic, see Fiser et al. (2010); Pouget et al. (2013); Aitchison and Lengyel (2017). The most intuitive idea, called *direct variable coding*, is that neurons did not represent probabilities, but instead directly represent quantities that they encode. For instance, neurons would spike more often when their encoded quantity (for instance light intensity) is high; see for example Olshausen and Field (1996). Another form of direct variable coding is sampling theories of the brain (Shi and Griffiths, 2009; Fiser et al., 2010; Berkes et al., 2011; Buesing et al., 2011; Probst et al., 2015; Orbán et al., 2016; Aitchison and Lengyel, 2016; Echeveste et al., 2020), which assume that responses of neurons reflect a sample of a possible value taken by the variable, that is, of the posterior distribution associated to this variable.

However, it is now widely known that it is not necessarily the case, even in sensory areas like the primary visual cortex Knierim and van Essen (1992). *Predictive coding* theories (Rao and Ballard, 1999; Friston and Kiebel, 2009) hypothesize that some neurons represent *prediction errors*, that is, the mismatch between the sensory input conveyed by feedforward connections (from sensory cortical areas) and the prediction of this sensory input conveyed by feedback connections (from areas higher in the cortical hierarchy). Rao and Ballard (1999), as most

articles on the topic, proposes a model with two classes of neurons: some encoding the variable directly, and some other encoding for prediction errors.

An alternative view is that neurons encode for *parameters* of the posterior probability distribution (see Fiser et al. (2010); Raju and Pitkow (2016)), not the variable (or the prediction error on the variable) itself. The most simple example of it is the *probability code* hypothesis, which claims that the firing rate of a neuron is proportional to the posterior probability of a variable, or of some range of this variable; see Lee and Mumford (2003). A highly related representation is the *log-probability code* (see for instance Barlow (1969)), according to which the firing rate is proportional to the log-posterior probability of the variable. An example of it is probabilistic population codes (PPC) (Zemel et al., 1998; Ma et al., 2006), which decomposes the log-probability into a sum of basis functions and proposes that a population of neurons encodes for parameters of these basis functions.

The implementation problem The second problem is how neural circuits perform probabilistic inference (*implementation problem*), once the type of representation has been decided (Denève et al., 2001; Rao, 2006; Ma et al., 2006; Beck et al., 2008; Lochmann and Denève, 2011; Moreno-Bote and Drugowitsch, 2015; Orbán et al., 2016; Spratling, 2016). Each kind of representational code has it advantages and drawbacks, and types inference problems on which they fit naturally, leading to a straightforward neural implementation proposition of these problems. For example, probability codes easily perform sums of probability distributions (which Bayes' rule requires for marginalization) by simply summing neural activities. Similarly, log-probability codes easily multiply probability distributions (which Bayes' rule requires for evidence integration and cue combination) also by summing neural activities. In the case of sampling codes, implementing Markov Chain Monte Carlo (MCMC) methods like Gibbs sampling (Geman and Geman, 1984) also naturally lead to neural implementations of inference. Finally, in both probability codes and log-probability codes, the neural implementation of Bayesian inference is straightforward: the circuit directly performs the probabilistic computations required from a particular inference algorithm.

In particular, message-passing algorithms appear to be natural candidates for the neural implementation of probabilistic inference (Pitkow and Angelaki, 2017; Parr et al., 2019). Indeed, a possibility is that the brain contains networks whose structure mirrors probabilistic graphs and that the network activity corresponds to the message-passing algorithm in the underlying graph (Shon and Rao, 2005; Steimer et al., 2009; Parr and Friston, 2018). Such algorithms involve an exchange of information between nodes of a graph representing the probability distribution, which can be seen as spikes sent between populations of neurons encoding for particular features. Chapter 4 discusses the neural implementation of a particular message-passing algorithm, *Belief Propagation*, and its variant *Circular Belief Propagation*, which are both formally introduced in the next section.

1.4 Belief Propagation and Circular Belief Propagation

We first provide some background on the Belief Propagation algorithm, an approximate inference method aiming at computing approximate marginals $p_i(x_i)$ of a given distribution $p(\mathbf{x})$.

1.4.1 Gentle introduction to Belief Propagation

The Belief Propagation (BP) algorithm or sum-product algorithm (Pearl, 1988) is a message-passing algorithm performing approximate inference on a probabilistic graph or *factor graph*. To do so, BP spreads probabilistic information everywhere in the graph via messages which

are sent between nodes of the graph. The algorithm consists of repeating an update operation on messages. Repeating this update operation allows messages to spread knowledge, which is initially only available at the extremities of the graph through the unitary potentials $\psi_i(x_i)$. For simplicity, we consider here an example of probability distribution whose corresponding probabilistic graph has no cycles (see Figure 1.1A).

BP consists of repeating an update process of messages being exchanged between nodes x_i (see Figure 1.1B) where representing nodes are variables of the distribution $p(\mathbf{x})$. The message from node x_i to node x_j represents probabilistic (i.e., unsure) information about the receiver node x_j given all the local evidence available at node x_i and the relation between variables x_i and x_j. More precisely, the message from node x_i to x_j is computed based on three components. First, the messages collected by x_i from all its neighbors except x_j; indeed, x_i should convey to x_j information unknown to x_j so it does get mixed with new information. Second, the message from x_i to x_j depends on external information ψ_i about the state of x_i. Finally, the message from x_i to x_j depends on the pairwise potential ψ_{ij} describing how x_i and x_j depend on each other; for instance it could encode for the fact that it does not usually rain when the sun is shining, or that flu is associated with fever.

Once the external information has been transmitted to all the nodes of the graph, running the update operation once again would not modify the value of the messages. The algorithm is said to have converged. The marginal probability of x_i is simply obtained by combining all the messages received by node x_i (that is, from all its neighbors, in addition to the external message representing ψ_i).

For a particular probability distribution whose corresponding probabilistic graph has no cycles, as in the example taken in Figure 1.1, the marginal probabilities computed by the BP algorithm are in fact equal to the exact marginals obtained by computing directly $\sum_{\mathbf{x}\setminus x_i} p(x_1, \ldots, x_n)$ (brute force method). The intuitive reason why marginals are exact for acyclic graphs is the following: the message sent from x_i to x_j is the information brought by all the part of the graph connected to x_i, and not only x_i itself: in Figure 1.1B, the message from x_1 to x_2 represents all the information from the subgraph composed x_1, x_5 and x_6 about the state of variable x_2 (see Figure 1.1B).

Although the algorithm was initially designed for probability distributions represented by acyclic graphs, BP can be applied to cyclic graphs by extension, by using the same update operation as for acyclic graphs (Frey and MacKay, 1998). Cycles in a probabilistic graph represent cyclic conditional dependencies between variables (concept A causes concept B which causes concept C, itself causing concept A). In this case, the BP algorithm is sometimes called the *Loopy* Belief Propagation algorithm: indeed, messages travel through loops or cycles of the graph; see also Figure 3.1. The BP algorithm does not necessarily converge on such general graphs, and when it converges it produces incorrect marginals which are more or less accurate depending on the probability distribution (see legend of Figure 3.1). The reason why BP is not exact for cyclic graphs is that the same evidence travels in the network multiple times because of loops, and is mistaken for new evidence (Pearl, 1988). BP has been extensively studied empirically (Murphy et al., 1999; Weiss, 2000; Mooij and Kappen, 2004; Litvak et al., 2009), which has led to a better understanding of the convergence and performance of BP, depending on the cylic graph. It turns out that when (loopy) BP converges, it produces beliefs which are a good approximation of the true marginals. However, when it does not converge, oscillating beliefs can have very little to do with the correct marginals. The convergence (or absence of convergence) of BP in a cyclic graph depends on the graph topology and in particular, the size of the network, the average degree and the size of the cycles (long cycles are better for convergence). But the graph topology is not the only criterion: a given graph structure with different values of ψ_i and ψ_{ij} can lead to BP oscillating or instead converging. The quality of BP is nearly independent of the network size.

Importantly, for sparse cyclic graphs, BP usually performs very well, which allows us to use this algorithm for many concrete applications like computer vision and medical diagnosis. In fact, BP was used in the early 1990s without strong theoretical insights as to why it achieved such good practical results. The BP algorithm admits as a particular case the famous Kalman Filter (Yedidia et al., 2003). Furthermore, BP applied to a particular bipartite graph is equivalent to Low Density Parity Check Codes (Gallager, 1962), an example of error-correcting code (Berrou et al., 1993) used in some Wifi standards and which has been adopted as the fifth generation mobile communication (5G) standards (Sun et al., 2019).

1.4.2 Less gentle introduction to Belief Propagation

We continue this introduction by providing the mathematical definition of the Belief Propagation algorithm. This section requires a stronger mathematical background of the reader.

1.4.2.1 Definition

Belief Propagation is a variational inference method which performs approximate inference on a probabilistic graph. It approximates the marginals of the distribution by making variable nodes x_i share all the probabilistic information available with the rest of the network. It does this by sending messages to other variable nodes. These messages represent all probabilistic information (observed variables, prior distribution over variables) brought from a part of the network to node x_j. The algorithm consists of running iteratively the following update message equation on the graph, where we consider here pairwise factor graphs or Markov networks:[2]

$$m_{i\to j}^{\text{new}}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}^{\text{old}}(x_i) \tag{1.3}$$

where $\mathcal{N}(i)$ is the set of neighbors of node x_i in the graph. Messages are for instance initialized uniformly over the nodes ($m_{i\to j}(x_j) = 1/\mathcal{N}(j)$). Once messages have converged, approximate marginal probabilities (or *beliefs*) are computed as:

$$b_i(x_i) \propto \psi_i(x_i) \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i) \tag{1.4}$$

Note the form of the update in Equation (1.3), which explains the alternative name of Belief Propagation: the sum-product algorithm. An crucial feature of the BP algorithm is the message exclusion principle ($k \in \mathcal{N}(i) \setminus j$): to compute $m_{i\to j}$, all messages coming to node x_i are taken into account and combined, except the message in the opposite direction $m_{j\to i}$.

As stated above, (Loopy) BP has initially been used because of its empirical performance, but lacked a deep theoretical comprehension as to why the algorithm worked so well in practice. However, in the early 2000s, a theoretical foundation of BP emerged (Yedidia et al., 2001; Heskes, 2002) based upon the approximation of the variational distribution $b(\mathbf{x}) \approx p(\mathbf{x})$ as if its associated probabilistic graph were acyclic (*Bethe approximation*), as further explained in the following section. This initial work was subsequently completed with various results on the convergence properties of BP (Tatikonda and Jordan, 2002; Tatikonda, 2003; Ihler et al., 2005; Mooij and Kappen, 2005, 2007b; Knoll and Pernkopf, 2017; Leisenberger et al., 2021), the goodness of BP through the estimation of the BP error (Wainwright et al., 2003; Taga and Mase, 2006; Ihler, 2007; Mooij and Kappen, 2009; Shi et al., 2010), and the properties of the Bethe free energy and

[2] For a more mathematically accessible version of the message update equation, refer to the binary case described in section 1.4.4 and particularly to Equation 1.14.

Algorithm 1 Belief Propagation algorithm in a pairwise factor graph

for all directed edges $i \to j$ **do**
  $m_{i\to j}(x_j) \leftarrow$ some distribution {Initialize the messages}
end for
repeat
  for all directed edges $x_i \to x_j$ **do**
    $m^{\text{new}}_{i\to j}(x_j) \leftarrow \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i)$ {Update the messages}
  end for
  $m \leftarrow m^{\text{new}}$
until convergence
for all nodes x_i **do**
  $b_i(x_i) \leftarrow \psi_i(x_i) \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i)$ {Compute the beliefs}
end for

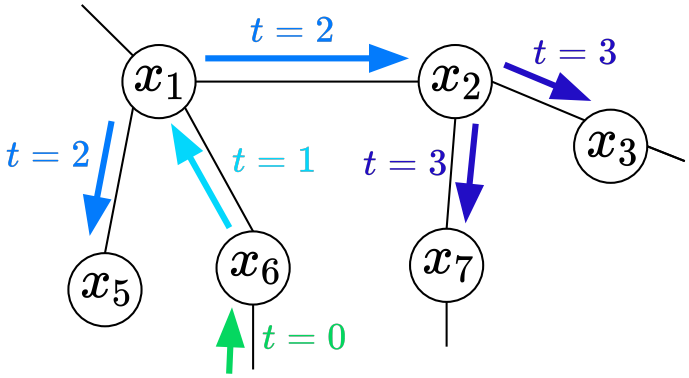

Figure 1.2: **Propagation of probabilistic information by the Belief Propagation algorithm**. The graph is the same as in Figure 1.1. The BP algorithm consists of running the same update equation on messages until convergence of the messages. Initially, information is only available locally at the nodes through unitary potentials ψ_i (external information about variable x_i). In an acyclic graph, the information propagates into the network without "coming back", because of the absence of cycles.

its fixed points (Heskes et al., 2002; Heskes, 2004; Watanabe and Fukumizu, 2009; Watanabe, 2011; Weller et al., 2014; Knoll and Pernkopf, 2017).

1.4.2.2 Theoretical idea behind Belief Propagation

Here we provide the theoretical background underlying the Belief Propagation algorithm defined by Equations (1.3) and (1.4), respectively message update equation and expression the belief (approximate marginal probability).

Problem We consider here a probability distribution $p(\mathbf{x})$ composed of pairwise factors, with continuous and/or discrete variables $x_1, x_2, \ldots, x_n$. However, for completeness we also consider the general case (any probability distribution written as a Markov random field) in section 5.2,

where we state and demonstrate message update equations of BP applied to a general factor graph.

First, we start by writing the general variational problem: given the true probability distribution $p(\mathbf{x})$, we want to find an approximating probability distribution $b(\mathbf{x})$, whose marginals are easier to compute than those of $p(\mathbf{x})$. In practice, in order to find $b(\mathbf{x})$, we want to minimize the Kullback–Leibler divergence between $p(\mathbf{x})$ and $b(\mathbf{x})$:

$$D_{KL}(b\|p) = \sum_{\mathbf{x}} b(\mathbf{x}) \log\left(\frac{b(\mathbf{x})}{p(\mathbf{x})}\right) \tag{1.5}$$

Minimizing $D_{KL}(b\|p)$ will result in a probability distribution $b(\mathbf{x})$ "close" to the original distribution $p(\mathbf{x})$.

The true probability distribution $p(\mathbf{x})$ is written $p(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z}$ where Z is a normalization constant ensuring that $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$.

$$\implies D_{KL}(b\|p) = \sum_{\mathbf{x}} b(\mathbf{x}) \log(b(\mathbf{x})) + \log(Z) + \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x}) \tag{1.6}$$

which can also be written

$$D_{KL}(b\|p) = -S_b - F_b + U_b \tag{1.7}$$

where $S_b = -\sum_{\mathbf{x}} b(\mathbf{x}) \log(b(\mathbf{x}))$ is the *variational entropy*, $F = -\log(Z)$ is the *Helmholtz free energy*, and $U_b = \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x})$ is the *variational average energy*. The *Gibbs free energy* (also called the *variational free energy*) is defined as $G = U_b - S_b$, which we want to minimize in order to minimize the KL divergence.

However, given the probability distribution $p(\mathbf{x})$ and the variational distribution $b(\mathbf{x})$, it can be tricky to compute the entropy term S_b. This is contrary to U_b, which can be decomposed very easily:

$$\begin{aligned} U_b &= \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x}) \\ &= -\sum_{\mathbf{x}} b(\mathbf{x}) \sum_{i,j} \psi_{ij}(x_i, x_j) - \sum_{x} b(\mathbf{x}) \sum_{i} \psi_i(x_i) \\ &= -\sum_{i,j} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \psi_{ij}(x_i, x_j) - \sum_{i} \sum_{x_i} b_i(x_i) \psi_i(x_i) \end{aligned} \tag{1.8}$$

Proposed solution As it is not possible to easily compute S_b (the entropy of $b(\mathbf{x})$), Belief Propagation estimates it as if the factor graph representing $b(x)$ were a tree (i.e., were acyclic). This means that:

$$b(\mathbf{x}) \approx \prod_{i,j} \frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \prod_i b_i(x_i) \tag{1.9}$$

where $b_{ij}(x_i, x_j) \equiv \sum_{\mathbf{x} \setminus (x_1, x_2)} b(\mathbf{x})$ and $b_i(x_i) \equiv \sum_{\mathbf{x} \setminus x_i} b(\mathbf{x})$. This is equivalent to approximating the entropy S_b of $b(\mathbf{x})$ as follows:

$$\begin{aligned} -S_b &= \sum_{\mathbf{x}} b(\mathbf{x}) \log(b(\mathbf{x})) \\ &\approx \sum_{x} b(x) \sum_{(i,j)} \log\left(\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)}\right) + \sum_{x} b(x) \sum_{i} \log(b_i(x_i)) \\ &\approx \sum_{(i,j)} \sum_{(x_i, x_j)} b_{ij}(x_i, x_j) \log\left(\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)}\right) + \sum_{i} \sum_{x_i} b_i(x_i) \log(b_i(x_i)) \end{aligned} \tag{1.10}$$

which gives the following approximation of the Gibbs Free Energy, known as the *Bethe Free Energy*: $G \approx G_{\text{Bethe}}$, where:

$$\begin{aligned} G_{\text{Bethe}} = &\sum_{(i,j)} \sum_{(x_i, x_j)} b_{ij}(x_i, x_j) \log\left(\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)}\right) - \sum_{(i,j)} \sum_{(x_i, x_j)} b_{ij}(x_i, x_j) \log\left(\psi_{ij}(x_i, x_j)\right) \\ &+ \sum_{i} \sum_{x_i} b_i(x_i) \log\left(b_i(x_i)\right) - \sum_{i} \sum_{x_i} b_i(x_i) \log\left(\psi_i(x_i)\right) \end{aligned} \tag{1.11}$$

The Belief Propagation algorithm thus consists of minimizing the Bethe Free Energy G_{Bethe} to find some probability distribution $b(\mathbf{x})$ and eventually, the marginals of $b(\mathbf{x})$ which are hypothesized to be close to the marginals of the initial distribution $p(\mathbf{x})$.

From Gibbs free energy approximation to messages Appendix A shows that the minimization of the Bethe Free Energy given in Equation (1.11) leads to the BP algorithm shown in Algorithm 1. The messages $m_{i \to j}$ are related to the Lagrange multipliers of the constrained optimization problem. Indeed, the goal of the algorithm, as seen above, is to try and to minimize the Bethe Free Energy under the constraints $\sum_{x_i} b_i(x_i) = 1$ and $\sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j)$ (as $b(\mathbf{x})$ is a probability distribution). The message update equation in the Belief Propagation algorithm is not a gradient descent procedure on the Bethe Free Energy, but instead a fixed-point equation. In other words, the fixed points of BP correspond to stationary points of the Bethe free energy (Yedidia et al., 2003). In fact, it was later shown that stable fixed points of BP are minima of the Bethe Free Energy (Heskes et al., 2002). In the case where the system has one fixed point and this fixed point is stable, the BP procedure is guaranteed to converge to the unique fixed point, that is, to the global minimum of the Bethe Free Energy. However, in the general case, BP might converge only to a local optimum of the Bethe free energy, or not converge at all.

Note that in the case where the initial distribution can be represented by a tree, Equation (1.9) is not an approximation, which explains why BP is exact when applied to acyclic probabilistic graphs.

1.4.3 Circular Belief Propagation

Building on the previous section, we provide here a definition of the Circular Belief Propagation algorithm.

1.4.3.1 Definition of Circular BP

The Circular BP algorithm was initially defined in Jardri and Denève (2013a) as an extension to the BP algorithm. Its message update equation is the same as the one for BP, with the exception

of parameter $\alpha_{i\to j}$, the *loop correction factor*:[3]

$$m_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i,x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{i\to j}} \tag{1.12}$$

which, in the case $\boldsymbol{\alpha} = \mathbf{1}$, is BP (see Equation (1.3)). The only difference with BP is the final term $m_{j\to i}(x_i)^{1-\alpha_{i\to j}}$, meaning that the message in the opposite direction $m_{j\to i}$ is partly taken into account for the computation of $m_{i\to j}$. Beliefs are computed the same way as for BP:

$$b_i(x_i) \propto \psi_i(x_i) \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i) \tag{1.13}$$

Note that $\alpha_{i\to j}$ is assigned to a *directed* edge (i,j). However, in most of this thesis, we will consider $\boldsymbol{\alpha}$ to be a symmetric matrix: $\alpha_{i\to j} = \alpha_{j\to i}$, which will be written α_{ij} in this case (α_{ij} is assigned to the *undirected* edge (i,j)).

For a definition of Circular BP applied on any factor graph (not necessarily pairwise as considered throughout this thesis), see section 5.2.

Algorithm 2 Circular Belief Propagation algorithm in a pairwise factor graph

for all directed edges $i \to j$ **do**
  $m_{i\to j}(x_j) \leftarrow$ some distribution {Initialize the messages}
end for
repeat
  for all directed edges $x_i \to x_j$ **do**
    $m^{\text{new}}_{i\to j}(x_j) \leftarrow \sum_{x_i} \psi_{ij}(x_i,x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}}$ {Update the messages}
  end for
  $m \leftarrow m^{\text{new}}$
until convergence
for all nodes x_i **do**
  $b_i(x_i) \leftarrow \psi_i(x_i) \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i)$ {Compute the beliefs}
end for

1.4.3.2 Theoretical motivation behind Circular Belief Propagation

The Circular BP algorithm was defined in Jardri and Denève (2013a) with the goal of modeling aberrant beliefs in general, and *psychosis* in particular (within a model called the "Circular Inference" model). Psychosis is a mental disorder defined by a loss of contact with reality characterized by seeing or hearing things that do not exist (*hallucinations*), or believing in things in a unshakable manner despite being contradicted by rational arguments (*delusions*). See more about the link between Circular BP and psychosis in section 1.5.

The theoretical intuition behind the change appearing in Equation (1.12) is the following (for the biological intuition behind the introduction of parameter α, seen as a level of excitation-inhibition imbalance, see section 1.6). Instead of having node x_i only send to node x_j all information it collected from its neighbors (except, of course, x_j) as in BP, node x_i also sends information coming from x_j ($m_{j\to i}$) "weighted" by some factor $1-\alpha_{ij}$ (power weight in the

[3] For a more mathematically accessible version of the equation in the binary case, refer to Equation 1.18.

initial formulation, multiplicative weight in the log-domain as shown in the following section) which is called the level of *circularity*. If the circularity $1 - \alpha_{ij} \neq 0$, then node x_j receives from node x_i some information that it already knew (as it previously sent it to node x_i via $m_{j\to i}$), in addition to information collected by node x_i from its other neighbors. The message sent from x_i to x_j thus comes back to x_i.

In other words, probabilistic information, instead of being properly spread out throughout the network, is being reverberated and treated multiple times. As shown in Figure 1.4, this generally leads beliefs to be overconfident, meaning that the approximate marginals are closer to the extreme values (0% and 100%) than they should be if we were to apply exact inference.

1.4.4 BP and Circular BP for binary distributions

1.4.4.1 BP for binary distributions

Throughout this thesis (with the exception of chapter 5)), the probability distributions $p(\boldsymbol{x})$ on which we would like to perform inference are assumed to be distributions over binary variables: $x_i \in \{-1, +1\}$. In this case, BP, defined in Equation (1.3), takes a very simple form in the log-domain (the proof is left to the reader):

$$M_{i\to j}^{\text{new}} = f_{ij}\left(\sum_{k\in\mathcal{N}(i)\setminus j} M_{k\to i} + M_{\text{ext}\to i}\right) \tag{1.14}$$

where $M_{i\to j} \equiv \frac{1}{2}\log\left(\frac{m_{i\to j}(x_j=+1)}{m_{i\to j}(x_j=-1)}\right)$ represents the information about variable x_j brought by variable x_i. $M_{\text{ext}\to i} \equiv \frac{1}{2}\log\left(\frac{\psi_i(x_i=+1)}{\psi_i(x_i=-1)}\right)$ represents alternatively prior information over a variable x_i, or here a noisy sensory observation providing information about the variable (e.g. auditory and visual sense; *ext* stands for "external"). f_{ij} is a sigmoidal function given in the general case by:

$$f_{ij}(x) = \frac{1}{2}\log\left(\frac{\psi_{ij}^{1,1}e^{2x} + \psi_{ij}^{1,0}}{\psi_{ij}^{0,1}e^{2x} + \psi_{ij}^{0,0}}\right) \tag{1.15}$$

where $\psi_{ij}^{x_i,x_j}$ is simply a notation for $\psi_{ij}(x_i, x_j)$.

However, in simulations, we consider a particular class of distributions over binary variables called the *Ising models* (also known as *Boltzmann machines*) from statistical physics (Ising, 1925; Baxter, 1982). In Ising models, pairwise factors take a specific form: $\psi_{ij}(x_i, x_j) \propto \exp(J_{ij}x_ix_j)$. Function f_{ij} then takes a simple form (see also Mooij and Kappen (2007b)):

$$f_{ij}(x) = \phi^{-1}\big(\phi(J_{ij})\phi(x)\big) \tag{1.16}$$

where ϕ is the hyperbolic tangent tanh. f_{ij} is a sigmoidal function controlled by parameter J_{ij}; see Figure 1.3B. Function f_{ij} is different from identity because there is some bounded "trust" between nodes x_i and x_j, which comes from the parameter J_{ij} being bounded.

Equation (1.14) only involves log-messages M (after convergence of the messages, beliefs are computed according to Equation (1.13)). This equation can be rewritten using log-messages M and log-beliefs B:

$$\begin{cases} M_{i\to j}^{\text{new}} = f_{ij}(B_i - M_{j\to i}) & \text{(1.17a)} \\ B_i = \displaystyle\sum_{j\in\mathcal{N}(i)} M_{j\to i} + M_{\text{ext}\to i} & \text{(1.17b)} \end{cases}$$

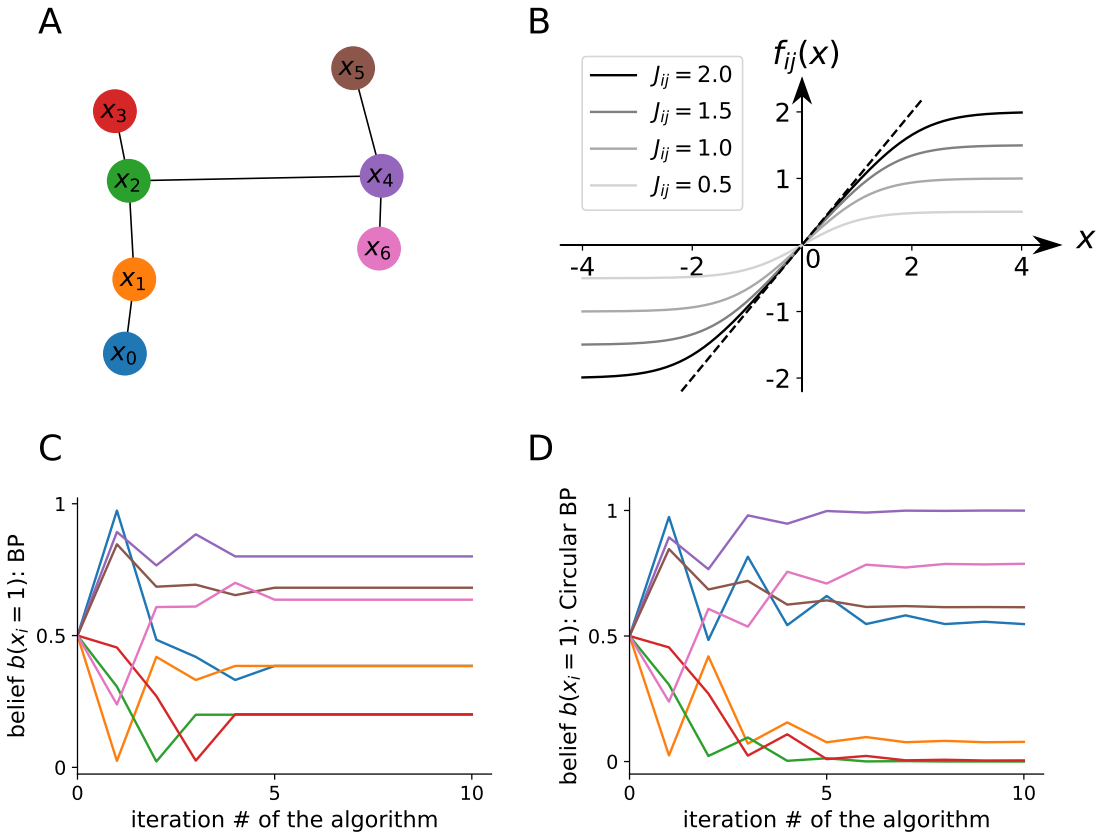

FIGURE 1.3: **Running the Belief Propagation (BP) algorithm and its variant, the Circular Belief Propagation (CBP) algorithm**. **(A)** Example of acyclic graph taken for the simulation. In our example, the probability distribution corresponds to an Ising model: pairwise potentials $\psi_{ij}(x_i, x_j) \propto \exp(J_{ij} x_i x_j)$ and unitary potentials $\psi_i(x_i) \propto \exp(M_{\text{ext}\to i} x_i)$ where $x_i \in \{-1; +1\}$ (binary case). J_{ij} is generated randomly ($\sim \mathcal{N}(0, 3)$), as well as $M_{\text{ext}\to i}$ ($\sim \mathcal{N}(0, 2)$). **(B)** The update function f_{ij} for both BP and CBP (see Equations (1.17a) and (1.19a)) is a parametric sigmoidal function close to the hyperbolic tangent. The parameter J_{ij} represents the level of trust between variables x_i and x_j. **(C)** Belief Propagation is a message-passing algorithm which consists of running the update equation (1.17a) until convergence of the messages. The approximate marginals, or beliefs, are defined by Equation (1.17b). Here the beliefs found by BP are exact as the graph has no cycles. **(D)** The Circular Belief Propagation algorithm is a parametric form of the Belief Propagation algorithm with parameter α_{ij} assigned to each edge (i, j) of the graph. It is identical to BP for $\boldsymbol{\alpha} = \mathbf{1}$. In the simulation, $\boldsymbol{\alpha}$ is taken uniformly over the edges, equal to 0.5.

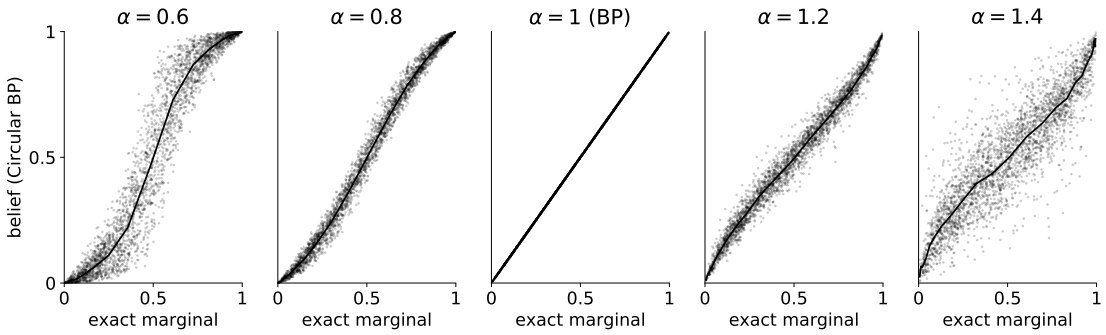

FIGURE 1.4: **Circular Belief Propagation ($\alpha < 1$) produces overconfident beliefs in acyclic probabilistic graphs, whereas Belief Propagation ($\alpha = 1$) is exact**. In this example, weights of the (acyclic) graph J_{ij} are taken as positive, and Circular BP is applied to the graph with $\boldsymbol{\alpha}$ taken uniformly over the edges. Circular BP was initially developed with the idea of $\alpha < 1$ which i to an imbalance between excitation and inhibition in favor of excitation causing an amplification of information, also known as double-counting, and eventually leading to overconfident beliefs (or marginals). On the contrary, BP would correspond to a perfect balance between excitation and inhibition. For information, we also picture here the situation $\alpha > 1$ for which the network becomes underconfident. The graph was randomly generated and randomly weighted. One point corresponds to the approximate marginal under Circular BP versus exact marginal, for a given node and example. The full line represents the average of all points.

where $B_i \equiv \frac{1}{2}\log\left(\frac{b_i(x_i=+1)}{b_i(x_i=-1)}\right)$ is by definition half of the *log odds* (*odds* is a synonym for likelihood ratio or probability ratio). The approximate marginal probabilities are given by $b_i(x_i = \pm 1) = \sigma(\pm 2B_i)$, i.e., $b_i(x_i) \propto \exp(B_i x_i)$.

Equation (1.17a) means that node x_i sends to x_j everything it knows (B_i) except what x_j communicated to x_i ($M_{j\to i}$). In the case of acyclic graphs, this strategy is optimal to spread information in the network, that is, without self-reinforcement of beliefs (case where some information would be communicated from x_i to x_j, then back from x_j to x_i, …). This intuitively explains why BP is exact in this case.

However, BP can perform poorly in cyclic graphs (Murphy et al., 1999; Weiss, 2000); see chapter 3.

1.4.4.2 Circular BP for binary distributions

Similarly to BP, the Circular BP algorithm (Jardri and Denève, 2013a) is written in the binary case very simply:

$$M_{i\to j}^{\text{new}} = f_{ij}\left(\sum_{k\in\mathcal{N}(i)\setminus j} M_{k\to i} + (1-\alpha_{ij})M_{j\to i} + M_{\text{ext}\to i}\right) \tag{1.18}$$

or equivalently,

$$\begin{cases} M_{i\to j}^{\text{new}} = f_{ij}(B_i - \alpha_{ij}M_{j\to i}) & \text{(1.19a)} \\ B_i = \sum_{j\in\mathcal{N}(i)} M_{j\to i} + M_{\text{ext}\to i} & \text{(1.19b)} \end{cases}$$

where

$$f_{ij}(x) = \frac{1}{2}\log\left(\frac{\psi_{ij}^{1,1}e^{2x} + \psi_{ij}^{1,0}}{\psi_{ij}^{0,1}e^{2x} + \psi_{ij}^{0,0}}\right) \tag{1.20}$$

in the general case, and

$$f_{ij}(x) = \phi^{-1}\big(\phi(J_{ij})\phi(x)\big) \tag{1.21}$$

in the specific case of Ising models for which $\psi_{ij}(x_i, x_j) \propto \exp(J_{ij}x_i x_j)$. Function f_{ij} for Circular BP has the same expression as function f_{ij} used in BP. The only difference between Circular BP and BP is that Circular BP has an additional parameter $\boldsymbol{\alpha}$ in Equation (1.19a). BP corresponds to the particular case $\boldsymbol{\alpha} = \mathbf{1}$. Note that we considered the special case $\alpha_{i\to j} = \alpha_{j\to i} \equiv \alpha_{ij}$ above (i.e., we assumed that $\boldsymbol{\alpha}$ is a symmetric matrix).

1.5 Circular BP as model of impaired behavior

The Circular Belief Propagation algorithm was proposed in Jardri and Denève (2013a) as way of modeling aberrant percepts, overconfidence in patients with schizophrenia in probabilistic tasks and resistance to illusions, as well as the learning of wrong associations and of false beliefs (the resulting model is called the "Circular Inference model"). This leads us in the next paragraph to describe what schizophrenia (SCZ) is.

Schizophrenia and the psychosis continuum Schizophrenia a serious mental illness that interferes with a person's ability to think clearly, manage emotions, make decisions and relate to others (definition from the National Alliance on Mental Illness). The disorder generally appears at the beginning of adult life, and affects around 1% of the world's population during their lifetime. It is characterized by a variety of symptoms, which have been categorized into three clinical dimensions: *positive symptoms*, *negative symptoms* and *disorganization*. People with disorganization are people who struggle to remember things, organize their thoughts or complete tasks. Negative symptoms stand for an absence of normal functions; examples include apathy, lack of motivation and interest, and displaying little feeling in emotional contexts. Finally, the positive symptoms correspond to thinking or behavior that the person with schizophrenia did not have before becoming ill. This includes *hallucinations* (seeing things or smelling things that others cannot perceive, and most commonly, hearing voices) and *delusions* (false beliefs which do not change, even when new ideas or facts are presented to the person - most commonly paranoid thoughts). Contrary to the other clinical dimensions, which take place constantly, positive symptoms occur intermittently, producing a state called *psychosis*, also known as a *psychotic episode*.

However, diagnosing schizophrenia is not straightforward. The diagnosis is currently defined on a pure clinical basis (that is, based on the person's report of symptoms and not neural data, for instance), in the absence of reliable biological markers. This leaves a part of subjectivity and therefore leads to a certain inter-rater variability: different diagnoses can potentially be made by different psychiatrists for the same individual. People diagnosed with schizophrenia have a strong clinical heterogeneity by nature. Indeed, when based on categorical disease classifications, e.g., International Classification of Diseases (ICD) or Diagnostic and Statistical Manual of Mental Disorders (DSM), schizophrenia is characterized by the conjunction of psychiatric symptoms (e.g. delusions, hallucinations, disorganization, etc.), none of them being specific, but easing clinical diagnosis by defining a threshold on the psychosis spectrum (see Figure 1.5). In particular, although the most specific symptom of schizophrenia is auditory hallucinations (followed by delusions), voice hearers are not necessarily diagnosed with schizophrenia; see the Hearing Voices

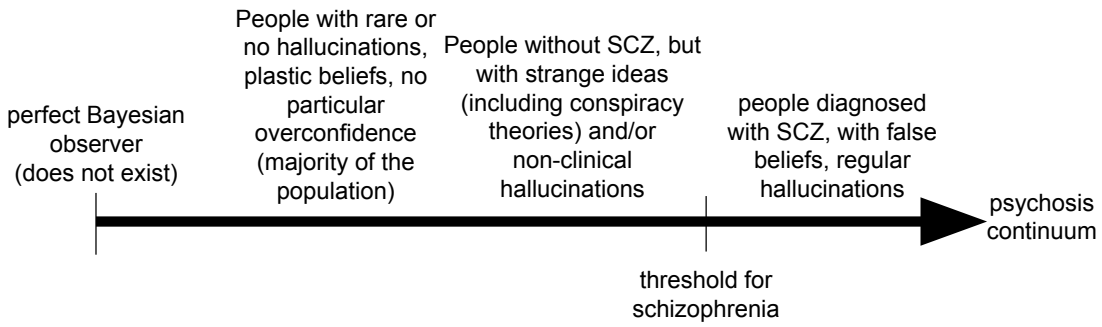

Figure 1.5: **Illustration of the idea of psychosis continuum or psychosis spectrum**. SCZ stands for schizophrenia.

community. On the contrary, patients diagnosed with schizophrenia do not necessarily hear voices but may instead experience non-sensory delusions such as persecutory beliefs.

In addition to the clinical symptoms described above, the behavior of people with schizophrenia differs in many ways from people without. A common experimental result differentiating people with SCZ from others is the *jumping-to-conclusions* effect. This effect is defined by the tendency to make decisions more quickly, on the basis of less evidence, and with more confidence. The jumping-to-conclusion effect has been associated with the presence of delusions (Huq et al., 1988; Garety et al., 1991; Moritz and Woodward, 2005; Speechley et al., 2010) and is quite logically frequently observed in SCZ and paranoia (Averbeck et al., 2011; Evans et al., 2015). Note here that the confidence in ones' beliefs does not directly relate to the accuracy of the belief: one might have the correct belief (e.g., option A is more probable than option B) given all the evidence at hand, but be more/less confident than what the evidence truly suggests.

Treating SCZ requires a multi-axis strategy, gathering medications, cognitive therapy and remediations, social supports, etc. Despite a recognized effectiveness, a significant number of patients who undergo treatment may still experience some symptoms (including psychotic episodes), even if these symptoms are less severe or frequent. That is why the comprehension of schizophrenia through basic science and theoretical research needs to advance, to better detect the disease, allow living better with this complex illness, and propose better treatments, ideally curing it completely.

Circular Inference as model of the mental health continuum As stated above, Circular Inference (based on the Circular BP algorithm) is a model of schizophrenia and psychosis. (Jardri and Denève, 2013a) introduces Circular BP and proposes that this algorithm is used to perform inference in the brain. Jardri and Denève also provide a sketchy idea of how a network of excitatory and inhibitory populations might carry out this algorithm, without proposing a neural implementation per se. The algorithm, as shown in Figure 1.4, causes overconfidence in the beliefs for $\alpha < 1$. The *jumping-to-conclusions* phenomenon is interpreted as a misinterpretation of nearly uninformative evidence (noise or non-significant information) because of the amplification of the signal. *Disorganization* is explained by the strong temporal oscillations arising in the algorithm in the event of contradicting evidence, where the beliefs alternate between opposite conclusions as the algorithm runs (before a very slow convergence). Furthermore, the article tackles the problem of learning of causal links (strength of relationships between variables) through a Hebbian-like learning rule, yet without precise neural interpretation. As it relies on an impaired

inference, the learning is also affected. The article shows how relationships between concepts can be learned despite being unrelated in reality, which is seen as the formation of a delusional idea. Finally, the authors show how reverberating specifically the sensory information or the prior knowledge allows for even more flexibility in the type of false inferences performed by the network, as a result of an imbalance between the effective "weighting" of the prior versus of the sensory evidence. *Hallucinations* and resistance to certain perceptual illusions, both observed in schizophrenia (Notredame et al., 2014), are an example of such false inferences. More specifically, Jardri and Denève propose that positive symptoms of schizophrenia originate from the reverberation of sensory evidence (which amounts to giving more weight to the sensory evidence with respect to the prior). Denève and Jardri (2016) develops further the link between the Circular Inference model and experimental evidence, with a focus on the notion of hierarchy (in neural circuits and in the associated generative model). For instance, the increased activation (measure by the BOLD signal in fMRI) observed during hallucinations in specific brain regions called *association sensory cortices* was related to the amount of belief update at the nodes located at the middle of the hierarchy during a simulated "hallucination" in the network. Additionally, Jardri et al. (2017) tests experimentally the ideas behind the Circular Inference model, that is, the corruption of the sensory evidence by the prior knowledge (or the reverse, or both). The paper models the confidence levels reported by people with and without schizophrenia in a probabilistic task (variant of the "beads task" (Huq et al., 1988)) which consists of combining two kinds of probabilistic information thought of prior and sensory evidence. To do so, a proxy equation, which qualitatively corresponds to the idea behind Circular BP (but is different from the Circular BP algorithm), is put forward. Fitting the model to the reported beliefs shows a correlation between the amount of ascending loops (controlling for the reverberation of the sensory evidence information) and the severity of the positive symptoms. Unexpectedly (because the Circular Inference was initially designed to account for positive symptoms of schizophrenia), the amount of descending loops also correlates with the severity of the negative symptoms, and the total amount of loops correlates with the severity of disorganized symptoms. Importantly, the model captures inter-participant variability and not only group averages. A predictive coding-like model involving modified effective weights to the prior or sensory evidence could not explain, contrary to Circular BP, the jumping-to-conclusions effect observed in the confidence levels even for small amounts of sensory evidence. Overall, the article provides an interesting link between the Circular Inference model and the entirety of the symptoms types observed in schizophrenia. More recently, Simonsen et al. (2021) (see also its commentary Seriès (2021)) uses Circular Inference to model a probabilistic task involving social cognition. Precisely, the authors used the same proxy equation as in Jardri et al. (2017) to model a social version of the beads task, where participants had to report their level of confidence based on some "sensory evidence" and the confidence levels of other people. Similarly to the results of Jardri et al. (2017), the amount of sensory loops correlated with the amount of positive symptoms (including hallucinations, delusions, and bizarre behavior). Additionally, contrary to Jardri et al. (2017), the amount of sensory (and not prior) loops correlated with the amount of negative symptoms (including anhedonia-asociality, avolition-apathy, and attention).

Furthermore, Circular Inference could model pathologically-induced psychosis and among others, the effects of the ketamine drug, a non-competitive NMDA receptor antagonist which has been proposed as pharmacological model of schizophrenia, as its produces behavioral and cognitive disturbances that are consistent with symptoms of schizophrenia (Krystal et al., 1994; Adler et al., 1999; Newcomer et al., 1999) and has been reliably associated with delusional thinking Corlett et al. (2006). Note that Bayesian models have already been used to account for the effects of ketamine (Corlett et al., 2007, 2016; Salvador et al., 2020). By extension, the Circular Inference model could be applied to the anti-NMDAR encephalitis auto-immune

disease, which affects NMDA receptors like ketamine and is accompanied by symptoms such as initial psychosis and long-lasting memory deficits, which are also features of schizophrenia (Kayser and Dalmau, 2016). Likewise, Bayesian models have been used to understand the effects of anti-NMDAR encephalitis (Stein et al., 2020).

However, the Circular Inference model is not specific to schizophrenia and psychosis: it could explain some suboptimal choices made by humans and animals in the general case (where *optimality* is defined with respect to the laws of probability) Intriguingly, people without schizophrenia also show signs of circular inference in the probabilistic tasks of Jardri et al. (2017) and Simonsen et al. (2021), although in lower quantities than in people with schizophrenia. This suggests the presence of circular reasoning in the general population. However, Jardri et al. (2017) reports an absence of significant correlation between the non-clinical delusional score (measured by the PDI scale (Peters et al., 2004)) and the amount of loops in the group of people without schizophrenia.

Related to the last point, Circular Inference could be a useful model for the formation and maintenance of conspiracy beliefs and similarly, of persecutory beliefs or unshakable strange beliefs in general (work in progress). Such beliefs may be observed in people exhibiting some vulnerabilities to uncertainty while being below the threshold for the diagnosis of schizophrenia. This "vulnerability to uncertainty" can be revealed in probabilistic tasks where uncertainty is high: the Circular Inference model indeed predicts that in the face of ambiguity, differences between perfect Bayesian observers ($\alpha = 1$) and people suffering from psychosis ($\alpha < 1$) are maximal. Examples of such tasks can be found in bistable perception experiments, one of which is modeled using Circular BP in section 2.2.

Moreover, the Circular Inference model could help understand the effects of psychedelic drugs. These effects, which differ from but remind the symptoms of schizophrenia, include distorted perception, visual and multimodal hallucinations and synaestesia. In fact, the Circular Belief Propagation algorithm provides an account of "classic psychedelics" (serotonergic agonists like DMT, LSD, and psilocybin), by showing that descending loops give rise to crossmodal percepts and stronger illusions (Leptourgos et al., 2021). This contrasts with ascending loops explaining jumping to conclusions, unimodal hallucinations and reduced vulnerability to illusions as observed in schizophrenia.

Finally, the Circular Inference model can potentially be applied to other diseases or neurodevelopmental disorders. One example is the autism spectrum disorder, which shares certain symptoms with schizophrenia, might be explained using similar biological arguments (e.g., the excitatory-inhibitory imbalance (Rubenstein and Merzenich, 2003)), and are often modeled similarly (Lanillos et al., 2020). A recent study (Chrysaitis et al., 2021) tests the presence of circular reasoning in autism by using the same probabilistic task as in Jardri et al. (2017). The modeling did not show any sign of circular inferences in the behavioural data, which (of course) does not mean that this link between the Circular Inference model and autism should not be investigated further. Other psychiatric disorders have been related to the idea of imbalance between excitation and inhibition, including attention-deficit hyperactivity disorder (ADHD) (Karch et al., 2012), depression (Luscher et al., 2011) and bipolar disorders (Sakai et al., 2008), and therefore could also possibly be explained using this model. Lastly, epsilepsy is not a psychiatric disorder but has consistently been related to impaired inhibitory function (Treiman, 2001), and people with epsilepsy have an increased risk of schizophrenia and of schizophrenia-like psychosis (Cascella et al., 2009). This opens the question of whether the Circular Inference model could be used to model this neurological disorder.

Cortical hierarchy and the different types of information loops in Circular BP The Circular BP algorithm is defined as a specific impairment of standard BP, whereby the message going from the sender node i to the receiver node j gets reverberated back to the sender i (and

is later counted as if it initially came from node j). A consequence of this is that information travelling in the probabilistic graph gets overcounted (self-amplification phenomenon, also known as positive feedback). As several other Bayesian models of the brain, the Circular Inference model (Jardri and Denève, 2013a) goes beyond the idea of multidirectional reverberation of information by investigating whether reverberation of the sensory evidence specifically (or alternatively, of the prior knowledge only) could potentially explain the symptoms of schizophrenia. This idea builds upon the notion of cortical hierarchy.

Felleman and Van Essen (1991) proposes a detailed cortical organization of the visual area and somatosensory/motor areas of the macaque monkey, by splitting subregions into a hierarchical graph composed of 14 levels. This graph is built from the anatomical definition of what a feedforward or feedback connection is: a feedforward pathway goes from superficial layers to layer 4, and a feedback pathway goes from deep or superficial layers to outside layer 4. The great majority of pathways involve reciprocal connections between areas. Intriguingly, the resulting hierarchical organization of the cortex defined by anatomy (feedforward versus feedback pathways) corresponds to a hierarchy of concepts: edges and lines are encoded in the primary visual cortex V1 (at the bottom of the hierarchy), shapes in V2 (above V1), objects in V4 (above V2), up to faces in the inferior temporal cortex IT (above V4). This shows that visual processing is hierarchical, building features of higher complexity while going up in the hierarchy. A consequence of this is that receptive fields increase in size and complexity with the level in the cortical hierarchy (Hubel and Wiesel, 1962). An extension of this idea is that the entire brain, that is, not only the visual and somatosensory/motor cortices, has a hierarchical organization (Hilgetag and Goulas, 2020).

It has been hypothesized that cortical hierarchies implement a model of the world's causal structure which translates into a hierarchical probabilistic graph, describing how noisy sensory inputs are caused by hidden states of the environment (e.g., presence of a car or not). In that respect, the notion of cortical hierarchy is equivalent to the notion of hierarchy between nodes of the probabilistic graph, given the assumption that lower-level concepts are encoded in brain regions which are located lower in the cortical hierarchy, as it is the case for vision. According to this view, the brain performs hierarchical Bayesian inference on this causal model of the world (Lee and Mumford, 2003; Summerfield and Koechlin, 2008; Rohe and Noppeney, 2015; Diaconescu et al., 2017) in the wild variety of tasks involving making probabilistic inference, which includes visual perception. The type of pathway (feedforward or feedback) determines the direction of information flow: sensory evidence, coming from lower sensory areas as V1, is propagated toward higher areas through feedforward connections, and prior information, encoded in higher areas, propagates through feedback connections down the hierarchy (as most connections are reciprocal in the cortex). The hierarchy in the probabilistic graph is defined by the level of the concept encoded by variable x_i versus x_j. This is a rather natural concept in vision, where low-level features like angles can be combined to create more evolved shapes, and eventually, objects or faces. It generalizes to any Bayesian network, an oriented graph which describes the causal dependencies between concepts. For instance, a disease is composed of (i.e., "causes") particular symptoms and therefore the variable encoding for the presence of disease is above the variable encoding for the presence of a particular symptom.

Considering the reverberation of the sensory evidence only (or of the prior knowledge only) in the Circular BP algorithm requires differentiating the case where node i is above node j in the (hierarchically organized) probabilistic graph from when i is below j in the update of the message $m_{i\to j}$ (see Equation (1.12)). This requires $\alpha_{i\to j}$ to depend on the relative position of node i and node j in the hierarchy. We talk of "ascending" loops to mean that information going up the cortical hierarchy (i.e., sent from node j to i where i is above j) is reverberated back (to node j). Therefore, the message $m_{i\to j}$ contains partly the message in the opposite direction

$m_{j\to i}$. In other words, sensory evidence travelling up the hierarchy is reverberated back, thus corrupting the prior knowledge. The amount of ascending loops is controlled by $\alpha_{i\to j}$ for i above j in the hierarchy (a presence of ascending loops means that $\alpha_{i\to j} \neq 1$). On the contrary, "descending" loops involve information going down the hierarchy (high-level prediction) being reverberated back to the top and treated as sensory evidence. This corresponds to $\alpha_{i\to j} \neq 1$ for i below j in the hierarchy). Mathematically, having potentially different amounts of ascending and descending loops is equivalent to allow $\boldsymbol{\alpha}$ from Circular BP to be an asymmetric matrix: $\alpha_{i\to j} \neq \alpha_{j\to i}$.

Ascending loops and descending loops might account for different phenomena (Jardri and Denève, 2013a). The effect of ascending loops is to create overconfidence in sensory observations and therefore explain several features of schizophrenia like the so-called "positive" symptoms. These involve the jumping-to-conclusions effect, the resistance to prior-based perceptual illusions. Jardri et al. (2017) evens shows a correlation between the level of symptoms of people with schizophrenia and the amount of ascending loops in a probabilistic task similar to the well-known beads task (Huq et al., 1988). On the contrary, descending loops were used to model the effects of psychedelic drugs (Leptourgos et al., 2021), that is, crossmodal percepts and stronger illusions, and could be related to negative symptoms of schizophrenia (Jardri et al., 2017). It is hypothesized in Jardri and Denève (2013b) that ascending and descending loops must have different anatomical substrates and could be differently affected by lesions, deficits in GABA receptors or neuromodulation.

As previously stated in section 1.4.3.1, some parts of the thesis differentiate ascending from descending loops (e.g. section 2.2), whereas some others (in fact most others) don't and consider $\boldsymbol{\alpha}$ to be a symmetric matrix: $\alpha_{i\to j} = \alpha_{j\to i} \equiv \alpha_{ij}$.

1.6 Motivation behind Circular BP: excitation-inhibition imbalance

The underlying biological idea behind the Circular Inference model is that the level of circularity $1 - \alpha$ controls for the amount of "excitation-inhibition imbalance". This requires to define the notion of balance between excitation and inhibition, and of its disturbance.

E-I balance Two processes fight in the brain: excitation and inhibition. Neural circuits, which in general do not show runaway excitation and do not fall silent either, are said to balance excitation and inhibition: we talk of "E-I balance". The resulting state is controlled by the relative numbers and activities of excitatory (typically glutamatergic) and inhibitory (typically GABAergic) neurons (Rubenstein and Merzenich, 2003). Interestingly, experiments show a balance over time: when excitation to a given neuron increases, inhibition increases proportionally. This is also true over space. Experiments show that the ratio between the number of excitatory and inhibitory synapses is constant, both across dendritic branches of a single neuron and across neurons (Liu, 2004), which is referred to as *structural balance*. Furthermore, the E/I ratio (ratio of excitatory to inhibitory inputs), measured by the ratio of excitatory to inhibitory synaptic strengths, is found constant across neurons (Liu, 2004; Xue et al., 2014), which is referred to as *functional balance*.

There are two opposing theories on the degree of balance between excitation and inhibition, loose balance (Ahmadian and Miller, 2021) or tight balance (Denève and Machens, 2016), depending on whether the sum of excitatory and inhibitory inputs to a neuron is comparable in size or much smaller than the excitatory and inhibitory inputs. These opposing theories are associated to different predictions. The debate remains open for now as measuring excitatory and inhibitory currents simultaneously in a given neuron is impossible; instead, researchers measure independently the average excitatory and the average inhibitory conductances. Despite

the development of new tools (e.g. Trakoshis et al. (2020)), there is currently a lack of robust biomarkers of the E-I balance that would be non-invasive and applicable in humans on a large scale.

However, the idea of balance between excitatory and inhibitory processes is now commonly accepted. Inhibition is seen as a way to help the network being efficient (low spiking rates corresponds to low energy consumption by the brain), robust to noise, and fast. It is also hypothesized to help reduce randomness in cortical operations (Wehr and Zador, 2003). The apparent random firing of cortical neurons has been seen as a consequence of the balance between excitatory (positive) and inhibitory (negative) synaptic currents: indeed, in this case, the network evolves on a chaotic or quasi-chaotic attractor. The importance of the E-I balance appears in biophysically realistic models of working memory, in particular because it is a necessary condition for the stability of the attractor network (Wang, 2006; Loh et al., 2007; Rolls et al., 2008; Murray et al., 2014). Lastly, inhibition prevents an explosion of cortical activity which might take place with excitatory neurons only, and which reminds the epileptic seizures (bursts of activity seen in epilepsy).

E-I imbalance If a change occurs to excitation and/or inhibition, neural circuits reach a new state (or a new set of states) while in general maintaining stability and activity in the network (Sohal and Rubenstein, 2019). This is referred to as "E-I imbalance", or disturbance of the E-I balance. For instance, if the level of excitation increases, then the network activity increases until more inhibition is recruited, which produces a new balanced state. As pointed out by Sohal and Rubenstein (2019), this definition of the alteration in E-I balance suggests that excitation and inhibition are unidimensional entities, which is an oversimplified view because there are different subtypes of excitatory or inhibitory neurons. However, this simple concept can help understand the mechanisms underlying neuropsychiatric disorders, particularly autism spectrum disorder and schizophrenia (Foss-Feig et al., 2017).

The "imbalance" between excitation and inhibition in favor of excitation may lead to to hyperexcitability of the cortex and increased neuronal noise according to Rubenstein and Merzenich (2003). More generally, it has been hypothesized that the balance between excitation and inhibition is abnormal in many neuropsychiatric conditions, including schizophrenia and autism (Rubenstein and Merzenich, 2003; Gao and Penzes, 2015; Canitano and Pallagrosi, 2017; Sohal and Rubenstein, 2019; Trakoshis et al., 2020). More precisely, there is strong biological evidence for an increased E/I ratio in schizophrenia (and other neuropsychiatric disorders) in the prefrontal cortex, as a result of impaired functioning of inhibitory (GABAergic) interneurons which causes cortical disinhibition (Lewis et al., 2005; Yoon et al., 2010; Marín, 2012; Selten et al., 2018).

The hypothesis of an increased E/I ratio in schizophrenia is backed by experiments in which a modification of the E/I ratio changed the behavior of healthy animals or humans, in ways which remind the symptoms of schizophrenia. For instance, an increase of the balance between excitation and inhibition in the medial prefrontal cortex of the mouse induced using optogenetic techniques causes social deficits (Yizhar et al., 2011). Notably, these social deficits partly disappear after a subsequent elevation of inhibition by increasing the activity of particular inhibitory neurons: PV interneurons (see also Lewis et al. (2012)). Similarly, an increase in activity of PV interneurons improves the impaired social behavior in an animal model of autism (Selimbeyoglu et al., 2017). Moreover, the ketamine drug, which is hypothesized to increase the E/I ratio by perturbing NMDA receptors (a type of glutamate receptors) on inhibitory interneurons, causes cognitive impairments frequently observed in schizophrenia like working memory deficits (Murray et al., 2014), impaired decision-making (Lam et al., 2017), but also other schizophrenia-like

symptoms like perceptual aberrations, delusional ideas, thought disorder and changes in affect (Krystal et al., 2005).

In the biophysical models of working-memory mentioned above, an impairment in the level of excitation (through NMDA or AMPA) and/or inhibition (through GABA) leads to schizophrenia-like symptoms. This involves working-memory impairments (unstable memories for decreased excitation; see Murray et al. (2014)), cognitive impairments (jumping from one thought to another as a result of the flattening of the attractor landscape by decreasing both excitation and inhibition), and potentially positive symptoms of schizophrenia like hallucinations (see Jardri and Denève (2013b) for a review of attractor models of hallucinations).

Circular BP and E-I imbalance The idea behind the Circular Belief Propagation algorithm is the one of a lack of control in information transmission, which corresponds at the intuitive level to a disturbance of the excitation/inhibition balance in favor of excitation (Jardri and Denève, 2013a,b; Jardri et al., 2016). A perfect balance ($\alpha = 1$), which amounts to BP, controls exactly for the loops of information and therefore avoids reverberations (at least for acyclic probabilistic graphs which was the only case considered before this thesis). Increased excitation relative to inhibition corresponds to $\alpha < 1$ and is pictured as information (exchanged between pyramidal cells) not being properly properly controlled by inhibitory interneurons, leading to reverberation of signals between excitatory populations of neurons (information reverberation). Finally, increased inhibition relative to excitation corresponds to $\alpha > 1$ (this hypothesis is not considered to model positive symptoms of schizophrenia with the Circular Inference model).

In local cortical circuits, pyramidal cells (excitatory neurons) are interconnected in a positive-feedback manner, and GABAergic neurons (inhibitory neurons) operate a negative-feedback loop (Shu et al., 2003). A potential network that might implement the algorithm could be composed of excitatory and inhibitory populations, where excitatory populations are responsible for exchanging information, and inhibitory populations are responsible for controlling this exchange (remove the effects of positive feedback). The state of the network (approximate marginals after convergence of the algorithm) is modified as a consequence of the impairment in the algorithm (circularity or lack of inhibitory control). However, the network remains stable and active, that is, excitation does not run away and inhibition does not shut down the network activity completely.

1.7 Circular BP in the realm of computational psychiatry models

The Circular Inference model, based on the Circular Belief Propagation algorithm, is not the only computational model of psychosis, schizophrenia, and psychiatric diseases in general. Here we mention the different classes of models of schizophrenia and autism. For excellent reviews on the subject, see Valton et al. (2017), which discusses normative and mechanistic models of schizophrenia, and Lanillos et al. (2020), which focuses on mechanistic models of both schizophrenia and autism.

Bayesian models of mental disorders relying on Circular BP (Jardri and Denève, 2013a) or predictive coding (Adams et al., 2013) are at the heart of this thesis (see more on the predictive coding theories of the brain and their link with Circular BP in section 5.3.4). However, Bayesian models (which are by nature normative) are not the only way of modelling psychiatric disorders. In fact, mechanistic approaches exist to understand mental disorders, as well as other normative approaches than Bayesian models. In fact, distinct hypothesizes have been made concerning the origin of schizophrenia, including the dopamine hypothesis, the glutamate hypothesis, the GABAergic hypothesis, and the disconnection hypothesis (Valton et al., 2017). These hypotheses gave rise to a huge diversity of models which includes biophysical models, reinforcement learning models, Hopfield networks and other artificial neural networks, etc. At the extreme opposite end

from normative Bayesian models, we can mention the biophysical models of working-memory (Wang, 2006; Murray et al., 2014; Loh et al., 2007; Rolls et al., 2008). These models incorporate different types of neurotransmitters and base the network on experimentally-measured (concentrations/)synaptic connections. For instance, Wang (2006) uses integrate-and-fire neurons (both excitatory pyramidal cells and inhibitory interneurons) and models their modulation by NMDA receptors. Loh et al. (2007); Rolls et al. (2008) go a step further by including not only NMDA receptors but also AMPA and GABA to their model.

Note that these hypotheses (dopamine, glutamate, GABAergic or disconnection) are not necessarily mutually exclusive, as all systems are linked. For example, an alteration in the dopaminergic system (dopamine hypothesis) could lead to reduced synaptic connectivity or abnormal functional connectivity (disconnection hypothesis), possibly as a way to compensate for the dysfunction initially introduced (Lewis and Gonzalez-Burgos, 2006). Finally, these hypotheses of course relate to Bayesian models impairments, although currently only hypothetically. Circular Inference for instance relies on the hypothesis of excitation-inhibition imbalance, which is tightly linked to the GABAergic hypothesis of schizophrenia, but also relates to the dopamine and glutamate hypotheses (see Jardri and Denève (2013a)). Furthermore, the work presented in section 2.3 (Bouttier et al., 2021) shows that an impairment of the level of excitation-inhibition balance leads to abnormal functional connectivity, which is consistent with the dysconnection hypothesis of schizophrenia.

In this "jungle" of models, the mathematics are different, levels of advancement are different (some remain for now at the algorithmic level without proposed implementation), and the exact processes modelled can be different as well (for instance, a significant proportion of models only account for the positive symptoms of schizophrenia, whereas others model disorganization (Lanillos et al., 2020)). Taken together, these models, as diverse as they may be, contribute towards the same goal of understanding psychiatric disorders, by trying to explain experimental brain and behavioral data collected on people suffering from these disorders, as well as animal models or pharmacological models of the disease.

1.8 Unanswered questions and aims of this thesis

The Circular Inference model (Jardri and Denève, 2013a), based on the Circular Belief Propagation algorithm, has been proposed to account for psychosis and subsequently developed and tested in several articles (Denève and Jardri, 2016; Jardri et al., 2017; Leptourgos et al., 2017; Chrysaitis et al., 2021; Leptourgos et al., 2021). Nevertheless, several questions remain unanswered before one can state with more certainty that it is a potentially good model of the way the brain implements probabilistic inference. The first limitation of the model is that it considers a specific (and restrictive) class of probability distributions, whose associated graphical models are acyclic, pairwise, and represent exclusively binary variables. The second limitation, linked to the first one, is that Circular Inference is only a model of impaired behavior and considers that $\alpha = 1$ (BP) symbolizes optimal inference, as it would be the case for a perfect Bayesian observer. However, for distributions with loops, not only Circular BP (that is, $0 \leq \alpha < 1$) but also BP itself may carry out really poor inferences and sometimes be unstable, a fact that the current model completely disregards. The third limitation of the model is that the change introduced in Circular BP based upon BP is not normative and instead rather ad-hoc. The fourth limitation is that the few propositions of neural implementation are either too vague or formulated on a modified Circular BP (see section 4.1), and make implementation hypotheses without testing them by simulating the network. Finally, the model lacks confrontation to experimental brain data.

The work presented here aims at filling (some of) these gaps. All the scientific questions tackled in this thesis point in the same direction, that is, the long-term validation or rejection of the model. Circular Inference is a normative model (see section 1.1 for a definition) and therefore it relies on abstract assumptions (here, that the Circular Belief Propagation algorithm is used by the brain to perform inference). As any normative model, it is important to validate or reject it, and a good way to progress toward this goal is to test its capabilities by stepping out of its comfort zone (which is behavioral modelling in this case). Another important step is to generate predictions, be it behavioral or most importantly, neural. This is a common criticism towards all Bayesian models of the brain (including predictive coding and BP), which cannot be proven false as long as they stay conceptual. The only way to generate such predictions is to build a precise enough proposition of implementation of the model in neural circuits.

This thesis is organized as follows. In **chapter 2**, we provide further evidence for Circular BP as a model of suboptimal behavior. First, we use the algorithm to model bistable perception, a badly understood phenomenon for which the Circular Inference model was initially not designed. The model naturally captures various aspects of bistability, including Levelt's laws, which is not the case for a model with proper loop correction, that is, without circularity. This leads the way to fitting bistable perception (behavioral) data to the model. Second, we propose a large-scale model of the brain by hypothesizing that the generative network mirrors the small-world anatomical structure found in brain imaging experiments. Using this large-scale model, we replicate using numerical simulations particular disturbances in the functional connectivity observed in schizophrenia.

In **chapter 3**, we show that the Circular BP algorithm can be used as a model of optimal behavior (thus answering our interrogation about BP not being a good model of optimal behavior). We develop extended Circular BP, an algorithm which naturally generalizes Circular BP and relate it to existing approximate inference algorithms, therefore providing a normative foundation to the Circular Inference model. We show that this extended Circular BP algorithm significantly outperforms BP in cyclic probabilistic graphs, and performs approximate inference with a very impressive quality even for fully dense probabilistic graphs.

In **chapter 4**, we propose a neural implementation of Circular BP applied to probability distributions over binary variables, and recover the effects of Circular BP using all this implementation. More precisely, we map the algorithm onto a network composed of rate units, which implements the algorithm exactly. Furthermore, we describe a more biologically-plausible implementation composed of spiking units (integrate-and-fire neurons), which approximate Circular BP very well. We also provide a biologically plausible learning algorithm for parameters of Circular BP, inspired by the principles of excitation-inhibition balance and efficient information transmission.

Finally, in **chapter 5**, we generalize the restrictive case of binary variables as used previously. We formulate the Circular BP algorithm on general factor graphs instead of pairwise graphs. We further investigate a special case of the algorithm where variables are continuous instead of discrete: Gaussian Circular BP. This leads us to propose a rate model implementation of the algorithm, which is very similar to the one proposed for binary variables. Finally, we relate Gaussian Circular BP to predictive coding theories of mental disorders.

Chapter 2

Circular Belief Propagation as model of suboptimal behavior

Summary of Chapter 2

We provide in this chapter further evidence for Circular BP as a model of suboptimal behavior. First, we use the algorithm to model bistable perception, a badly understood phenomenon for which the Circular Inference model was initially not designed. The model naturally captures various aspects of bistability, including Levelt's laws, which is not the case for a model with proper loop correction, that is, without circularity. This leads the way to fitting bistable perception (behavioral) data to the model. Second, we propose a large-scale model of the brain based on Circular BP by hypothesizing that the generative network mirrors the small-world anatomical structure found in brain imaging experiments. Using this large-scale model, we replicate using numerical simulations particular disturbances in the functional connectivity observed in schizophrenia, as well as targeted overactivation in specific brain regions.
This chapter is based on the two following articles. Section 2.2 corresponds to *A functional theory of bistable perception based on dynamical circular inference*, by P. Leptourgos, V. Bouttier, R. Jardri and S. Denève (2020), PLoS Comput Biol 16(12): e1008480. Section 2.3 corresponds to *Circular inference predicts nonuniform overactivation and dysconnectivity in brain-wide connectomes*, by V. Bouttier, S. Duttagupta, S. Denève and R. Jardri (2021), Schizophrenia Research.

2.1 Introduction

After presenting the Circular Belief Propagation algorithm and its link to psychiatric disorders and excitation-inhibition balance, we tackle in this chapter the modeling of "suboptimal behavior" with the Circular BP algorithm.

An *optimal* evaluation of one's confidence is simply the estimation of a marginal probability using the laws of probability (including Bayes' rule) given all the pieces of evidence at hand: optimality is defined here in the Bayesian sense. This means most of the time perceiving the right object or taking the right decision. However, this can also account for illusions (Geisler and Kersten, 2002; Lee and Mumford, 2003) like motion illusions (Weiss et al., 2002) in which the prior knowledge biases the sensory information and can fool us to taking wrong decisions with

respect to the sensory evidence alone, but optimal decisions with respect to the sensory evidence and prior together.

Suboptimality is by definition the contrary of optimality. By nature, it is impossible to show that a given human is a suboptimal Bayesian observer, because it would mean showing that this human takes optimal decisions in all possible situations. However, behavioral experiments suggest that no human is a perfect Bayesian observer, as our decisions are sometimes suboptimal. We focus in this chapter on two concrete examples of such suboptimality: the bistable perception phenomenon and schizophrenia.

Bistable perception, a phenomenon observed in the general population consisting of spontaneous alternations between two possible interpretations of a single situation, is by nature a suboptimal process. The brain should indeed perceive both possible percepts at the same time, and does not. We assume here that perception is done subconsciously , which implies that what we perceive at a given time (e.g., the Necker cube from above) is the exact translation of our beliefs (here, that means that the probability that the Necker cube is from above is above 50%).

Schizophrenia is a mental disorder which is linked to suboptimality as well. Symptoms of schizophrenia are in fact manifestations of extreme suboptimality. For example, visual or auditory hallucinations can be seen as a wrong integration of sensory evidence and prior information, where highly unreliable and/or non-significant information ("noise") is perceived as if the information were strong. Delusions - having abnormal beliefs which do not change or hardly change despite contrary evidence - is a manifestation of suboptimality as well, because Bayes' rule states that beliefs should be updated if new information gets added. Finally, other differences experimentally observed between people with and without schizophrenia suggest that the probabilistic inference system differs in many ways between the two populations. An example of that is the jumping to conclusions phenomenon - taking decisions faster and being overconfident about it - which is observed in schizophrenia and can be seen as a wrong (i.e., suboptimal in Bayesian terms) accumulation of evidence over time.

Altogether, this chapter, which shows that bistable perception and some characteristics of schizophrenia can be accounted using Circular Belief Propagation, provides further evidence for the algorithm to be a good model of suboptimal behavior.

2.2 Circular BP as model of bistable perception

This section 2.2 corresponds to the following published article: *A functional theory of bistable perception based on dynamical circular inference*, by P. Leptourgos, V. Bouttier, R. Jardri and S. Denève (2020), PLoS Comput Biol 16(12): e1008480.

2.2.1 Abstract

When we face ambiguous images, the brain cannot commit to a single percept; instead, it switches between mutually exclusive interpretations every few seconds, a phenomenon known as *bistable perception*. While neuromechanistic models, e.g., adapting neural populations with lateral inhibition, may account for the dynamics of bistability, a larger question remains unresolved: how this phenomenon informs us on generic perceptual processes in less artificial contexts. Here, we propose that bistable perception is due to our prior beliefs being reverberated in the cortical hierarchy and corrupting the sensory evidence, a phenomenon known as "circular inference". Such circularity could occur in a hierarchical brain where sensory responses trigger activity in higher-level areas but are also modulated by feedback projections from these same areas. We show that in the face of ambiguous sensory stimuli, circular inference can change the dynamics of the perceptual system and turn what should be an integrator of inputs into a bistable attractor

switching between two highly trusted interpretations. The model captures various aspects of bistability, including Levelt's laws and the stabilizing effects of intermittent presentation of the stimulus. Since it is related to the generic perceptual inference and belief updating mechanisms, this approach can be used to predict the tendency of individuals to form aberrant beliefs from their bistable perception behavior. Overall, we suggest that feedforward/feedback information loops in hierarchical neural networks, a phenomenon that could lead to psychotic symptoms when overly strong, could also underlie perception in nonclinical populations.

2.2.2 Introduction

All perceptual systems have one fundamental goal: to interpret the surrounding environment based on unreliable sensory evidence. In most cases, this task is performed very accurately, and the correct interpretation is found. Sometimes, perceptual systems fail to detect any meaningful interpretation (e.g., when sensory evidence is too degraded) or converge to the wrong interpretation (e.g., visual illusions (Weiss et al., 2002; Notredame et al., 2014)). Finally, a third possibility occurs (mainly in lab conditions (Arnold, 2011)) when ambiguity is high; the system detects more than one plausible interpretations but instead of committing to one interpretation, it switches every few seconds, a phenomenon known as *bistable perception* (Blake and Logothetis, 2002). Despite ongoing scientific efforts, there has been no unanimous agreement either on the causes of bistability or on its functional role.

The dominant mechanistic view on bistable perception suggests that it results from the competition between different neuronal populations, each of them encoding a different interpretation of the sensory signal (Blake, 1989). The two populations suppress each other via lateral inhibition, while some form of slow negative feedback (e.g., spike frequency adaptation or synaptic depression) acts on the dominant population, weakening the interpretation that is currently perceived (Lago-Fernández and Deco, 2002; Laing and Chow, 2002; Wilson, 2003; Noest et al., 2007; Wilson, 2007; Vattikuti et al., 2016). Additionally, injected noise renders irregular switching and in some models, it can even be the driving force of oscillatory behavior (Moreno-Bote et al., 2007; Shpiro et al., 2009; Panagiotaropoulos et al., 2013; Huguet et al., 2014). Although these models have proven quite successful in describing different experimental observations (and linking them to the underlying neural mechanisms), they do not address functional considerations about bistable perception.

To overcome this issue, other groups suggested functional models of bistability, largely based on the idea that the brain is an inference machine and perception is equivalent to a probabilistic process (e.g., (Brascamp et al., 2018); see also (Hohwy et al., 2008; Weilnhammer et al., 2017) for predictive coding, or (Sundareswara and Schrater, 2008; Reichert et al., 2011; Gershman et al., 2012) for sampling). However, some crucial questions remain largely unanswered from a purely normative perspective, namely, (1) why would a system form such strong percepts based on ambiguous sensory evidence, but only in some cases, and why do the percepts persist in such a way instead of switching rapidly, and (3) how the behavior of individuals in bistable perception tasks may predict their performance in other probabilistic inference tasks.

In the present paper, we address the problem of bistable perception by proposing a functional model with a well-defined interpretation in terms of generic neural processes. Based on previous experimental findings, we suggest that bistability could be a perceptual manifestation of circular inference (CI), a form of belief propagation in which priors and likelihoods are reverberated in the cortical hierarchy and consequently corrupted by each other (Jardri and Denève, 2013a; Denève and Jardri, 2016). More specifically, bistable perception could be imposed by the presence of "descending loops", where high-level beliefs are combined with sensory representations (through feedback connections), and subsequently reinforce themselves (through feedforward connections).

This results in the perceptual system "seeing what it expects" instead of the truly ambiguous image (Leptourgos et al., 2017). Of note, previous work from our group linked CI with pathological brain function, as in the case of schizophrenia (Jardri et al., 2017) but also to a smaller extent with physiological functioning (Leptourgos et al., 2020b).

In the following sections, we derive the dynamics of inference in the presence of ambiguous sensory stimuli and inference loops. The consequence of CI is to replace what is normally a slow temporal integration of unreliable sensory evidence with a bistable attractor switching between two highly trusted interpretations. We demonstrate that such a model can reproduce well-known qualitative aspects of bistability, including the four Levelt's laws and the stabilizing effect of intermittent presentation, while it also makes testable quantitative predictions (e.g., about the behavior of patients suffering from schizophrenia). Since circularity arises from an imbalance between neural excitation and inhibition in recurrent brain circuits (Leptourgos et al., 2017; Jardri et al., 2016), our approach bridges normative interpretations of bistable perception with plausible underlying neural mechanisms.

2.2.3 Methods

Here, we introduce a CI model of bistable perception and highlight its underlying functional assumptions. For reasons of clarity, we refer to the example of the Necker cube, an ambiguous 2D figure which is compatible with 2 different 3D cubes and generates bistability: a cube that is "*seen from above*" (later called the SFA interpretation) and a cube that is "seen from below" (later called the SFB interpretation) (Fig 2.1A). Note that the model can be generalized to any other stimuli inducing perceptual rivalry.

2.2.3.1 Generative model

Our model postulates that bistable perception is triggered by the same mechanisms and computations that underlie normal perception. There is accumulating evidence that the brain uses its cortical hierarchy to represent the causal structure of the world (Friston, 2008; Clark, 2013). Brain circuits invert this "generative model" to find the most likely interpretation of the noisy sensory information. In other words, perception can be viewed as an instance of hierarchical Bayesian inference (Friston, 2008; Mathys et al., 2014) (Fig 2.1A). A particularly striking example of this inferential process is 3D vision (such as the perception of the Necker cube). The brain has no direct access to the 3D structure of the perceived object. In contrast, it receives low-level 2D sensory information from the retina. In such a context, the task of the perceptual system is to extract valuable depth cues and combine them with high-level prior knowledge, to make "educated guesses" about the 3D object. Evidence suggests that this is a gradual process (Finlayson et al., 2017), with different brain regions representing features of different complexity; the lower levels of the visual cortex represent the basic features of the stimulus such as contours and orientations while higher levels are responsible for more abstract information such as the 3D organization of the stimulus (Felleman and Van Essen, 1991; Lee and Mumford, 2003).

In the case of the Necker cube, a veridical percept would correspond to a 2D drawing of crossing lines. The presence of illusory depth cues forces the brain to consider a 3D structure. Nonetheless, since the cues are ambiguous and contradictory, the 2D projection of the hypothetical 3D stimulus is compatible with different objects, including the SFA and SFB interpretations mentioned previously. The two interpretations are considered mutually exclusive, an assumption that corresponds with the epistemological truth that two different 3D objects cannot occupy the same space (Hohwy et al., 2008). It is interesting to note that in a more general sense, the Necker cube is compatible with an infinity of3D objects, among which the brain represents only the two

symmetrical cubes. This reduction of possible causes could be the result of hyperpriors used by the brain and is not considered in the current model.

We formalize this inference problem with a simple graphical model, a chain with 2 latent variables and one sensory observation (Fig 2.1A). This "generative model" summarizes assumptions made by our sensory system on the underlying causes of natural inputs, which may significantly differ from the artificial data presented in a laboratory setting.

The sensory observation (S) represents the basic features extracted by visual receptors (edges, contrast, etc.). For simplicity, S is assumed to be a scalar drawn from two probability distributions, one for each configuration of the cube, as illustrated in Fig 2.1A and 2.1B (red and blue dotted distributions; $P(S|X_{2D} = 1) \neq P(S|X_{2D} = 0)$). These distributions have different means $\pm\mu_{\text{int}}$ and the same variance σ^2_{int}. The difference in these two distributions considers the fact that natural 3D objects have true depth cues (disparities, shadows, occlusion, etc.), predicting different likelihoods for the two interpretations. Note that completely ambiguous stimuli (i.e., falling in the perfect overlap between the two distributions) are, in fact, rarely encountered in nature.

The next variable X_{2D} is binary and represents an intermediate level of complexity in the perceptual hierarchy (e.g., the 2D surfaces and their orientation). Finally, the binary variable X_{3D} represents the final 3D cube configuration, with values 0 and 1 corresponding to SFB and SFA respectively. w_S corresponds to how reliably X_{3D} predicts X_{2D}.

$$w_s = P(X_{3D} = 1|X_{2D} = 1) = P(X_{3D} = 0|X_{2D} = 0) \tag{2.1}$$

We also assume that the environment has some volatility, e.g., objects are not permanently present, but occasionally appear or disappear. Thus, X_{3D} can randomly switch at any time, as represented by two rates of change, from 0 to 1 (r_{on}), and from 1 to 0 (r_{off}). For the sake of simplicity in the notation, we will replace X_{3D} at time t by X_t, representing the 3D configuration of the cube (SFA or SFB) at time t.

$$r_{\text{on}}dt = P(X_t = 1|X_{t-dt} = 0) \tag{2.2}$$

$$r_{\text{off}}dt = P(X_t = 0|X_{t-dt} = 1) \tag{2.3}$$

Note that if we use $r_{\text{on}} \neq r_{\text{off}}$, one of the two interpretations becomes more probable that the other. This is very useful in the case of the Necker cube, where people usually prefer the SFA interpretation, according to a general prior to view things from above ($r_{\text{on}} > r_{\text{off}}$) (Mamassian and Landy, 1998).

Now that we have described the generative model, i.e., the internal model used by the brain to perceive objects in the real world, we have to consider the artificial stimulus provided during a bistable perception experiment. The Necker cube is very unnatural in the sense that it contains no real depth cue. Thus, the sensory information it provides is assumed to be sampled (independently at each time step) from a Gaussian distribution with mean μ_{noise} ($\mu_{\text{noise}} = 0$ (Gaussian process without drift) if the cube is completely unbiased and $\mu_{\text{noise}} \neq 0$ (Gaussian process with drift), if there are visual cues supporting one of the two configurations, e.g., different contrast for the edges) and variance σ^2_{noise} (Fig 2.1B; black and gray distributions).

The ultimate goal of the perceptual system is to infer X_{3D} using the noisy measurements and any available prior knowledge (for more information about the generative model, see Appendix 1).

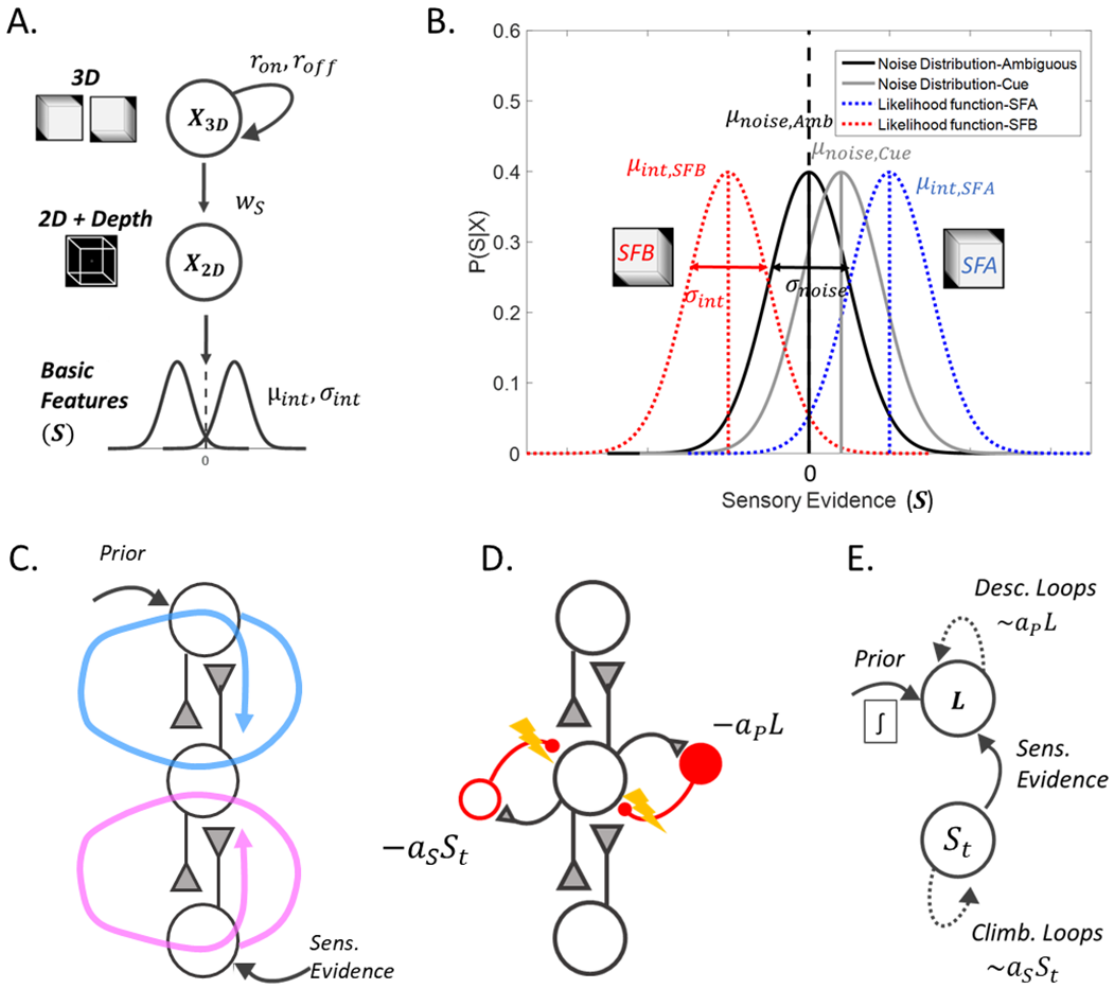

Figure 2.1: **Normative model for how 3D objects result in particular sensory inputs, and putative neural implementation of the corresponding perceptual inference**. **(A)** The internal model is a simple Bayesian generative model, where 3D objects predict the 2D image, and the 2D image predicts low-level sensory inputs. The brain interprets the depth cues (basic features) as indicative of real depth. Consequently, it first reconstructs the 2D figure and from that, it infers the 3D object. Note that in reality there is one single 2D stimulus (the Necker cube drawing) containing contradictory depth cues. **(B)** Close-up on the assumed "basic feature" distributions (likelihood) compared to the real input distributions. The brain interprets the depth cues as meaningful, predicting separate input distributions for the two cubes (SFA, SFB; two objects cannot occupy the same space), which corresponds to two non-overlapping likelihood distributions in the internal model (dotted red and blue distributions). In the totally ambiguous case (cube with no extra cues), the real input is sampled from a distribution with mean 0 (black). Visual cues shift this input distribution toward mostly positive or negative values. Crucially, there is a discrepancy between the real input and the input assumed by the internal model. This, together with the loops, predicts the suboptimal inference at the heart of bistable perception. **(C)** A simplified neural implementation of hierarchical perceptual inference. Reciprocal connections can combine bottom-up sensory evidence with top-down priors at all levels of the hierarchical representation. Unfortunately, this also creates redundant information loops, ascending (magenta arrow) and descending (blue arrow). **(D)** The brain can cancel these loops by using inhibitory interneurons and maintaining a tight E/I balance. If this balance is impaired, however, there will be some residual loops, parameterized by a_P (descending loops, amplifying prior beliefs) and a_S (ascending loops, amplifying the sensory evidence). L is the log-ratio of the belief. **(E)** From the Bayesian model of panel (A), we derived an attractor model that performs inference in the presence of loops. The model accumulates noisy evidence while descending loops add positive feedback and ascending loops increase the sensory gain.

Temporal dynamics of inference

We show in Appendix 1 that exact inference implements a leaky integration of the noisy sensory input (Fig 2.1E), i.e.

$$\frac{dL}{dt} = -\Phi(L) + w_{\text{int}} S \tag{2.4}$$

where $w_{\text{int}} = \frac{2\mu_{\text{int}}}{\sigma^2_{\text{int}}}(2w_s - 1)$ represents the overall reliability of the sensory input (as assumed by the generative model). L is the log-odds ($L = \log\left(\frac{P(X_t=1|S_{0\to t})}{P(X_t=0|S_{0\to t})}\right)$. The nonlinear leak term $\Phi(L)$ depends on the transition rates, i.e.,

$$-\Phi(L) = (r_{\text{on}}e^{-L} - r_{\text{off}}e^{L}) + (r_{\text{on}} - r_{\text{off}}) \tag{2.5}$$

As a result of this leak, in the absence of sensory evidence, the log-odds go back to the constant prior value $\log(r_{\text{on}}/r_{\text{off}})$. This relaxation is faster for larger volatility in the environment (higher transition rates). In the presence of reliable and unambiguous sensory input (e.g., when adding visual cues, i.e., $\mu_{\text{noise}} \neq 0$), L integrates out the noise and eventually reaches high (positive) or low (negative) values, corresponding to high levels of confidence in favor of the SFA or SFB configurations. However, in the presence o fa completely ambiguous sensory input, L integrates unbiased noise ($\mu_{\text{noise}} \neq 0$) and constantly hovers around the prior value, rarely reaching a sustained high level of confidence in either of the two configurations.

Dynamics notably change in the presence of CI. CI is defined in the context of hierarchical probabilistic inference but can also be understood intuitively as a consequence of feedforward/ feedback loops in brain circuits (Fig 2.1C). Bottom-up sensory evidence (from S to X_{2D}) and top-down prior information (from X_{3D} to X_{2D}) have to be combined to compute the probability of intermediate representations (X_{2D}), a task presumably performed by feedforward (bottom-up) and feedback (top-down) connections converging on the same intermediate "2D" sensory area (Douglas et al., 1995). This hypothesis is supported by the experimentally observed top-down modulation of sensory neuron responses by higher-level interpretation of the image (Hupé et al., 1998; Bullier et al., 2001; Manita et al., 2015). However, feedforward connections between the "2D" and "3D" areas also communicate this modulated sensory response back to the "3D" areas. While this modulation does not bring any "new" information, it could nevertheless be mistaken for additional sensory evidence supporting the current interpretation. In fact, without dedicated control mechanisms, feedforward/ feedback loops would systematically result in CI in the underlying perceptual process. We found previously that while this can, in theory, be avoided by maintaining a tight excitatory/ inhibitory balance in brain circuits (Fig 2.1D), human subjects show some level of circularity in their probabilistic reasoning, which is aggravated in individuals suffering from schizophrenia (Jardri et al., 2017; Leptourgos et al., 2020b).

Here, we quantify the strength of CI by two variables representing the level of"ascending" (also called "climbing" (Jardri and Denève, 2013a); a_S) and "descending" loops (a_P). Descending loops represent to what extent top-down modulation of sensory responses is misinterpreted by upstream (higher-level) neurons as new sensory information, forcing the perceptual system to "see what it expects". Vice-versa, ascending loops represent to what extent intermediate sensory responses are misinterpreted by downstream (lower-level) neurons as prior knowledge, even when they do not provide them with any new information (Fig 2.1C). This forces the perceptual system to "expect what it sees" and over-interpret weak sensory inputs.

If CI is introduced in the model, the dynamics of perceptual integration changes as follows (Fig 2.1E):

$$\frac{dL}{dt} = -\Phi(L) + aL + w^*_{\text{int}} S \tag{2.6}$$

Variable	Description	Link to other variables
μ_{noise}	Drift of sensory evidence	-
σ_{noise}	Standard deviation of sensory evidence	-
μ_{int}	Mean of likelihood function	-
σ_{int}	Standard deviation of likelihood function	-
w_S	Feed-forward weight	-
a_P	Descending loops	-
a_S	Ascending Loops	-
r_{on}	Transition rate $(0 \to 1)$	-
r_{off}	Transition rate $(1 \to 0)$	-
w_{int}	Sensory gain without ascending loops	$w_{\text{int}} = \frac{2\mu_{\text{int}}}{\sigma^2_{\text{int}}}(2w_S - 1)$
w^*_{int}	Sensory gain with ascending loops	$w^*_{\text{int}} = w_{\text{int}}(1 + 2w_s a_s)$
b	Bias	-

Table 2.1: **Parameters of the model**

Note that the new auto-amplification term $aL = 2a_P w_S L$ (due to the corruption of the sensory evidence by the prior belief) is proportional to the strength of descending loops a_P and the assumed reliability of the sensory information, w_S. If a is large enough, this amplification term may exceed the leak term, at least in a certain range of confidence near $L = 0$. This leads, as we will see, to bistable dynamics. Importantly, this term not only depends on the strength of the descending loops but also on the reliability of the sensory input (assumed by the generative model). Bistable dynamics occur only for large w_S, which we may interpret as a typically highly reliable input (such as 2D drawings of 3D objects) as opposed to typically unreliable inputs (e.g., low contrast or degraded stimuli). This may explain in part why bistable perception is a relatively rare phenomenon in natural (non-laboratory) settings.

In contrast, ascending loops amplify the weight of the sensory evidence according to their strength, i.e., $w^*_{\text{int}} = w_{\text{int}}(1 + 2w_S a_S)$. In particular, ascending loops affect the dynamics only if a sensory stimulus is present and tend to destabilize the percept by increasing the gain of the noise injected into the dynamical system.

Note that without loss of generality, this model of perceptual dynamics can be reduced to 4 free parameters: the two transition rates r_{on} and r_{off}, the auto-amplification a and the overall gain of the sensory inputs w_{int}.

Perceptual decision Finally, we require a model of perceptual decision, which can predict the current percept from the confidence. For simplicity, we assume a maximum-a-posteriori (MAP) decision criterion, which means that decisions are made according to the sign of L (SFA if $L > 0$; SFB if $L < 0$). The MAP decision criterion results in optimal behavior when the goal of the system is to maximize accuracy, as in the case of perception.

Simulations For all the simulations, we used the Euler–Maruyama algorithm. The time step was fixed at $dt = 0.01s$. Both the standard deviation of the noise σ_{noise} (real model) and of the likelihood function σ_{int} (internal model) were equal to 1. The mean of the likelihood function $\pm\mu_{\text{int}}$ was also fixed at ± 1. $\mu_{\text{noise}} = 0$ for the completely ambiguous case and $\mu_{\text{noise}} \neq 0$ when sensory evidence was biased. The initial belief in all simulations was $L_0 = 0$. A summary of the parameters can be found in Table 2.1.

2.2.4 Results

As a first step, we highlight the importance of the descending loops in the generation of bistable perception from a phenomenological and mechanistic point of view. Subsequently, we illustrate how CI replicates some of the most seminal features of bistable perception, such as Levelt's laws but also some counterintuitive findings, including stabilization of perception after a brief disappearance of the stimulus. Finally, we present further consequences of the model, notable predictions about the performance of schizophrenia patients exposed to bistable stimuli.

Strong descending loops induce bistable perception An example of model dynamics in response to a continuous presentation of a Necker cube, in the presence of strong descending loops is shown in Fig 2.2A and 2.2C. With descending loops, the percept switches between two highly trusted interpretations (for example, $L = 4$ corresponds to probability 0.98 in favor of SFA; see also Appendix 3). Periods with low confidence are short and limited to sudden perceptual switches, induced by the noisy input. These switches occur at apparently random times, resulting in an exponential decay observed in the distribution of dominance durations (Fig 2.2E). When there is a bias (e.g., $r_{\text{on}} > r_{\text{off}}$), one of the two configurations (e.g., SFA) becomes more likely and is perceived more often (Fig 2.2C). However, the shape of the dominance durations remains similar for the two configurations, even if the durations of the preferred configurations are longer overall. It's worth-highlighting that the stronger interpretation is also perceived with higher confidence, a prediction that could be tested in future studies. For comparison, we also show the dynamics of the model without descending loops ($a_P = 0$) (Fig 2.2B and 2.2D). The resulting system is equivalent to a hidden Markov model (HMM), with transition rates r_{on} and r_{off} (Denève, 2008), and has only one stable state corresponding to the prior. As a result, the confidence behaves similarly to a leaky random walk. Since the leak maintains L close to zero, the system rarely attains high levels of trust in either configuration, which may preclude the emergence of strong and stable percepts in the absence of descending loops (instead, low confidence might give rise to mixed percepts (Knapen et al., 2011)).

Dependency of bistability on the parameters Due to its simplicity, the model dynamics can be analyzed more formally. This has the advantage of generalizing the model and providing a general view on the dependency of bistable perception on prior assumptions about the external world and on the strength of ascending and descending loops.

These dynamics can be represented by an energy landscape plotting the "potential" (the temporal integral of the dynamic Equations (2.1-2.6)) as a function of the current state L. The relationship between the energy landscape and stability of a dynamical system is shown in Fig 2.3A and 2.3B, while the actual energy landscape of the model for different parameter settings is shown on Fig 2.3C and 2.3D. In the absence of inputs, L always decreases toward the lower potentials in these energy landscapes, until it reaches a stable fixed point corresponding to a local minimum in the potential, also called an "energy well" (Fig 2.3A). The presence of a noisy input introduces random perturbations which might allow L to temporarily climb the barrier between two wells, thus switching to a different stable state (Fig 2.3B).

Without the descending loops, the model is equivalent to an HMM. Importantly, an HMM acts as a leaky integrator with only one stable fixed point (the prior) determined by the 2 rates (volatility):

$$L_{St,a=0} = \log\left(\frac{r_{\text{on}}}{r_{\text{off}}}\right) \tag{2.7}$$

This can be visualized by observing that the corresponding energy landscape contains a single energy well (Fig 2.3C, dashed line). As long as the descending loops are weak compared to the

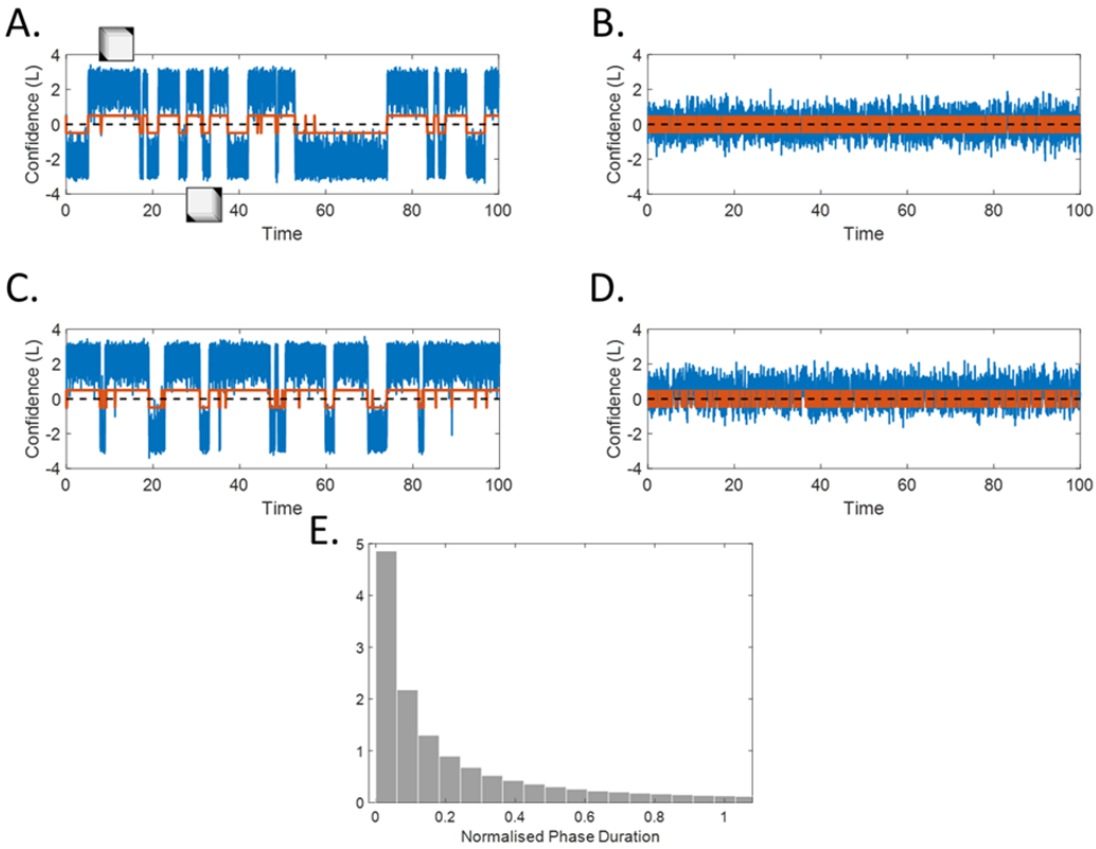

Figure 2.2: **Examples of model dynamics**. **(A)** Model with descending loops ($a_P = 1.5$), unbiased ($r_{\text{on}} = r_{\text{off}} = 0.5$), with sensory gain $w_{\text{int}} = 0.8$. The model received an ongoing, ambiguous, white noise input with standard deviation $\sigma_{\text{noise}} = 1$. Blue line: L (log-ratio of the belief / confidence), red line = percept, dashed line = decision threshold). (B) Model with no descending loops (same parameters as in (A.) except $a_P = 0$). (C) The same model as (A), but with a preference for the "SFA" configuration (transition rates changed to $r_{\text{on}} = 0.52$, $r_{\text{off}} = 0.48$). (D) The same model as (B), with $r_{\text{on}} = 0.6$, $r_{\text{off}} = 0.4$. (E.) Phase-duration histogram (No loops; unbiased). The dynamical circular inference model (with/without loops; with/without bias) predicts exponential distribution of phase-durations. Gamma-like distributions, often observed in bistable perception experiments, can be obtained by adding filtered noise, adaptation-like mechanisms or more complex decision criteria to the model (see Discussion).

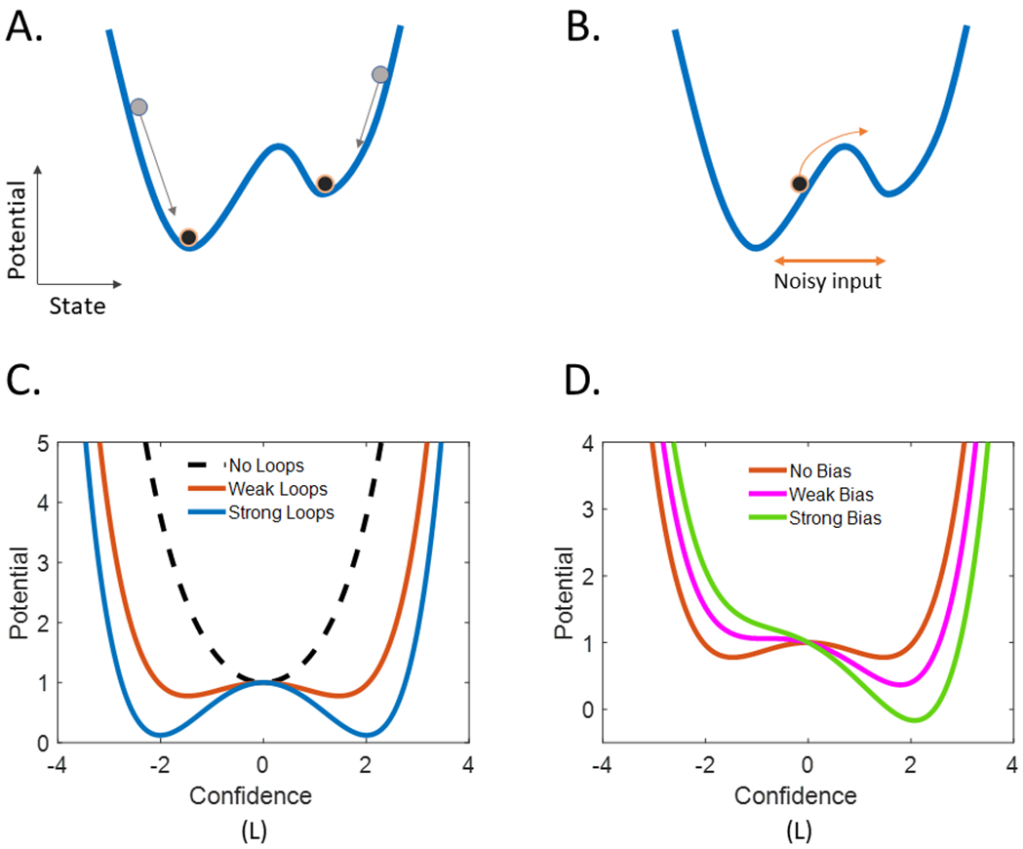

FIGURE 2.3: **Energy landscapes of the model with and without descending loops**. **(A)** Schema illustrating the relationship between wells in the energy landscape (potential = integral of the dynamic equation, in blue) and stable states. Gray and black dots represent the initial and final state from two different initial states. In the absence of external input, dots can only decrease. **(B)** Schema illustrating how noise can force the state to climb an energy barrier (a hill in the energy landscape) and switch to a different stable state. **(C)** Energy landscape of the model with no descending loops (dashed, $a_P = 0$), and two increasing levels of descending loops (red: $a_P = 1$, blue: $a_P = 1.3$). Descending loops generate a bistable attractor, whose stable fixed points correspond to (strong beliefs about) the two interpretations (blue). In contrast, a system with no loops has only one attractor, the prior, (equal to 0 in this unbiased scenario). **(D)** Energy landscape for different biases, no bias (red: $r_{\text{on}} = r_{\text{off}} = 0.5$), weak bias (magenta: $r_{\text{on}} = 0.55, r_{\text{off}} = 0.45$) and strong bias (light green: $r_{\text{on}} = 0.6, r_{\text{off}} = 0.4$). Note that for stronger biases, the non-preferred configuration becomes unstable.

leak, the prior remains the only fixed point of the system and is stable. For example, with $r_{\text{on}} = r_{\text{off}} = r$, this remains true up to the value:

$$a_p^{Pf} = \frac{r}{w_s} \tag{2.8}$$

At this value, the system undergoes a pitchfork bifurcation (Fig 2.4A; see also Appendix 2). The preexisting fixed point becomes unstable and 2 additional attractors are generated, given

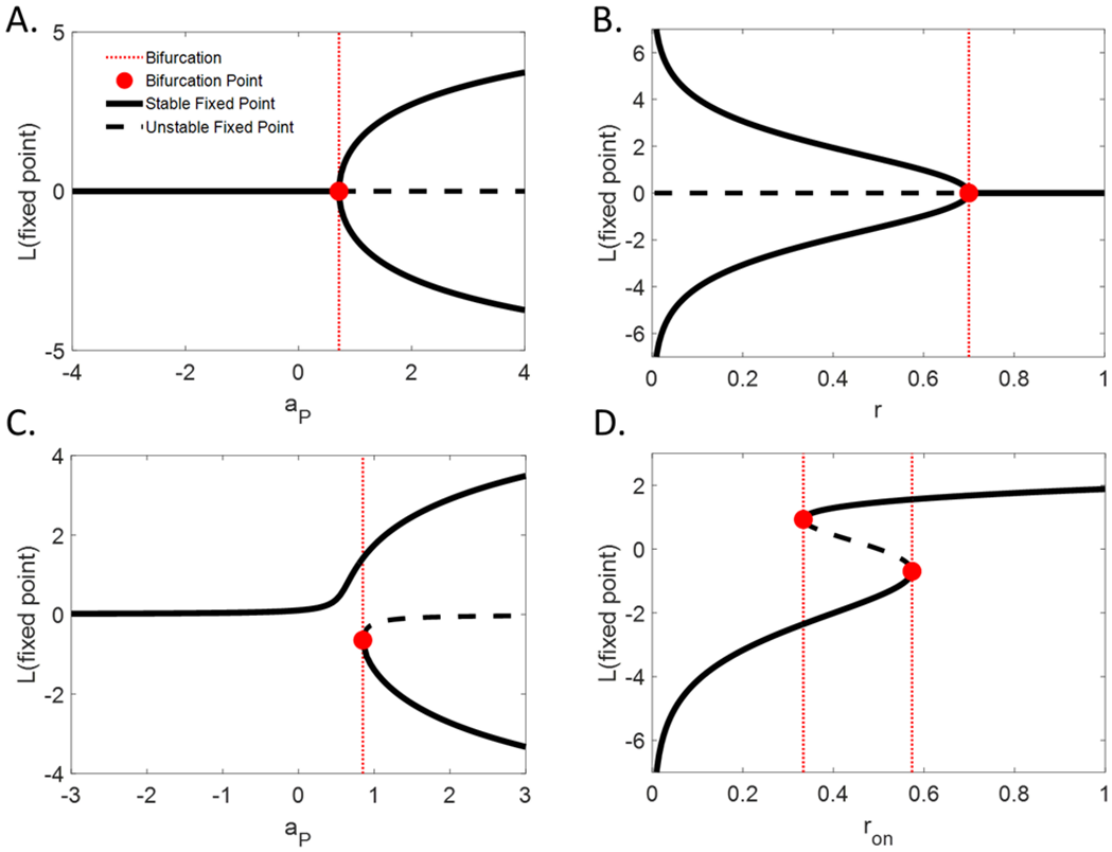

Figure 2.4: **Phase diagrams of the model dynamics**. **(A)** Stable fixed point (plain), unstable fixed point (dashed) and bifurcation point (red dot) as a function of a_P for an unbiased system ($r_{\text{on}} = r_{\text{off}} = r$). **(B)** Stable fixed point, unstable fixed point and bifurcation points as a function of r. **(C)** The same as (A) for a biased system ($r_{\text{on}} > r_{\text{off}}$). **(D)** The same as (B) but as a function of r_{on}, r_{off} being fixed at 0.5. Note that bistability can exist in a narrow range around symmetry. **(A,B)** Pitchfork bifurcation for symmetrical systems. **(C,D)** Saddle-node bifurcation for asymmetrical systems.

by the 2 symmetrical, nonzero solutions of the equation $-\Phi(L) + a_L = 0$ (Figs. 2.3C and 2.4A). The stronger the descending loops (or the weaker the leak), the further apart the 2 symmetrical attractors are, resulting in more highly trusted configurations, which are also more stable since the energy barrier is harder to cross.

Adding bias to the system ($r_{\text{on}} \neq r_{\text{off}}$; e.g., SFA bias in Necker cube) creates an asymmetry in the energy landscape (Fig 2.3D). A saddle-node (SN) bifurcation occurs when the loops become strong enough to overcome the leak (Fig 2.4C; for a mathematical description of the SN bifurcation, see Appendix 2). However, bistability can only exist in a narrow range of biases (i.e., the difference between the two transition rates r_{on} and r_{off}), more particularly in the range constrained by the 2 SN bifurcation points (one for $r_{\text{on}} > r_{\text{off}}$ and one for $r_{\text{on}} < r_{\text{off}}$; Fig 2.4D). These two bifurcations represent points at which the bias becomes strong enough to ensure that only one of the two configurations (the most likely one a-priori) can be stably perceived.

Our analysis suggests that descending loops can constitute a crucial part of the machinery of

a system exhibiting bistable perception. When they are strong enough to overcome the effect of the leak, they generate a bistable attractor, implementing a memory-like mechanism that pushes the belief toward more extreme values based on the previous observations. This helps the system make decisions and act upon them in the absence of fully convincing evidence.

Until now, our analysis focused mainly on the effects of the descending loops. However, ascending loops play an important role as well. According to (6), ascending loops increase the gain of the sensory evidence (noise) (Fig 2.3B), which consequently acts by destabilizing perception and reducing the effect of the bias on predominance.

In conclusion, this analysis demonstrates that robust bistable perception requires a very specific set of conditions. It can only exist if there is a combination of (1.) reliable sensory inputs (large w_S), (2.) stimuli that are assumed to be stable (i.e., small transition rates r_{on} and r_{off}, that are dominated by descending loops), (3.) at least two probable interpretations, even if one can dominate the other (i.e., r_{on} and r_{off} relatively close to each other, leading to a weak bias). Given these stringent conditions, it is not surprising that bistability is rather uncommon in everyday life and occurs mainly for artificial stimuli chosen to obey these requirements.

In the next sections, we explore the predictions of the model regarding well-known psychophysical features of bistable perception.

Levelt's laws An important qualitative aspect of bistable perception is Levelt's laws. These laws constitute a set of 4 psychophysical propositions relating the strength of the bistable stimulus to the phenomenology of binocular rivalry (Levelt, 1966), and more generally of bistable perception (Klink et al., 2008). Despite some recent modifications in their formulation (to account for new experimental data (Shpiro et al., 2007; Brascamp et al., 2015b)), Levelt's laws remain fundamental to our understanding of the machinery of bistability and an important crash-test for any potential model. We will present one by one the four revised propositions (as described in (Klink et al., 2008) and not in Levelt's original monograph (Levelt, 1966)) and will critically discuss them through the prism of the dynamical circular inference (dCI) model.

1st Levelt's law. The first proposition links the stimulus strength with the predominance of each interpretation. It postulates that increasing the stimulus strength of one perceptual interpretation increases the predominance of this perceptual interpretation (Klink et al., 2008). For example, adding a cue to the Necker cube helps the relevant interpretation gain more perceptual dominance compared to its rival. Although in modern terminology, proposition 1 sounds more like a tautology, it is still useful for detecting stimulus features (or parameters of the model) that affect the strength of an interpretation (Brascamp et al., 2015b). Within our model, we can parameterize the strength of the sensory evidence by adjusting the drift μ_{noise} of the Gaussian noise, which biases the sampling of evidence (Fig 2.1B). As expected, the more positive the drift the closer the relative predominance goes to 1 (the opposite for negative drift) (Fig 2.5A), in agreement with the first proposition.

2nd Levelt's law. The second proposition is less intuitive than the first and posits that manipulating the stimulus strength of one perceptual interpretation of a bistable stimulus does not influence equally the average dominance duration of both interpretations, but mainly affects the persistence of the stronger interpretation (Klink et al., 2008; Moreno-Bote et al., 2010). For example, increasing the strength of a visual cue in the Necker cube example mainly affects the mean dominance duration of the corresponding interpretation. The dCI model is fully compatible with Levelt's second law, as presented in Fig 2.5B; making the drift more positive (bias for SFA) predominantly affects the mean phase duration of the SFA interpretation (the opposite happens for a negative drift and the SFB interpretation). Indeed, the drift acts as an additional bias term in (4)/(6), which deepens the well of the strong interpretation, while making the other well shallower. This dual effect of the drift (not obvious in other models in which different

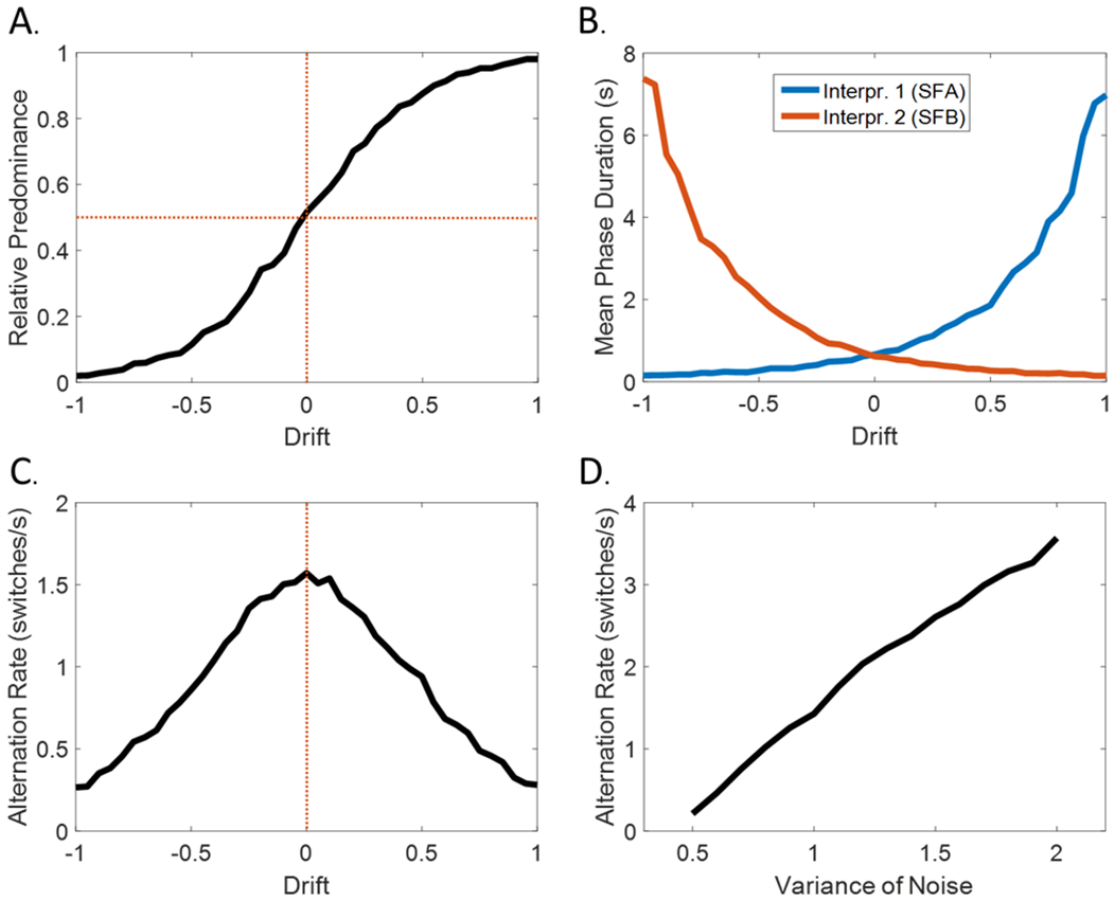

Figure 2.5: **Levelt's laws**. The circular inference model qualitatively reproduces the 4 Levelt's propositions (here: $w_S = 0.9; a_P = 1; r_{\text{on}} = r_{\text{off}} = 0.5$). **(A)** 1st proposition — Increasing the stimulus strength of one perceptual interpretation increases the predominance of this perceptual interpretation. **(B)** 2nd proposition — Manipulating the stimulus strength of one perceptual interpretation of a bistable stimulus does not equally influence the average dominance duration of both interpretations, but mainly affects the persistence of the stronger interpretation. **(C)** 3rd proposition — Increasing the difference in the stimulus strength between the 2 perceptual interpretations should result in a decrease in the perceptual alternation rate (i.e., maximum number of switches at equi-dominance). **(D)** 4th proposition — When we increase the strength of both interpretations, the number of switches increases.

variables represent the different interpretations, see also (Moreno-Bote et al., 2007)), along with the model's inherent non-linearity can explain Levelt's second law (Moreno-Bote et al., 2010).

3rd Levelt's law. Levelt's third proposition is closely related to the second proposition (Brascamp et al., 2015b) and suggests that increasing the difference in the stimulus strength between the 2 perceptual interpretations should result in a decrease in the perceptual alternation rate (Klink et al., 2008). In the Necker cube example, this proposition implies that adding a visual cue results in fewer switches. Importantly, the dCI model behaves exactly as the third proposition dictates. As shown in Fig 2.5C, the alternation rate achieves its maximum value for drift = 0 (completely ambiguous stimulus) and decreases symmetrically as the drift becomes more positive or negative, a direct con-sequence of the third law (Moreno-Bote et al., 2010).

4th Levelt's law. Finally, the fourth proposition goes one step further and discusses what happens to the alternation rate if we equally increase the strength of both interpretations. In this case, the number of switches increases, resulting in a higher alternation rate. Contrary to the 3 first propositions, the fourth proposition illustrates the effect of a simultaneous and equal manipulation of both interpretations (global stimulus strength). In the model, this should result in an increase in the mean of the absolute value of the sensory evidence, while it should have no effect on the mean of the sensory evidence per se. In other words, this global manipulation can be captured by a change in the variance in the noise distribution σ_{noise}. A higher variance results in more exploration of the energy landscape due to the noise. Consequently, as illustrated in Fig 2.5D, increasing σ_{noise} results in more switches, in agreement with Levelt's fourth law. In conclusion, the model obeys Levelt's laws regardless of the chosen parameters as long as

1. The sensory gain is high enough to induce transitions.

2. The bias is not strong enough to render one of the two configurations unstable.

Note that the respect of Levelt's laws is not sufficient to prove the presence of descending loops since the model without loops can also reproduce them (as long as the decision threshold is set appropriately). However, definite support for the existence of descending loops is provided by the stabilization of the percept by intermittent presentations of the stimulus, as described in the next section.

Intermittent presentation When an ambiguous stimulus is presented continuously, switches between competing interpretations occur randomly every few seconds, with consecutive phase durations being largely independent (Walker, 1975). Based on this observation, many researchers concluded that bistable perception is principally a memoryless process ((Lehky, 1995), see also Nawrot and Blake (1989); Pastukhov and Braun (2011)). Nevertheless, this conclusion contravenes another observation: the fact that people tend to perceive the same interpretation repeatedly when ambiguous stimuli are presented intermittently for a wide range of OFF-durations (intervals during which stimulus is absent) (Orbach et al., 1963; Leopold et al., 2002). This second observation forced researchers to assume the presence of some perceptual memory (Pearson and Brascamp, 2008), which manifests when the stimulus disappears from the screen. A variety of mechanisms implementing this memory have been proposed, including low-level mechanisms such as adaptation (combined with subthreshold effects; (Noest et al., 2007)), or high-level memory mechanisms located outside the extrastriate cortex (Leopold et al., 2002; Maier et al., 2003; Sterzer and Rees, 2008). The dCI model offers a different explanation for this stabilization effect, based on the descending loops.

In agreement with previously published experimental observations, our model predicts no significant correlation in the duration of successive phases (Walker, 1975; Lehky, 1995), as expected from a model that does not contain adaptation (or adaptation-like) mechanisms (Pastukhov and

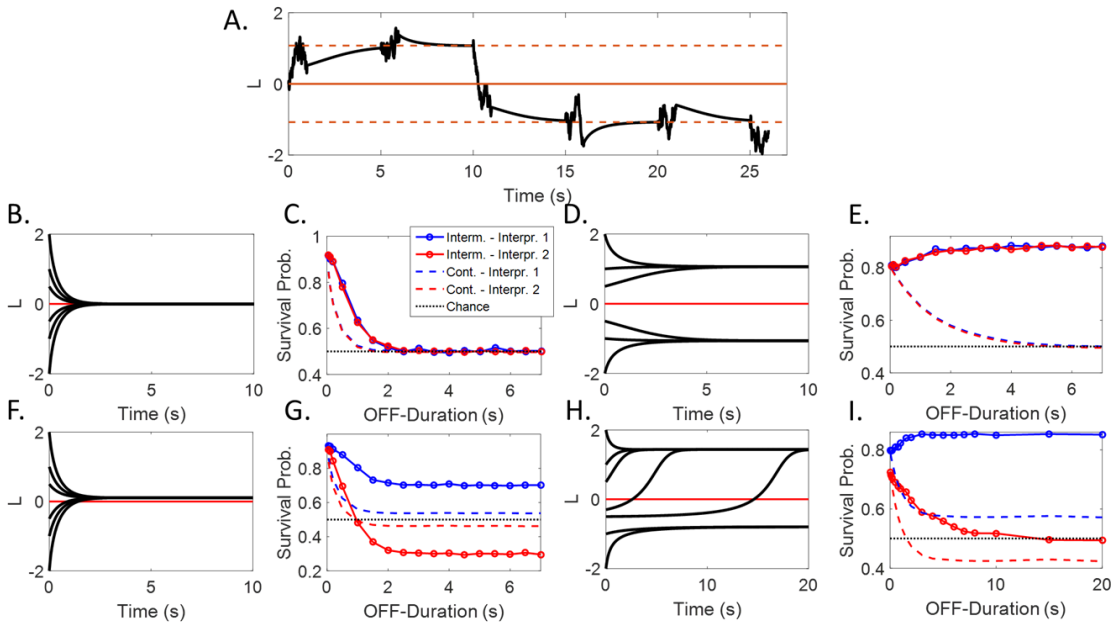

Figure 2.6: **Continuous vs intermittent presentation**. **(A)** An interpretation of the phenomenon, based on the circular inference framework. When the stimulus disappears, the belief converges to an attractor. The behavior of the system depends on the number and the value of the fixed points (here: $w_S = 1; a_P = 1.2; r_{\text{on}} = r_{\text{off}} = 1$ (symmetrical case) or $r_{\text{on}} = 1; r_{\text{off}} = 0.9$ (asymmetrical case)). **(B,C,F,G) No loops** — If there are no (descending) loops, when the stimulus disappears the beliefs converge to the prior (**(B) No implicit preference; (F) Implicit preference**). Consequently, for longer OFF-durations, the 2 survival probabilities (blue and red solid lines) either converge to 0.5 (**(C) No implicit preference**) or to symmetrical values (**(G) Implicit preference**). In both cases, the stimulus is not stabilized for longer intervals. Interestingly, it is more stable compared to a continuous presentation (dashed lines). **(D,E,H,I)** Descending loops–Descending loops generate a bistable attractor (**(D) No implicit preference (H) Implicit preference**). Crucially, when they are strong enough, they cause stabilization for longer intervals (**(E) No implicit preference (I) Implicit preference**). Furthermore, in the biased case, survival probabilities converge to asymmetrical values.

Braun, 2011). However, the model should be able to predict a stabilization effect, when the stimulus disappears for brief durations. To quantify stabilization, many studies referred to the alternation rate, which is the number of switches in a time interval (Orbach et al., 1963; Leopold et al., 2002; Kornmeier et al., 2007). However, this measure is not ideal as it can be affected by various confounding factors including different presentation durations and switches occurring during ON-durations (interval during which stimulus is present). Moreover, the alternation rate considers both interpretations together and obscures any possible asymmetries. Instead, we used the survival probability (SP) of each interpretation, which is the probability that the dominant percept at the end of an ON-duration will be dominant again when the stimulus reappears after the OFF-duration. Fig 2.6A illustrates our interpretation of the phenomenon (5 ON-OFF cycles, $a_P > 0$).

Without descending loops ($a_P = 0$), and in the absence of input (i.e., when the stimulus is

"OFF"), the belief progressively goes back to its prior value ($\log(r_{on}/r_{off})$) due to the leak (Fig 2.6B and 2.6F). For the unbiased system, the model predicts that both survival probabilities (SP) will decrease toward 0.5 (chance) with a time constant that depends on the transition rates (Fig 2.6C). An SP in a biased system would reach symmetrical points above and below chance, with the values depending on the strength of the bias (Fig 2.6G). The longer the OFF-duration, the less temporal dependency there would be between subsequent percepts. Thus, without descending loops, there could not be any stabilization of the percept by an intermittent presentation for long "OFF" durations. For comparison, SP is shown for the continuous case (stimulation is not interrupted; in which case, we measure the survival probability in constant intervals; dashed lines).

The descending loops ($a_P > 0$) change the behavior of the system. The phase portrait of this system is presented in Fig 2.6D and 2.6H. Instead of one single point where all the trajectories meet, now we observe 2 clearly distinct basins of attraction, symmetrical for an unbiased system and asymmetrical for a biased system. As a result, the temporal stability of the percept is drastically increased, especially for long "OFF" durations (Fig 2.6E). In biased systems, the level of stabilization depends on whether we consider the dominant or nondominant percept. The probability of persistence of the dominant percept (if biased) always converges to a higher probability than the nondominant percept. In the example shown in Fig 2.6I, only the dominant stimulus is stabilized by intermittent presentation, while the nondominant percept SP converges to a chance level. In other cases, both the dominant and nondominant percept can be stabilized. The stabilization of both percepts increases with the level of descending loops and decreases with sensory gain, as shown in the next section.

An important comment needs to be made. The current version of the model does not predict a destabilization occurring for small OFF-durations, usually for values below 500 ms, as reported in some studies (Kornmeier et al., 2007). Other models have attributed this observation to short-term sensory adaptation (Noest et al., 2007). To keep the model as simple as possible, we did not introduce sensory adaptation. However, such a short-term effect, occurring only at the time of stimulus presentation, would not affect the stabilization for long OFF-durations as predicted by the model with descending loops.

To summarize, dCI predicts the stabilization of bistable perception for longer OFF-periods. In addition, it makes specific predictions about the persistence of each interpretation separately, which could help to experimentally validate (or invalidate) this model.

Bistable perception as a tool for investigating mental illness So far, we have described a functional model of bistable perception, based on the notion of CI. Accumulating evidence supports the idea that circularity (and especially a small amount of descending loops) is a common property of the human brain, reflecting some inherent limitations of neural circuits (Jardri et al., 2017; Leptourgos et al., 2020b). However, it has also been suggested that CI could be the cause of several cognitive and/or perceptual disorders, including schizophrenia (Jardri and Denève, 2013a; Leptourgos et al., 2017). In a previous study, Jardri et al found that on average, patients with schizophrenia have stronger ascending loops compared to a group of matched healthy controls (Jardri et al., 2017). Additionally, it was evidenced that "positive" (i.e., psychotic) symptoms, including hallucinations and delusions, correlate with the amount of ascending loops (i.e., sensory evidence amplification), "negative" symptoms, including lack of motivation and anhedonia, correlate with the amount of descending loops (i.e., prior amplification), and finally, cognitive disorganization correlates with the total amount of loops ($a_S + a_P$). Considering these previous findings, an interesting question is what does the current dCI model predict the behavior of schizophrenia patients exposed to bistable stimuli?

Fig 2.7A and 2.7B illustrates the effect of ascending loops on the bias (relative predominance) and stability (mean phase duration). As previously shown, ascending loops increase the gain of the noise, facilitating the jumps between the 2 attractors. Consequently, our model predicts that patients with more severe hallucinations and delusions should be less biased in their responses (both due to inherent priors and visual cues) but also less stable (especially the interpretation that is supported by the visual cue). Specifically, the effect of ascending loops on relative predominance, although it might seem counterintuitive (over-counting of sensory evidence leads to a smaller effect of that evidence), illustrates the detrimental effect of the higher gain of noise on the accumulation of evidence.

In contrast, descending loops deepen the wells of the energy landscape and consequently, they produce the exact opposite effects. As shown in Fig 2.7C and 2.7D, the prediction would be that they increase both the bias and the stability of schizophrenia patients with more severe negative symptoms.

Similar stabilization and destabilization effects as a function of the level of ascending and descending loops are predicted for intermittent presentation (Fig 2.7E and 2.7F). In particular, increasing ascending loops (and thus, the sensory gain), leads to destabilization of both the dominant and nondominant percept (more precisely, both SP get closer to 0.5; Fig 2.7E). This effect is in agreement with recent experimental results on schizophrenia patients (Schmack et al., 2013, 2015). In contrast, increasing descending loops stabilizes first the dominant percept, and then both the dominant and nondominant percepts (Fig 2.7F).

Finally, note that these predictions are not only qualitative but also quantitative. The results in Fig 2.7, as well as the shape of the stabilization curves in Fig 2.6, depend on 4 free parameters, the transition rates, overall descending loop strength a and sensory gain w_{int}, all specifically related to generic parameters of perceptions applicable to many behavioral tasks. This could provide a foundation for parametric study of natural variation in the general population and psychiatric disorders, generalization over the results of different experiments (e.g., probabilistic decision tasks versus bistable perception), and raise the possibility of finding specific neural correlates of these variations (e.g., levels of E/I balance, effective connectivity between highlevel and low-level areas, etc.) (see Appendix 4).

2.2.5 Discussion

In the present paper, we demonstrated that bistable perception could arise in a perceptual system where feedback based on the current beliefs corrupts the sensory inputs. In this scenario, expectations are reverberated back up and considered several times (forming descending information-loops), suboptimally amplifying prior beliefs and causing the system to «see what it expects» (Leptourgos et al., 2017). The emerging dynamical system can explain various intriguing features of bistable perception, including its mere existence. It artificially inflates the accumulated noisy information, leading to a system that perceives clearly, persistently and in alternation the two potential interpretations, with high levels of conviction. Such a dCI model is compatible with Levelt's laws and accounts for the stabilization of the percepts when the stimulus is presented intermittently.

Importantly, this model allowed us to make new predictions regarding bistable perception in physiological and pathological conditions. Each free parameter has a clear interpretation in terms of perceptual inference, can be directly estimated from behavioral data (see Appendix 4), and can be generalized to predict behavior in other tasks (e.g., probabilistic decisions). Crucially, although descending loops could be necessary for bistability, they are not sufficient. Bistable stimuli need to lack crucial information that would clearly disambiguate them in a natural setting (such as depth cues). The perceptual system should expect the input distribution to

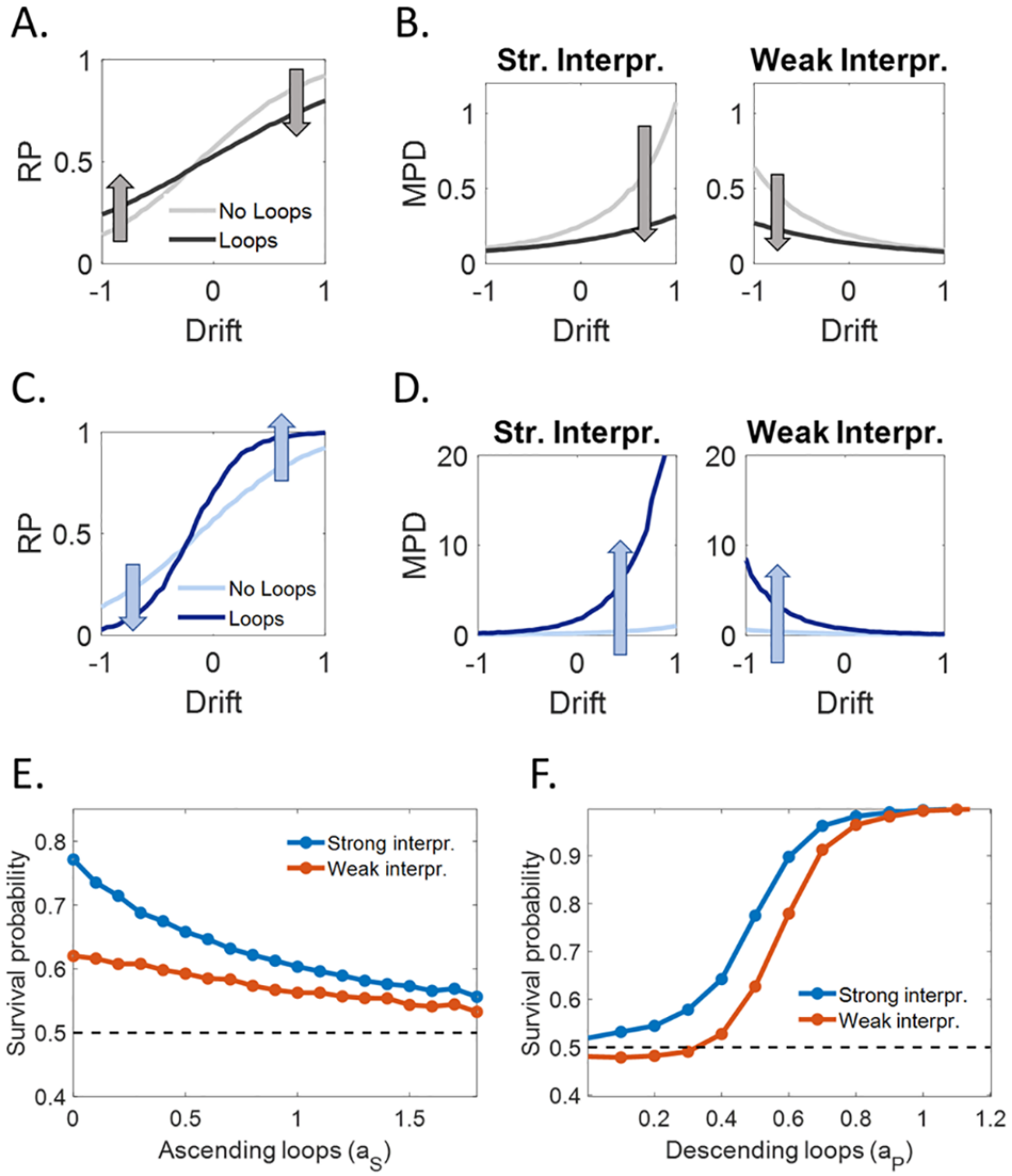

FIGURE 2.7: **Predicted effects of CI strength on bistable perception**. **(A)** Relative predominance (RP) as a function of the strength of sensory evidence in favor (positive drift) or against (negative drift) the preferred configuration (i.e., μ_{noise}) for increasing sensory gain (including ascending loops), from light to dark gray. **(B)** Mean phase duration of the preferred and non-preferred configuration. **(C)** The same as (A) but with no ascending loops and increasing descending loops, from light to dark blue. **(D)** The same as (B), with no ascending loops and increasing descending loops. **(E)** The probability of persistence of the preferred (blue) and non-preferred (red) configuration during the intermittent presentation of an ambiguous stimulus (stimulus duration 200 ms, OFF-duration 5 s) as a function of the ascending loops a_S ($a_P = 0.5$). **(F)** The same as (E), but as a function of the descending loops a_P ($a_P = 0$). All the other parameters were kept constant across simulations: $w_S = 1; r_{\text{on}} = 0.5; r_{\text{off}} = 0.48$.

differ between the two interpretations (otherwise they would be uninformative and disregarded) even if this is not the case for artificial stimuli used in bistable experiments (Fig 2.1B). Of note, completely ambiguous stimuli are, in fact, very rare (Arnold, 2011; Kersten et al., 2004) and unlikely to be learned from experience.

From the point of view of the underlying dynamics of perception, descending loops have important consequences beyond bistability. Due to their inherently stabilizing effect, a perceptual system can switch from a pure Bayesian integrator to a bistable attractor. By changing just the strength of descending loops, the perceptual system can transit between two decision-making strategies: Integration to bound (Ratcliff et al., 2016; Palmer et al., 2005) and attractor dynamics (Bitzer et al., 2015; Wang, 2002).

Beyond our model, various other implementations have been proposed to account for the unique characteristics of bistable perception. Mechanistic models have either focused on neural mechanisms (Laing and Chow, 2002; Wilson, 2003, 2007) and/or on more abstract dynamical systems (Lago-Fernández and Deco, 2002; Noest et al., 2007; Moreno-Bote et al., 2007). Nevertheless, those models are usually designed on an ad hoc basis and remain largely descriptive. With few exceptions (e.g., (Moreno-Bote et al., 2010)), they are agnostic regarding the functional implication of bistability for perception and decision in general. In other words, although they may address the «what» questions (mechanisms and implementations), they are not addressing the «why» questions (epistemological questions).

To answer the second type of question, other groups have proposed functional models of bistable perception that approach the problem in a top-down fashion (Hohwy et al., 2008; Weilnhammer et al., 2017; Sundareswara and Schrater, 2008; Reichert et al., 2011; Gershman et al., 2012; Dayan, 1998; Albert et al., 2017). Like ours, those approaches focus on the type of problems that perceptual systems usually encounter (e.g., deal with uncertainty) and impose functional limitations (e.g., Markovian statistics, approximate Bayesian inference (Bishop, 2006)). However, some of these models are abstract and do not specify neural mechanisms. Others are more complex and contain large numbers of free parameters, rendering them difficult to (in)validate experimentally.

In particular, an interesting model that bears some similarity with the dCI model was described by Hohwy and Friston (Hohwy et al., 2008) and formalized by Weilnhammer and colleagues (Weilnhammer et al., 2017). Like dCI, it relies on a message passing algorithm, but instead of Belief Propagation, it is largely based on a simplified version of predictive coding (Friston, 2008; Rao and Ballard, 1999; Spratling, 2017) — predictive coding postulates that priors explain away sensory inputs while residual prediction error signals are fed-forward to higher regions to update beliefs. Importantly, top-down effects play a crucial role in both explanations of bistability. Instead of adding (descending) loops, the predictive coding model suggests that perception is biased by a stabilization prior, which depends on the current interpretation. This prior is constantly weakened by prediction errors emerging from evidence for the suppressed percept, via an exponential decay mechanism. A switch occurs when the evidence for the suppressed percept surpasses that for the dominant percept. Despite their similarities, the two models are not identical. While dCI is derived from first principles (inference in a Hidden Markov Model, corrupted by loops), the predictive coding model relies on a number of ad-hoc assumptions, that nuance its normative character. For example, the precision of the stabilization prior is renormalized after each switch, resulting in strong and stable percepts; this is an important assumption, yet it's difficult to interpret it from a normative perspective.

Furthermore, several models were based on the idea that inference is approximated by a sampling process, without explicit calculation and knowledge of the exact posterior distribution (Sundareswara and Schrater, 2008; Reichert et al., 2011; Gershman et al., 2012). In that case, bistable perception occurs because the perceptual system is assumed to take only one sample

at each time step, resulting in high temporal correlations between samples. This is, in fact, a nuisance in this kind of algorithm, predicting a highly suboptimal form of perceptual inference (e.g., it takes a very long time to infer the exact probability distribution, and the corresponding estimates are much more variable than a maximum-a-posteriori estimate). Because of this limitation, perceptual inference by sampling might be far less performant than belief propagation (even with loops), raising the question of why our perceptual system would choose such a strategy. Additionally, it remains unclear whether those models could account for less trivial experimental results, including stabilization under an intermittent presentation.

Note that in our case, bistable perception could also be seen as a suboptimality resulting from descending loops (i.e., the estimated probability are not the correct ones given the real sensory evidence and prior knowledge). However, we predict that it mostly affects perception in rather unusual cases, e.g., for a fixed level of descending loops, stimuli that are both expected to be very reliable (high w_S) and in reality are highly ambiguous (μ_{noise} close to zero). Consequently, this unusual stimulus does not fit our generative model (Beck et al., 2012). The effects could be far more subtle otherwise. In agreement with this hypothesis, we found that CI only rarely affects choices in randomly selected probabilistic inference problems (i.e., random graphs, see (Jardri and Denève, 2013a)). The dCI model presented in this paper is normative (i.e. derived from first principles; strictly speaking, normativity is violated due to the loops) but can also be seen as descriptive due to its closed-form solution. Switches in perceptual bistability are driven by noise in agreement with existing evidence (Shpiro et al., 2009; Panagiotaropoulos et al., 2013; Huguet et al., 2014). In contrast to models based on lateral inhibition between local populations, bistable perception is interpreted as a brain-wide phenomenon linked to inhibitory control of feedforward and feedback processes (as is generally required for hierarchical perceptual inference (Jardri and Denève, 2013a)). Its dynamical behavior has important similarities with that of other attractor models (Moreno-Bote et al., 2007), but the bistable attractor is hereby not imposed to explain certain features of bistability, but instead a direct consequence of the descending loops. In the same vein, our model makes a clear distinction between a bias induced by sensory evidence and bias resulting from the system's implicit preference (prior knowledge), thus enabling the generation of asymmetries in the absence of stimulation (intermittent presentation).

Another important feature of bistable perception, shared by human and nonhuman observers, is the distribution of dominance durations. Although there is considerable variability in the mean phase duration between participants (but also within participants and between conditions or stimuli), there is an impressive similarity in the shape of the distribution of phase durations, relatively well approximated by a gamma or log-normal distribution (Levelt, 1967; Zhou et al., 2004; Gigante et al., 2009) (but see also (Brascamp et al., 2005)). The dCI model, like all the noise-driven attractor models, generates exponential distributions of phase durations (Moreno-Bote et al., 2007). Several extensions of the model can engender gamma-like distributions, in which simple mechanisms are added on an ad-hoc basis. For example, one could assume that inference is preceded by filtering, which takes place at the very first levels of the sensory hierarchy (e.g. retina, LGN in case of visual inputs); filtered noise is smoother than Gaussian noise and precludes the occurrence of fast switches. Alternatively, one could introduce an adaptation-like mechanism (see also (Moreno-Bote et al., 2007)); in the dCI context, this could be implemented as time-dependent transition rates, e.g. as a form of learning. Finally, a third option is to replace MAP with a more complex decision criterion, e.g. a more conservative criterion, implemented as a moving threshold, where switches occur only when there is substantial evidence in favour of the opposite interpretation.

It has been argued that CI are linked at the neurophysiological level to an imbalance between neural excitation and inhibition in favor of excitation (Leptourgos et al., 2017; Jardri et al., 2016). This imbalance might concern only local microcircuits, encompassing pyramidal cells and

local interneurons (Fig 2.1D), or more global networks, potentially involving thalamocortical or corticostriatal long-range connections (Leptourgos et al., 2017). Although both are plausible implementations of loops, local interneurons make a better candidate in the particular case of bistable perception. Indeed, it has been argued that bistability is a rather low-level process mainly occurring within the visual cortex ((Blake and Logothetis, 2002; Brascamp et al., 2013, 2015a); but see (Lumer et al., 1998; Sterzer and Kleinschmidt, 2007), arguing for the involvement of high-level areas) while the involvement of local inhibition is also supported by pharmacological evidence (van Loon et al., 2013). Apart from normal brain functioning, CI has been used to account for clinical dimensions in schizophrenia (Jardri and Denève, 2013a; Jardri et al., 2017). Our model implies that generic mechanisms involved in hallucinations and delusions could also explain common perceptual phenomena, such as bistable perception, in agreement with the idea that psychosis may exist along a continuum with normal experience (Waters et al., 2016; Alderson-Day et al., 2017; Baumeister et al., 2017; Powers et al., 2017). Nevertheless, when and how exactly those mechanisms go awry and generate pathological symptoms remains an open question. In addition, the present model provides a dynamical system interpretation of CI models, relating them to other influential frameworks (Loh et al., 2007; Rolls and Deco, 2011; Adams et al., 2018). Could circularity offer a relative advantage to perceptual systems or is it simply a manifestation of the inherent limitations of neural systems? Our present results suggest that a system performing exact inference with ambiguous information could be more vulnerable to noise and have difficulties in forming stable percepts. Moderate descending loops could improve the system, allowing rapid and robust decisions even when evidence is not conclusive (after all, both "fighting" and "fleeing" are better than standing still; a similar explanation was suggested by Moreno-Bote and colleagues, who interpreted bistability as exploratory behavior under uncertainty (Moreno-Bote et al., 2010)). Moving a step further, a system with flexible descending loops (e.g., a system that can regulate its E/I balance through neuromodulators, such as dopamine, serotonin or acetylcholine (Lucas-Meunier et al., 2009; William Moreau et al., 2010)) could vary the perceptual strategy from impulsive to deliberative in accordance with task requirements. This suggestion, although speculative, could reconcile the present results with evidence showing a balance between excitation and inhibition at different scales (Wehr and Zador, 2003; Okun and Lampl, 2008; Xue et al., 2014) and is furthermore easily testable (e.g., by measuring E/I balance during bistability and during stimulation with unambiguous stimuli). In conclusion, we described bistable perception as a probabilistic inference process, under the influence of amplified priors due to the presence of descending loops in the cortical hierarchy. The model explains why bistable perception occurs in the first place and qualitatively predicts several of its properties. Additionally, it has important implications for the neural correlates of bistability and the relation between normal brain functioning and pathology, ultimately linking computation, behavior and neural implementation.

2.2.6 Supplementary material

The supplementary material of the article, which includes mathematical derivations, bifurcation analyses and investigates the parameter recovery of the model, can be found online at `https://doi.org/10.1371/journal.pcbi.1008480`.

2.3 Circular BP as model of schizophrenia

This section corresponds to the following published article: *Circular inference predicts nonuniform overactivation and dysconnectivity in brain-wide connectomes*, by V. Bouttier, S. Duttagupta, S. Denève and R. Jardri (2021), Schizophrenia Research.

The work characterizes the differences between psychosis/schizophrenia (modeled by the Circular BP algorithm) and normal functioning (modeled by BP). We point here at the underlying assumption that the Belief Propagation algorithm is a good model for "normal" functioning of the human brain. This assumption will be revised in chapter 3, where we propose that Circular BP with "inverted circular inference" is a better model than BP for the healthy brain: see discussion in section 3.7. However, this by no means implies that the work presented here is wrong: indeed, in the specific graphs considered, BP performs approximate inference with relatively high quality. More precisely, Mooij and Kappen (2004) shows that for scale-free networks, the validity of the BP approximation scales very well with network size (the intuition is that such networks resemble a forest of sparsely-interconnected hubs and that BP is exact on acyclic graphs). Additionally, as explained in section 2.3.8.2, probabilistic graphs were weighted in order for BP to be rather of good quality (for instance, sets of weights which involved frustration or bistable dynamics were not considered).

An idea naturally arising from the work is that if Circular BP was implemented in the brain, for instance through a rate model, then it would be possible to catch its signature through brain imaging. An interesting lead would be to fit the proposed neural model to data, therefore estimating the parameter α from Circular BP without behavioral studies but instead with neural data purely. As a reminder, this parameter relates to the level of "circularity" in the inference, defined by the distance between the parameter α used or fitted and the one achieving the best possible quality of inference; see chapter 3. Eventually, such fitted α could be related to experimental measures of the E-I imbalance (through the concentration of GABA and glutamate in the cortex), which was the initial intuition behind this parameter and more generally the Circular BP algorithm.

2.3.1 Abstract

Schizophrenia is a severe mental disorder whose neural basis remains difficult to ascertain. Among the available pathophysiological theories, recent work has pointed towards subtle perturbations in the excitation-inhibition (E/I) balance within different neural circuits. Computational approaches have suggested interesting mechanisms that can account for both E/I imbalances and psychotic symptoms. Based on hierarchical neural networks propagating information through a message-passing algorithm, it was hypothesized that changes in the E/I ratio could cause a "circular belief propagation" in which bottom-up and top-down information reverberate. This circular inference (CI) was proposed to account for the clinical features of schizophrenia. Under this assumption, this paper examined the impact of CI on network dynamics in light of brain imaging findings related to psychosis. Using brain-inspired graphical models, we show that CI causes overconfidence and overactivation most specifically at the level of connector hubs (e.g., nodes with many connections allowing integration across networks). By also measuring functional connectivity in these graphs, we provide evidence that CI is able to predict specific changes in modularity known to be associated with schizophrenia. Altogether, these findings suggest that the CI framework may facilitate behavioral and neural research on the multifaceted nature of psychosis.

2.3.2 Introduction

Cognitive dysfunctions (e.g., impaired attention, working memory, or abstract thinking) and aberrant beliefs and perceptions (e.g., delusions and hallucinations) are prevalent features of schizophrenia. Numerous studies have attempted to decipher the neurobiological bases of these symptoms mostly using brain imaging or pharmacological methods. However, given the com-

plexity of the results, psychosis retains much of its mystery. An essential difficulty is due to the absence of a dominant framework able to relate the widely different levels of analysis available. Computational approaches represent a nascent attempt at bridging these gaps (Adams et al., 2013; Anticevic et al., 2015; Fletcher and Frith, 2009; Krystal et al., 2017; Sterzer et al., 2018). In this paper, we propose a new computational method to relate psychosis with impaired global brain dynamics based on two simple hypotheses. First, the brain is an inference machine (Knill and Pouget, 2004; Lochmann and Denève, 2011). Second, psychosis is associated with imbalances between excitation (E) and inhibition (I) in local neural circuits (Foss-Feig et al., 2017; Jardri et al., 2016; Lisman, 2012; Sohal and Rubenstein, 2019).

We know that structural and functional brain networks exhibit massive changes in patients with schizophrenia (Brandl et al., 2019). This finding is compatible with the common theory assuming that psychotic disorders result directly from anatomical-functional dysconnections (Friston et al., 2016; Friston, 2020; Murray and Anticevic, 2017; Stephan et al., 2009; Yang et al., 2016) and that small functional dysfunctions can easily spread between linked elements within unimpaired complex networks (Carrera and Tononi, 2014; Fornito et al., 2015; Pantano et al., 1986; Price et al., 2001).

However, the exact mechanisms underlying this breakdown of integration between widely distributed brain areas are poorly understood. These impairments do not seem related to macroscopic lesions and are more likely related to subtle and diffuse deficits at the microscale (e.g., impaired neuromodulation or synaptic plasticity and E/I imbalances). Unfortunately, consensus regarding this topic is lacking.

How should the field of computational psychiatry proceed in the face of such uncertainty? One possible strategy is to directly explore the influence of different candidate mechanisms on brain circuits (e.g., using large-scale modeling of intact and impaired neural networks) to attempt to predict (nontrivial) neural and behavioral effects. Another strategy is to set aside the complexity of the real brain in favor of normative models of belief/behavior formation in humans before searching for signatures of these processes in neural signals. Finally, some recent approaches initially proposed normative models but further proposed (highly simplified) neural mechanistic models that may account for aberrant belief formation (Adams et al., 2013; Jardri and Denève, 2013a). Notably, these strategies are unlikely to succeed on their own. A successful model should quantitatively (and qualitatively) account for behavioral and neural data, even when tested outside of its "area of comfort" (e.g., the task it was specifically designed for).

A promising framework to achieve such a goal considers the fact that a major brain function is to build internal predictive representations of its uncertain sensory-motor environment (Doya et al., 2007). Roughly speaking, brain circuits would mirror an underlying hierarchy of causes with sensory inputs at the bottom and more abstract knowledge/ context at the top (Fig. 2.8a). Inference in such a system occurs by integrating information propagated in opposite directions within neural circuits with sensory information "climbing" the hierarchy through feedforward connections, while prior knowledge descends the hierarchy using feedback connections (Fig. 2.8b, see also "Summary of methods"). In their simplest expression, these models apply the Bayes theorem in which priors (top-down predictions) and likelihoods (bottom-up sensory information) are combined with weights corresponding to their reliability. This approach is equivalent to correcting the prior with a prediction error (see also Aitchison and Lengyel (2017) for a critical discussion regarding Bayesian inference and predictive coding).

While the specific neural mechanisms underlying inference are still highly controversial, the different corresponding computational models have much in common. The brain structure is assumed to represent an underlying probabilistic structure (Parr and Friston, 2018). Neural activity represents probabilities or probabilistic updates (Pouget et al., 2013). Connection strength represents how reliably variables are able to predict each other's state, and inference is performed

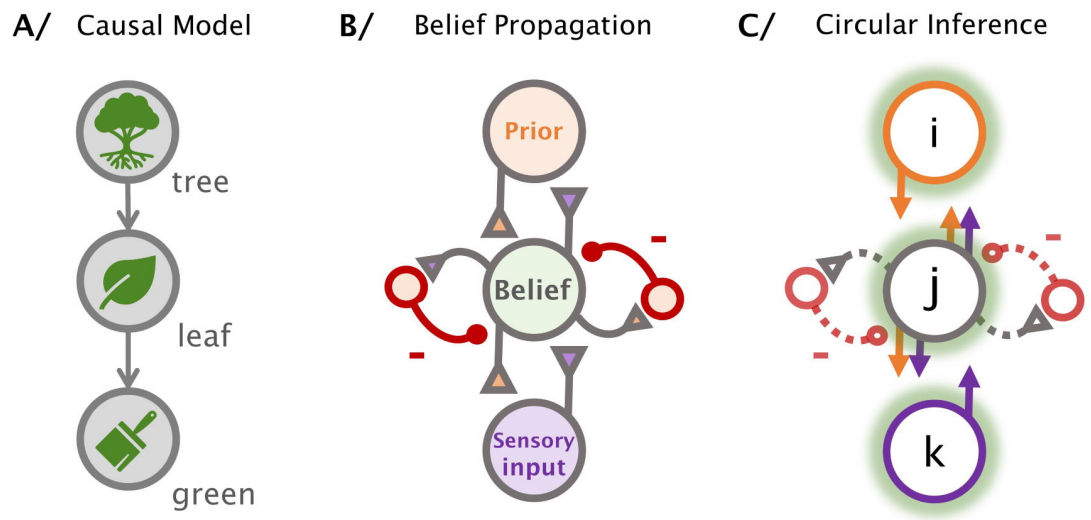

FIGURE 2.8: **Principles of Belief Propagation and Circular Inference**. **(A)** Toy example of a hierarchical causal model of three nodes representing hidden variables. The sensory input for the color green is given at the bottom of the hierarchy, while the prior expectation of a tree is given at the top. The beliefs in each node are shown in green. **(B)** A possible implementation of the belief propagation algorithm by a neural network. The information shared between the different nodes of the network is under the control of inhibitory interneurons (shown in red), which remove redundant information from messages. **(C)** In the case of circular inference, an impairment in the interneurons (dotted lines) causes an uncontrolled reverberation of messages in the network, leading to aberrant beliefs (depicted here with green halos).

by propagating local messages (beliefs, predictions, prediction errors, etc.) through these connections. For example, given some sensory evidence for the color green and a prior belief of walking under trees, the up and down propagation of messages allows computing the probability of perceiving leaves in the environment (Fig. 2.8a and b).

However, these computational models also differ in the assumed impairments at the roots of aberrant beliefs, such as those that may occur during psychosis. For instance, certain connection types could be disproportionately strong (e.g., an overweighting of top-down messages would result in priors dominating the percept (see Corlett et al. (2019) for a review), which corresponds to changes in the generative model (Parr et al., 2019). Alternatively, we hypothesize that the generative model is unchanged and that the inference mechanism (message-passing scheme) is dysfunctional as follows: messages could be uncontrollably reverberated and amplified through feedforward/ feedback loops (Fig. 2.8c) and, in turn, drive the perceptual content. Indeed, we previously showed that such a form of circular inference (CI) could be a direct consequence of impaired inhibitory control in hierarchical brain circuits (Denève and Jardri, 2016; Jardri and Denève, 2013a; Leptourgos et al., 2017).

If valid, such theoretical models should be able to capture individual behavior using a minimal set of parameters. For instance, we found that different levels of CI could account for non-pathological (e.g., illusions (Notredame et al., 2014), bistable perceptions (Leptourgos et al., 2020a,b)) and pathological behaviors, such as the heterogeneous features/dimensions of

schizophrenia (Jardri et al., 2017). Fewer studies validating probabilistic models at the neural level have been performed e.g. Powers et al. (2017). A major difficulty is that large-scale brain networks are far more complex than the simple hierarchical chains used in toy examples or to describe experimentally designed tasks.

The goal of this paper is to provide a proof of concept for extending the CI model to brain-wide neural activity with possible applications in the context of psychosis. More precisely, we show that it is possible to apply CI to simplified brain-like (abstract) graphs or brain-based connectomes (Bullmore and Bassett, 2011). We predicted the basic impairments due to CI in these graphs at both the activity level and the functional connectivity level. Finally, we compared these predictions with common fMRI findings of dysconnectivity, as observed in schizophrenia.

2.3.3 Summary of methods

This section succinctly describes how the graphs were generated and how their activity was stimulated. For more details, see Supplementary Material. The code (in Python) is available online at `github.com/VincentBt/`. We randomly generated modular small-world graphs (two common properties of brain-like networks - Fig. 2.9a, b, d), which we call "abstract graphs". Nodes within the graphs were assumed to receive randomly fluctuating, temporally smooth inputs (insets in Fig. 2.9c). The goal of using randomly generated graphs was to predict dynamic properties independent of the specific structure of the network. We used random input patterns to mimic resting-state brain activity as opposed to task-based functional patterns.

Inference was performed in these graphs by continuously propagating messages in multiple directions along the links using a local message-passing algorithm called belief propagation (BP) (Bishop, 2006; Friston et al., 2017). The confluence of messages in a given node was used to compute its belief, to be understood as a local estimate of the probability that the binary variable encoded by the node is 1 ($b = p(X = 1)$), given the currently available evidence. Thus, the nodes in the generated graph constantly attempt to reach an agreement by exchanging predictions regarding each other's states. The "sensory" evidence provided to the network (the random inputs) smoothly changed over time, as did the beliefs, as exemplified in Fig. 2.9c. Importantly, the reverberation of messages is avoided in BP by removing the message previously sent in the opposite direction from each message, which is carried out by inhibition in our proposed neural implementation (Fig. 2.8b). For example, when a tree predicts leaves, leaves should not subsequently predict a tree by total circular reasoning (this circularity is illustrated in Fig. 2.8c). To implement CI with increasing severity, we progressively decreased this correction using parameter α representing the level of inhibitory control and using values from 60% (strong circular reasoning - impaired inhibitory control) to 100% (BP - perfect inhibitory control).

We analyzed the statistics of the beliefs generated in these graphs for normal inference (BP) and increasing amounts of pathological inference (CI). As a sanity check, the predictions obtained by investigating the abstract graphs were reproduced using a specific but more realistic brain-based graph or "realistic connectome" by utilizing the set of reconstructed group-averaged fiber tracts from the open-access HCP-842 MRI atlas (Yeh et al., 2018) combined with a collection of 86 anatomical parcels (see Supplementary Table) taken from the AAL2 atlas (Rolls et al., 2015). We referred to a canonical division of nodes in a priori communities to define structural modules (also called clusters, communities, or groups) as proposed by Bertolero et al. (2018) (see Fig. 2.13).

Nodes are divided into the following three categories: connector hubs, local hubs, and other nodes. Connector hubs are nodes with connections that are diversely distributed across modules. Local hubs are nodes that are highly connected within their own module (and are not connector hubs). Other nodes are all nodes that are not connector hubs or local hubs. Connector hubs

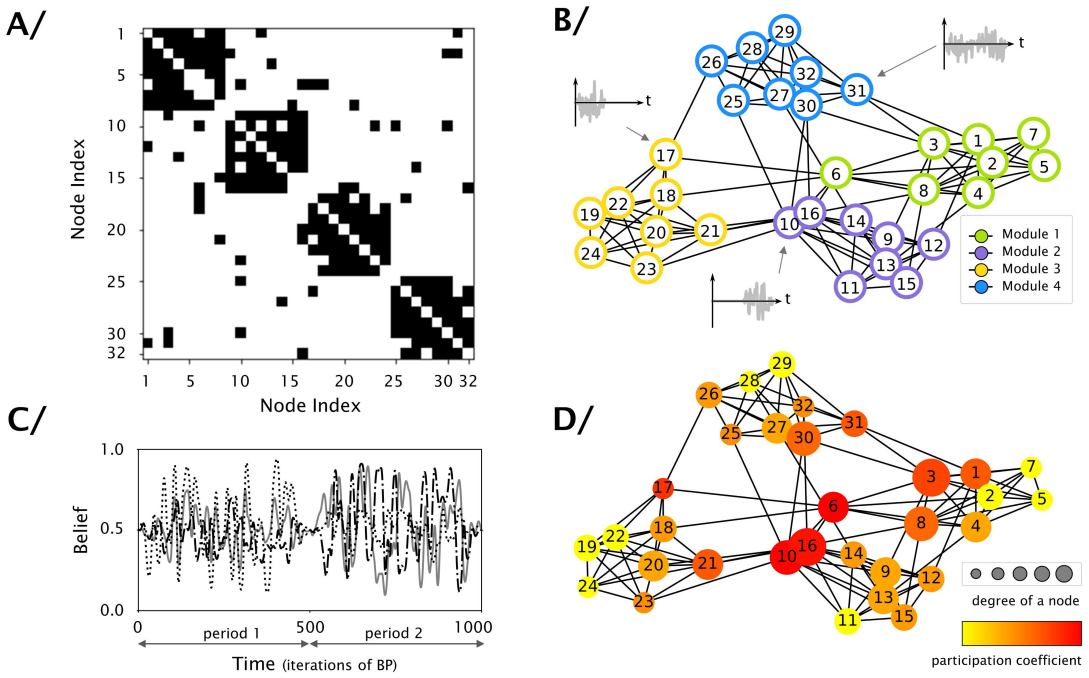

FIGURE 2.9: **Running belief propagation in abstract small-world networks**. Graphical networks are randomly generated with small-world properties and a modular structure consisting of 4 modules with 8 nodes per module. **(A)** Adjacency matrix of one of the networks generated. **(B)** Graphical representation of the network. Each of the 4 modules gathers nodes of a given color. All nodes receive a randomly fluctuating, temporally smooth input (insets). **(C)** Illustration of the temporal evolution of beliefs (probability estimates) in the network using proper inference (i.e., belief propagation). We present the beliefs in 3 nodes randomly selected from the graph. **(D)** Graphical representation of the same network presented in (B) but using a yellow-to-red color code to reflect the participation coefficient and node size for the degree.

are defined based on the participation coefficient (a measure of the diversity of intermodular connections of the node), and local hubs are defined based on the within-community strength (a measure of the locality of the node through its intramodular connections). See Supplementary Material for a formal definition of the node types and graph metrics.

2.3.4 Neural interpretation

To interpret the graph dynamics in neural terms, we need to decide how beliefs translate into neural activity, which is a topic that is still controversial. For simplicity, it was assumed that activity in a brain parcel covaried with the confidence level of the corresponding node (i.e., the absolute value of $\log(b/(1-b))$ where belief b is the probability that the binary node is in state 1 - see Supplementary Material for more detail). Thus, the more certain a node was of its variable state at a given time (in the case of binary variables, the closer its belief was to 0 or 1), the more active the node was. Note that some studies identified a link between neural activity and surprise (Schwartenbeck et al., 2016), which approximately corresponds to large temporal

fluctuations in beliefs. In our simulations, the two quantities were strongly correlated, and both predicted essentially the same effects on functional connectivity (see also Supplementary Material). Assuming a neural representation of surprise instead of beliefs would not change any of the conclusions presented here.

As previously mentioned, one cannot make a direct and naive parallel between BP (or CI) and the dynamics of a brain-like network with a matching connectivity structure. If the graph links are indeed analogous to recurrent connections between corresponding neural populations, reciprocal anatomical connections cause positive feedback, i.e., reverberation of messages. The cancellation of the reverberated part of the messages in BP is proposed to be carried out by inhibitory control, which can occur locally as shown in Fig. 2.8b or through long-range connections (Leptourgos et al., 2017). This would imply that BP corresponds to the dynamics of a superbalanced brain in which recurrent loops are constantly controlled by tight local inhibition (Denève and Machens, 2016). Presumably, relaxing this inhibitory control results in an increased CI, which is measurable at the behavioral level, even if signatures of this process at the neural level remain to be found.

2.3.5 Circular inference in abstract graphs

We first report the effects observed on randomly generated graphs. We observed that CI induces overconfidence and, thus, generates an excess of neural activity. On average, the CI-generated confidence levels are indeed higher as reflected by a sigmoidal relationship between the CI and BP-computed posterior probabilities (Fig. 2.10a, upper panel). Similarly, the distribution of beliefs among all nodes extends further towards extreme values at higher levels of CI (Fig. 2.10b, upper panel), and this result persists when bounding the belief under belief propagation between 0.4 and 0.6 (Fig. 2.15). Thus, while BP generates graded beliefs in proportion with the weak and/or contradictory evidence provided to the network (i.e., fluctuating inputs), CI causes more extreme levels of certainty.

In reality, this average relationship at the network scale hides a large amount of heterogeneity (one-way ANOVA, $F(2, 894) = 215$, $p < 0.001$). The effect in some nodes is much stronger than that in other nodes (dependent on the local structure of the network as described later in Fig. 2.11b). In the most affected nodes, CI causes beliefs to saturate to extreme values in a large portion of what should normally be their response range (Fig. 2.10c, upper panel). Thus, these nodes not only are aberrantly confident but also become insensitive to small fluctuations in their input messages and, thus, are presumably unable to transfer information to nodes downstream in the network. This finding suggests that CI not only causes overconfidence but also, somewhat counterintuitively, weakens the communication between nodes.

Upon closer examination, one finds that the variations in overconfidence induced by CI are explained by only a few properties characterizing the centrality of a node within the graph (Fig. 2.14). The nodes most affected by CIs are connector hubs whose connections are diversely distributed across modules (post hoc comparisons using t-tests for independent samples revealed that connector hubs exhibited significantly higher confidence than local hubs, $p = 2.58$ e-22, or other nodes, $p = 1.69$ e-88; it should be noted that local hubs, which are nodes that are highly connected within their own module, also significantly differ from the other nodes, $p = 8.6$ e-20; see also Fig. 2.11). For a formal definition of connector hubs, local hubs, and other nodes, see Supplementary Material. These results concerning overconfidence also apply to overactivation (excess of neural activity). Indeed, overconfidence and overactivation have the same definition in the model (see Supplementary Material). Consequently, there is an overall overactivation of the network, especially in the network hubs.

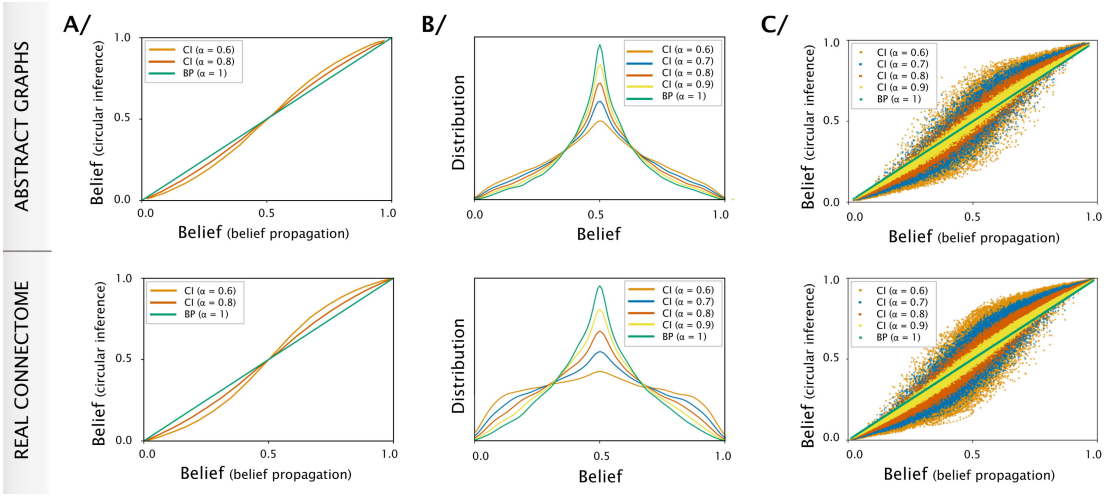

FIGURE 2.10: **Effect of circularity on beliefs in the abstract small-world networks and the real connectome**. The results based on belief propagation in randomly generated small-world modular graphs are presented in the upper panels, and those for the realistic connectome network are presented in the lower panels. **(A)** Plot of the posterior probabilities as measured by circular inference (CI) against the same probabilities from belief propagation (BP), averaged. Decreasing the level of inhibitory control α (i.e., increasing the level of CI) causes the nodes to have greater confidence compared to BP (where $\alpha = 100\%$ represents BP). **(B)** Distribution of beliefs in a single node while varying the degree of circularity. Lower inhibitory control causes more extreme beliefs. **(C)** A comparison of the beliefs under CI in one node against the same beliefs under BP. CI causes the nodes to saturate towards more extreme beliefs.

Since connector hubs exert maximal control over long-range communication within and between modules, one could expect that as the severity of CI increases, the network becomes more strongly modular with weaker functional interactions at long-range relative to short-range. These expected consequences of CI are confirmed when directly measuring functional connectivity based on the graph responses (Fig. 2.12a). Here, functional connectivity is defined as the amount of correlation between all pairs of nodes (see Supplementary Material). Measuring such functional connectivity provides an approximate idea of the underlying structure (anatomical connectivity) of the network. As expected, truly connected nodes exhibit strong correlations (increased by CI, see Fig. 2.16), while nodes separated by longer paths exhibit lower levels of correlation.

Finally, as predicted from the effects of CI on hubs, the degree of modularity of the network (defined as the intra-inter modular ratio of the functional connectome, see Supplementary Material) significantly increases with CI as functional connectivity increases within a given module but comparatively decreases between different modules (Fig. 2.12b). Thus, the network becomes less able to process information at the global scale while comparatively sparing intramodular (local) communication.

These predictions might appear counterintuitive given that reverberation could appear to increase rather than decrease the amount of global communication between nodes. However, these simulation results show that the crucial element in transferring information between different parts of the graphs is to keep beliefs graded and driven by external inputs and afferent messages rather than saturated by internal recurrent dynamics. Thus, CI paradoxically predicts both

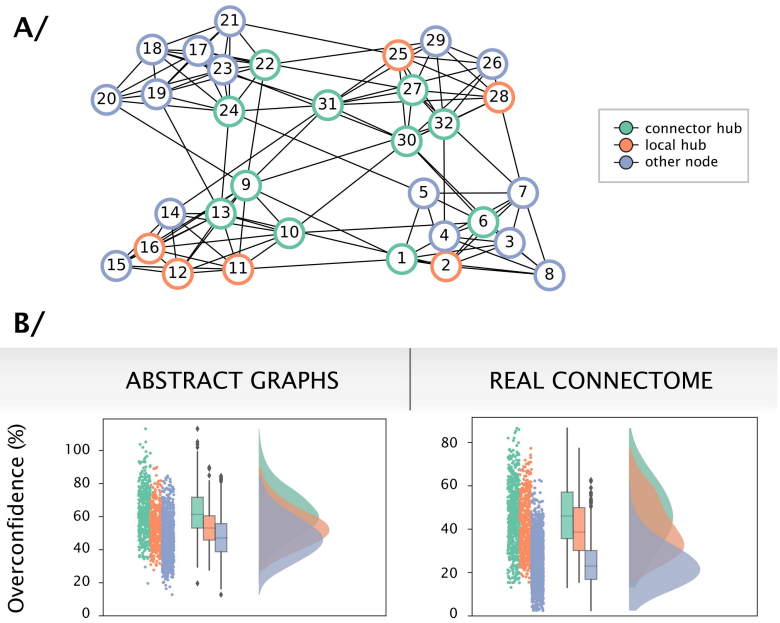

Figure 2.11: **Factors causing overconfidence due to circular inference**. **(A)** The following three types of nodes were considered in the graphs: connector hubs (nodes with connections that are diversely distributed across modules – shown in green); local hubs (nodes that are highly connected within their own module – shown in orange) and other nodes (shown in blue). **(B)** Overconfidence measured in the random graphs and realistic connectome according to the type of nodes (rain-cloud plots), for $\alpha = 60\%$. Connector hubs are significantly more overconfident than local hubs, which are significantly more overconfident than the other nodes in the network. The results are the same when examining overactivation.

overconfidence (especially in variables represented in associative or multimodal brain areas) and a deficit in global processing of information with a marked decrease in long-range (functional) connectivity relative to local connectivity.

2.3.6 Circular inference in a more realistic brain connectome

As observed in the randomly generated graphs, the implementation of CI in a more realistic connectome resulted in overconfidence, especially in the connector hubs (Figs. 2.10 and 2.11 lower panels, $F(2, 930) = 903$, $p < 0.001$; post hoc comparisons: connector hubs > local hubs: $p = 4.0$ e-12; connector hubs > other nodes: $p = 6.0$ e-323: local hubs > other nodes: $p = 3.5$ e-186). A linear model was trained to predict overactivation induced by circular inference (see Supplementary Material for a formal definition of overactivation) based on various graph measures characterizing the centrality of each node (which is a proxy for the amount of control a node has over communication in the network). This linear model trained on randomly generated graphs or "abstract graphs" accurately predicted the level of overactivation in each node in the connectome (up to a constant of proportionality, Pearson's r correlation = 0.97, $p = 1.8$ e-52; see also Fig. 2.17).

We ranked the 86 parcels of this connectome according to the level of overactivation caused by CI. Represented on the same scale, we can observe how strongly the degree centrality and,

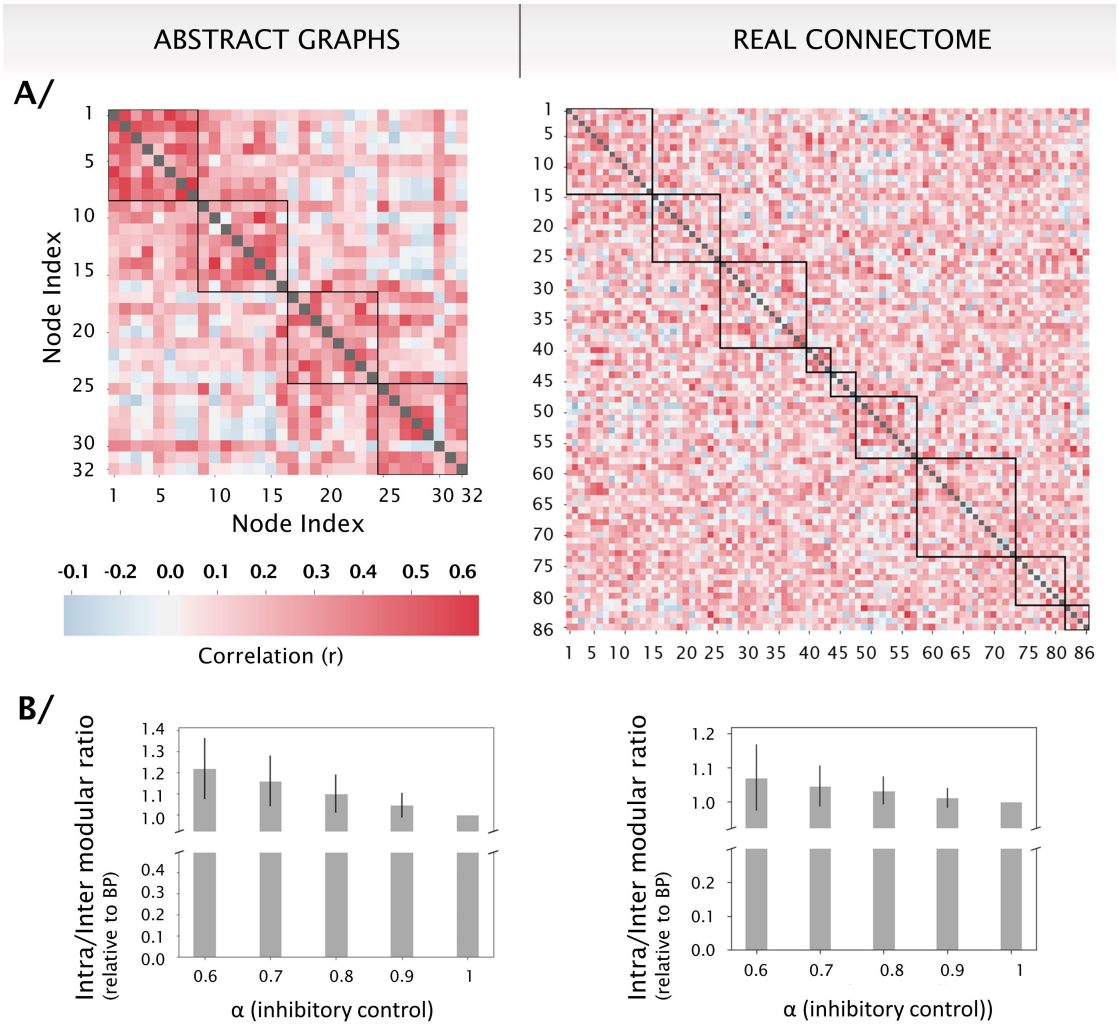

FIGURE 2.12: **Functional connectivity under circular inference**. The results for randomly generated small-world graphs are presented in the left panels, and those for the realistic connectome network are presented in the right panels. **(A)** Functional connectivity matrix network as measured by the activation function applied over the beliefs in the network. Regarding the real connectome, the modules are presented in the following order: auditory, sensorimotor, visual, dorsal attention, salience, frontoparietal, default-mode, subcortical, and finally, nodes not attributed to a specific module. **(B)** We explored the ratio between the number of intramodular connections and the number of intermodular connections in the functional network. BP was chosen as the reference to explore the impact of varying the degree of circularity. The ratio significantly increases when we decrease inhibitory control, rendering the network more modular. B left corresponds to A left (abstract graph), and B right corresponds to A right (real connectome).

to a lesser extent, the participation coefficient and the within-community strength correlated with overactivation (Fig. 2.18 - see Supplementary Material for a formal definition of these three graph metrics).

Finally, the results showing functional connectivity impairments in the abstract graphs were also observed in the real connectome as circular inference increases the modularity in the functional network (Fig. 2.12 lower panels), i.e., the long-range connections between modules were more strongly affected than the short-range connections within a module.

2.3.7 Discussion and perspectives

In this paper, we explored how disruptions of E/I homeostasis in a graph model of the brain can change the dynamics of the network and modify information propagation, eventually leading to psychotic-like symptoms. We developed a whole-brain computational model relating the probabilities encoded by a population to their neural activity using efficient coding principles. This normative approach allowed us to simulate the network dynamics and reproduce some results from the connectomics literature concerning psychosis, specifically the generation of overconfidence and overactivations (centered on hubs) and the inability to maintain an efficient modular small-world architecture (i.e., the increase in modularity in the functional connectome).

Breakdowns in the E/I balance at the microcircuit level are considered major alterations in neurodevelopmental disorders (Foss-Feig et al., 2017; Sohal and Rubenstein, 2019). Notably, an increased E/I ratio was not only proposed to drive psychotic features in schizophrenia but could also be involved in more acute disorders such as anti-NMDAR encephalitis (Parenti et al., 2016). The CI framework, which is based on impaired inhibitory control, appears particularly useful for modeling psychosis across diagnosis categories and has already received some behavioral support. For instance, the overconfidence due to CI is compatible with prior theoretical results (Jardri and Denève, 2013a) and the model fitting of the Jumping-to-conclusions reasoning bias in schizophrenia patients (Jardri et al., 2017), which is usually also correlated with delusional severity (Dudley et al., 2016; Glöckner and Moritz, 2008; Moritz and Woodward, 2005).

The present study extends this literature by showing that CI applied in a brain-like network can generate a nonuniform distribution of aberrantly strong beliefs. More specifically, we showed that belief saturation (excessively high levels of confidence) observed under CI in network hubs prevents the hubs from properly transferring information, which is expressed as intense overactivation. Previous reports have suggested that these hubs play a pivotal role in psychiatric disorders in general (Crossley et al., 2014; Fornito et al., 2015) and schizophrenia in particular (Crossley et al., 2016; Van Den Heuvel et al., 2013). For instance, based on fMRI symptom-capture studies, it is known that patients with psychosis exhibit specific patterns of hyperactivation during hallucinatory experiences (Ćurčić-Blake et al., 2017; Jardri et al., 2011; Sommer et al., 2008). These signal changes are localized in not only essential hubs that constitute a part of the speech-related network when hallucinations occur in the verbal domain (mainly in the inferior frontal gyrus and the temporoparietal junction) but also amodal epicenters involved in contextual memory, such as the hippocampal complex.

Thus, specific dysfunctions in connector hubs appear compatible with the clinical richness and cross-domain impairments of schizophrenia. This nonuniform distribution of beliefs / activations may account for the apparently unrelated observations that psychosis, on the one hand, can generate unshakable beliefs but, on the other hand, may result in impaired information processing.

As a consequence of such localized impairments in hubs, we observe a shift from global to local connectivity (increased short- relative to long-range functional connectivity) and widely distributed miscommunication in the network, which also appears compatible with previous

reports investigating schizophrenia (Li et al., 2017; Xiang et al., 2019; Zalesky et al., 2011). These changes in the brain network topology were previously found to be linked with psychotic symptoms. For instance, functional hyperconnectivity was observed between different parts of the language network (module AUD in Fig. 2.13) in patients suffering from auditory-verbal hallucinations, while hypoconnectivity was found with other distant brain areas (Shinn et al., 2013). Interestingly, these functional dysconnectivity patterns were found to be state-dependent and modulated by antipsychotic medication (Hadley et al., 2016).

The modular topology is known to physiologically vary across an individual's lifespan and notably optimize and correlate with cognitive efficiency during adolescence (Baum et al., 2017). Proper functioning of the brain (i.e., effective segregation and integration of information processing) depends on these short- and long-range connections. Interestingly, synaptic elimination and late E/I balance adjustments contribute to adolescent brain maturation (Selemon, 2013), which is considered to be a critical developmental period for schizophrenia onset (Paus et al., 2008; Rolls and Deco, 2011). Precisely, the transition to psychosis has been shown to be associated with a preferential reduction in long-range connections in patients compared with nonclinical relatives, who shared a genetic vulnerability to schizophrenia but did not develop the disorder (Guo et al., 2014).

It is usually well accepted that global changes in the structural connectivity of the brain represent a pathological hallmark of several neuropsychiatric disorders (Lord et al., 2017). However, our results also suggest that the inability to properly integrate information in different brain areas could be partially due to pure impaired dynamics (i.e., even without modifications of the structural connections), representing a particularly interesting process accounting for acute psychotic manifestations as such manifestations can be observed beyond the schizophrenia spectrum. These predictions of the CI model (in which we alter the inference mechanism but not the anatomical graph) are notably compatible with data from animals exposed to ketamine (an NMDA antagonist - (Voss et al., 2012)) or E/I changes following chemogenetic manipulation (Markicevic et al., 2020) in which a reduction in long-range connectivity was observed. These predictions are also consistent with recent brain imaging findings in patients suffering from anti-NMDAR encephalitis who exhibit multiple focal increases in neural activity (Miao et al., 2020) and a significant decrease in the strength of long-range connections (Peer et al., 2017), even though the structure of the anatomical graph is preserved. Of course, we cannot exclude the possibility that persistent functional changes may lead to plastic structural impairments in the long run if not properly fixed, which could be a nice development of the model.

We would like to acknowledge the preliminary nature of the present work. We intended to provide a proof of concept rather than a detailed framework of spontaneous brain dynamics based on the connectome as we use random connections, random inputs, etc. For instance, spontaneous activity is triggered in the graphs by using random inputs injected in all nodes. In reality, even spontaneous activity is likely to have anatomically and spatially more limited sources (Uddin, 2020). Better identification of these sources in the future might explain more specific patterns in brain activity, such as those in the default-mode network (Fox et al., 2005), and improve the performance of the model in predicting which brain areas should be overactivated. Furthermore, more detailed information could be obtained by using the weights of the connectome. Finally, the current model is restricted to binary variables; subsequent work could adapt the circular inference framework to continuous variables (see Supplementary Material).

In addition, several issues remain unanswered and may benefit from further clarification. Among these possible tracks, we should mention the effect of implementing different types of inference loops in the graph (e.g., ascending/descending), more precise exploration of the neural hierarchy, the trigger of specific subnetworks (e.g., thalamocortical loops or the hippocampal-prefrontal pathway) or even fitting fMRI data from patients with various psychotic symptoms,

e.g., suffering from schizophrenia or anti-NMDAR encephalitis.

Nevertheless, the current findings suggest that the same parametric model (CI) could fit behavioral (Jardri et al., 2017) and neural (this study) data and, thus, pave the way for transdiagnostically linking neural signals with psychosis.

2.3.8 Supplementary material

2.3.8.1 Belief Propagation and Circular Belief Propagation

The Belief Propagation algorithm Belief Propagation, also known as sum-product message passing, is a local message-passing algorithm that performs inference in graphical models (Bishop, 2006), meaning it computes the marginal probabilities of all nodes in the network based on external messages (sensory inputs, prior knowledge). It does so by sending messages across the network. In a factor graph, we distinguish messages sent from variable node (i, j, etc.) to factor node (I, J, etc.) from messages sent from factor node to variable node, whose equations updates are respectively (see for instance Jardri and Denève (2013a)):

$$\mu_{j\to I}(x_j) = \prod_{J\in\mathcal{N}(x_j)\setminus I} \mu_{J\to j}(x_j) \tag{2.9}$$

$$\mu_{I\to i}(x_i) = \sum_{x_{\mathcal{N}(I)\setminus i}} f_I(x_{N(I)}) \prod_{j\in\mathcal{N}(I)\setminus i} \mu_{j\to I}(x_j) \tag{2.10}$$

After convergence of the algorithm, beliefs (marginal probability estimates) are computed as follows:

$$b_i(x_i) = \frac{1}{Z} \prod_{I\in\mathcal{N}(x_i)} \mu_{I\to i}(x_i) \tag{2.11}$$

It has been shown that in the case of binary variables, pairwise factors and with specific factors on edges on the graph $f_{ij}(x_i, x_j)$ if $x_i = x_j$ ($= 0$ or 1) and $1 - w_{ij}$ otherwise, then BP takes a simple form in the log-domain which consists of the following updates equations for the messages and the log-odds (Jardri and Denève, 2013a; Mooij and Kappen, 2005):

$$M_{i\to j}^{t+1} = F_{ij}(L_i^t - M_{j\to i}^t)$$

$$L_i^{t+1} = \sum_j M_{j\to i}^{t+1} + M_{\text{ext}\to i}^{t+1}$$

where
$M_{i\to j}^t$ is the message from node i to node j at iteration t (transmits probabilistic information from i to j - here we consider identical transmission delays for all connections, corresponding to one iteration in the simulation).
$M_{\text{ext}\to i}^t$ is the external input to node i (Gaussian process, see paragraph A).
L_i^n is the *log-odds* of node i (log-odds of the probability that x_i is equal to 1 - also called *log-likelihood ratio*) at iteration n:

$$L_i = \log(b_i/(1 - b_i))$$

where b_i is the belief of node i, i.e. the probability that the binary variable x_i (represented by node i) is 1:

$$b_i = p(X_i = 1) = \sigma(L_i)$$

Finally, F_{ij} is a sigmoidal function with parameter w_{ij}:

$$F_{ij}(x) = \log\Big(\frac{w_{ij}e^x + (1-w_{ij}}{(1-w_{ij})e^x + w_{ij}}\Big) = 2\phi^{-1}\Big((2w_{ij}-1)\phi(\frac{x}{2})\Big)$$

where ϕ is the hyperbolic tanget function (tanh), $w_{ij} = w_{ji}$ is the "synaptic weight" between nodes i and j. It can be seen as a conditional probability and takes values between 0 and 1. For unrelated variables x_i and x_j, the edge between nodes i and j will be absent and $w_{ij} = 0.5$. $w_{ij} > 0.5$ corresponds to positive interaction and $w_{ij} < 0.5$ to a negative interaction between variables x_i and x_j. In its linear region (x small), f_{ij} has slope $2w_{ij} - 1 = 2(w_{ij} - 0.5)$, which supports the positive interaction / negative interaction view depending on the position of w_{ij} compared to 0.5.

The Circular Inference model Circular Inference (CI) is a model of psychosis based on the so-called Circular Belief Propagation algorithm. Circular BP is defined as a specific modification of the update equation of the Belief Propagation algorithm as a result of impaired inhibitory control (E-I imbalance). More specifically, messages from variable nodes to factor nodes are modified under CI (this is true for any graph, pairwise or not, and any type of variable, binary or not) as follows:

$$\mu_{j\to I}(x_j) = (\mu_{I\to j}(x_j))^{1-\alpha} \prod_{J\in\mathcal{N}(x_j)\setminus I} \mu_{J\to j}(x_j) \tag{2.12}$$

which, in the case of binary variables and pairwise factors, leads to the following expression of message $M_{i\to j}$ from factor node f_{ij} to variable node x_j:

$$M_{i\to j}^{t+1} = F_{ij}(L_i^t - \alpha M_{j\to i}^t) \tag{2.13}$$

where parameter α (which takes values in $[0, 1]$) represents the level of inhibitory control in the network. $\alpha = 1$ indicates normality (belief propagation) and lowering α increases the amount of circularity/reverberation in the network.

The original model distinguishes between the impaired feedforward inhibitory control and impaired feedback inhibitory control, so that if the feedforward direction of the edge (i, j) is defined as the direction from i to j, then: a) Feedforward messages ($i \to j$) are controlled by parameter α_d (level of inhibition for the descending loops in the network). b) Feedback messages ($j \to i$) are controlled by parameter α_c (level of inhibition for the climbing loops in the network). In our simulations, we took $\alpha_c = \alpha_d \equiv \alpha$, meaning that feedforward and feedback connections were identical. Values of α were taken ranging from 0.6 to 1, where $\alpha = 1$ indicates normal belief propagation.

Note that in the special case where $\alpha = 0$, circular inference takes a simple form:

$$\mu_{j\to I}(x_j) = \prod_{J\in\mathcal{N}(x_j)} \mu_{J\to j}(x_j)$$

and in the pairwise case:

$$M_{i\to j}^{t+1} = F_{ij}(L_i^t)$$

Several message-passing algorithms on factor graphs such as variational message passing take the form as above (full product, $\alpha = 0$), which could lead to build CI-like message-passing schemes with parameter α by using the partial product from above (CI). (or even BP-like message-passing schemes by taking $\alpha = 1$).

Extension to continuous variables In the article, we only consider probability distributions of binary variables. However, the theory could be extended to continuous variables. Indeed, one can write the circular inference model in the general case (see first equation in paragraph "The circular inference model" above) for any graph (pairwise or not) and any type of variables. Taking pairwise graphs as above, BP (and CI by introducing α) can be written in the case of continuous variables very similarly to BP in the binary case (see also Equations 10-12 of the supplementary material of Jardri and Denève (2013a)), where the update equations for the messages and the log-odds become:

$$M_{i\to j}^{t+1} = F_{ij}(L_i^t(X_i) - M_{j\to i}^t(X_i))$$
$$L_i^{t+1}(X_i) = \sum_j M_{j\to i}^{t+1}(X_i) + M_{\text{ext}\to i}^{t+1}(X_i)$$

and

$$L_i(X_i) = \log(p(X_i)) + \log(Z)$$

where Z is the normalization constant of BP. Finally, F_{ij} is a function of function (it applies to $g(X_i)$) which, among other things, performs an integration over X_i and returns a function of X_j):

$$F_{ij}(g(x_i)) = \log\Big(\int_{x_i} f_{ij}(x_i, x_j)\exp(g(x_i))dx_i\Big)$$

An extension of the present paper to continuous variables would be that the neuronal populations encode $\log(p(X))$ (the log of the whole probability distribution of variable X), as opposed to $\log(p(X=1)/p(X=0))$ in the binary case. In this case, the external input to the network would no longer be a scalar ($M_{\text{ext}\to i}(X_i=1) - M_{\text{ext}\to i}(X_i=0)$) but a function (that is, $M_{\text{ext}\to i}(X_i)$).

2.3.8.2 Simulating network activity

The belief propagation (BP) algorithm was executed over 1000 iterations, as well as its impaired version circular belief propagation, also called circular inference (CI). To do so, we first design the weighted graph and the external stimulus, run the algorithm, and keep for the analysis the non-pathological simulations.

Building the graph structure We used two types of graphs: "abstract graphs" (modular small-world graphs of small size) and "realistic connectomes" (brain-based graphs of bigger size) for sanity check. To randomly generate modular small-world networks, we used the function makeevenCIJ from Python's Brain Connectivity Toolbox (Muldoon et al., 2016). Each generated graph had 32 nodes with 4 modules of size 8 and 425 oriented edges. The networks were modified to remove unidirectional connections, after which additional 20 edges were removed at random to create some sparsity within the modules. The final graphs had between 220 and 250 oriented edges (to be divided by two for unoriented edges) and a density between 0.21 and 0.25.

In addition to the small-world graphs ("abstract graphs"), we also implemented a more realistic brain-like graph based on a preprocessed population template derived from data acquired by the Human Connectome Project (Yeh et al., 2018). Nodes were defined using 86 parcels (see Supplementary Table) taken from the AAL2 atlas (Rolls et al., 2015). Notably, we removed from the 120 original nodes, the cerebellum (18), vermis (8), and the orbitofrontal cortex (8 regions): indeed, these regions were significantly different from the rest of the graph, with very few connections between its containing nodes (orbitofrontal cortex) or on the contrary with nodes outside the anatomical area (cerebellum and vermis). We then estimated connection (edge) strengths between nodes of the connectome by using the magnitude of the along-track diffusion properties (reconstructed group-averaged fiber tracts from the HCP-842 tractography atlas - Yeh et al. (2018);

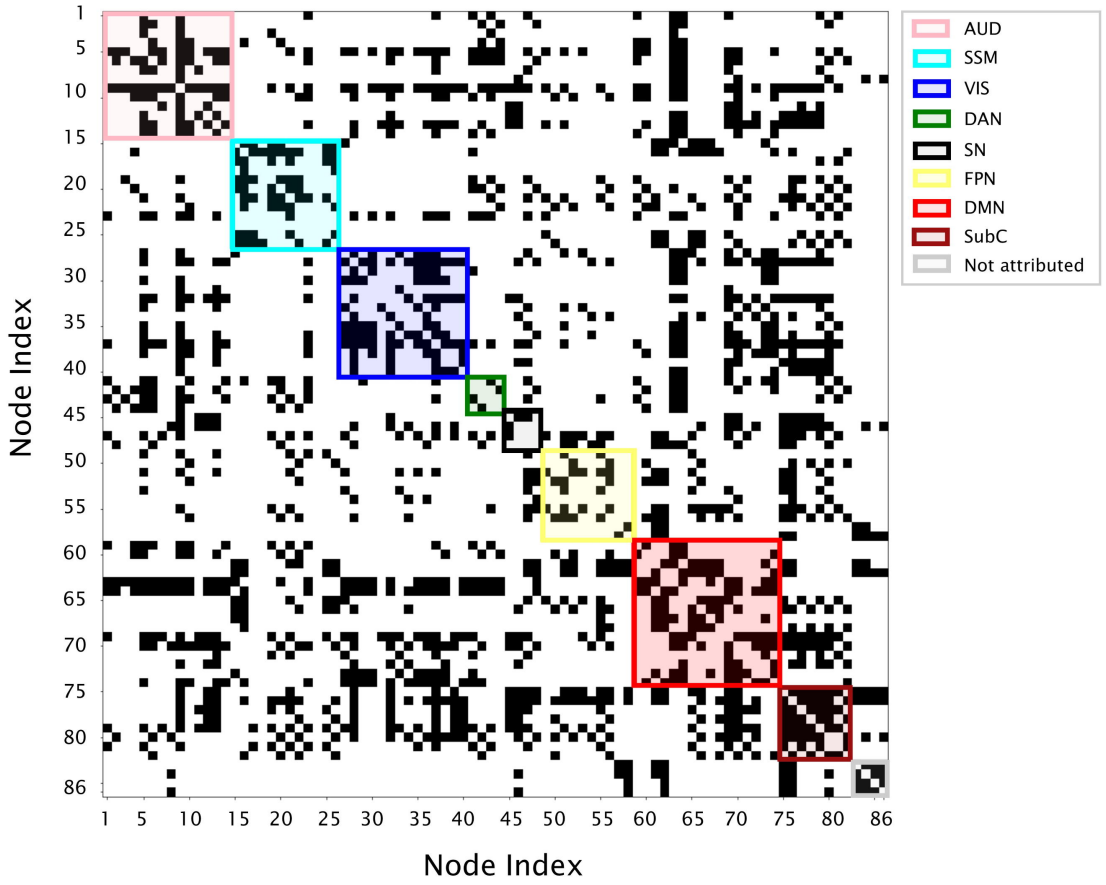

FIGURE 2.13: **Modules in realistic brain-based graphs**. AUD: auditory; SSM: sensorimotor; VIS: visual; DAN: dorsal attention; SN: salience; FPN: frontoparietal; DMN: default-mode; SubC: subcortical. A full list of the parcels' names (1 - 86) is available in the Supplementary Table.

accessible at http://brain.labsolver.org/diffusion-mri-templates/tractography). We applied a threshold to the probability of having a track of $p<0.0007$ to obtain a small-world graph and finally, removed unconnected nodes (here 1 node: Rolandic_Operc_R). The resulting graph has a density of 0.23. See the resulting connectivity matrix in Fig 2.13.

Graph weighting In both cases (abstract graphs and realistic connectome), each edge was given a symmetrical synaptic weight w, where w was selected randomly between two possible values which were symmetrical w.r.t. 0.5 in order to have balanced weights in the graph and thus avoid frustration or bistability (see section "Dynamical regime" below). Connection strengths can be seen as conditional dependencies between these variables. Please note that the graph weighting procedure is random. For this reason, results presented Figs 2.10, 2.11, and 2.12b were obtained by averaging the observations from approximately 50 graphs for each level of inhibitory control α.

Stimulation of the network The network was stimulated by external inputs $M_{\text{ext}\to i}(t)$. Nodes were stimulated using continuous but rapidly varying signals (Gaussian process with a RBF kernel) with mean $\mu = 0$ and variance $\sigma^2 = 30$ (Fig 2.9b). Interestingly, having non-zero means does not change the results presented in the article. Stimulating all nodes is crucial here - stimulating only some portion of nodes would introduce distortions in the activation pattern that have not much to do with the structure of the network but instead have to do with the choice of activated area(s).

Dynamical regime We wanted to avoid frustration (pathological oscillation of the beliefs) in the network, as well as bistable dynamics (where switches between the two stable states are triggered by the noisy external input). Both frustration and bistability are more probable to occur for strong interactions. For this reason, interactions between 2 connected nodes were chosen half positive (w>0.5) and half negative (w<0.5), and low enough (w close enough to 0.5). This, in addition to control for bistability and frustration, allowed to generate different graphs even in the case of the realistic connectome where unoriented edges were determined once for all. In practice, for the randomly generated modular small-world graphs with 32 nodes, connection strengths were selected at random for each edge with value w = 0.65 or 0.35. In the realistic graph, the higher number of nodes and connections by node was compensated by using lower interactions, and w was chosen randomly between 0.57 and 0.43. We excluded the graphs which (still) caused frustration or bistability with these parameters.

Interestingly, the presence of frustration depends on the technical details of BP. More precisely, these pathological oscillations can be fixed by considering belief propagation with momentum instead of normal belief propagation (Murphy et al., 1999), i.e. by replacing the messages sent at time t with a weighted average of messages at time t and t-1. This is the reason why we do not consider frustrated networks for the analysis. Besides, we are modeling psychosis, for which overactivations are relatively mild and not synchronized over areas. For example, per-hallucinatory signal changes correspond to sustained overactivations (clearly different from spikes of activity). Whether these frustrated states could be considered as an epileptic state (large-scale synchronous neuronal discharges) is not the subject of the present article.

2.3.8.3 Activity, overactivity, overconfidence

Neural activity a function of belief Each node in the network encodes a binary variable and tries to represent the belief $b_i = p(X_i = 1)$. To simulate neural activity based on the nodes' beliefs, we assumed that each (belief b_i) as follows: activity was high when the belief was close to 1 or 0, corresponding to a node i corresponded to a brain parcel, and activity in that parcel (a_i) was related to probability high level of certainty for the state of the corresponding variable. Reversely, activity was low when the belief was close to 0.5, corresponding to complete uncertainty.

$$a_i = |L_i| = \left| \log \left(\frac{b_i}{1 - b_i} \right) \right| \tag{2.14}$$

This choice was motivated by previous work on the correspondence between population activity and probability, efficient coding principles and probabilistic population coding (Boerlin et al. (2013); Denève and Machens (2016); Pouget et al. (2013) - see a full demonstration below in paragraph 4C "Demonstration: neural activity as function of belief"):

$$a_i = \left| \dot{L}_i + kL_i \right| \tag{2.15}$$

where we took $k = +\inf$ for simplicity, ignoring the derivative term. Thus, the activity is a weighted sum between the surprise (the derivative of the log-odds of the belief) and the belief itself.

However, as stated in the main text, the two quantities strongly correlated in our simulations and both predicted essentially the same effects on functional connectivity. In other words, changes in k do not modify qualitatively the results of the article.

Finally, the activation was averaged over 10 iterations for it to be similar to the real fMRI BOLD signal.

Overactivation and overconfidence The overactivation due to circular inference is a measure defined for each node and over the entire simulation. It is computed by comparing the temporal sum of the node activation a_i under CI ($\alpha < 1$) with the same under BP ($\alpha = 1$); the overactivation of node i is defined as follows:

$$\text{overactivation}_i(\%) = \frac{\sum_{t=1}^{T} a_i^t(CI) - \sum_{t=1}^{T} a_i^t(BP)}{\sum_{t=1}^{T} a_i^t(BP)} \times 100 \tag{2.16}$$

Similarly, the *overconfidence* of a node i is defined as follows:

$$\text{overconfidence}_i(\%) = \frac{\sum_{t=1}^{T} \left|L_i^t(CI)\right| - \sum_{t=1}^{T} \left|L_i^t(BP)\right|}{\sum_{t=1}^{T} \left|L_i^t(BP)\right|} \times 100 \tag{2.17}$$

Please note that with our choice to take $k = +\text{inf}$ in the simulations (see Paragraph "Neural activity as function of belief"), overactivation and overconfidence are the same.

Demonstration: neural activity as function of belief Here we demonstrate that the activation of a node (population of neurons) is linked to the log-odds L of the variable associated with the node, according to:

$$a = \left|\dot{L} + kL\right| \tag{2.18}$$

(see also Boerlin et al. (2013)).

We insist on the underlying hypotheses. First, the encoded variables are binary. Second, the neural activity observed comes from the neurons encoding the log-odds of the estimated belief: $L = \log\left(\frac{p(X=1)}{1-p(X=1)}\right)$ (scalar value).

Assumption 1: Population coding. The estimate of the log-odds, $\hat{L}$, is a weighted, leaky integration of the spike trains of the neurons which belong to the population ($\boldsymbol{o}(t)$ is a vector):

$$\dot{\hat{L}} = -k\hat{L} + \boldsymbol{\Gamma}^T \boldsymbol{o}(t) \tag{2.19}$$

Importantly, some of the decoding weights Γ_i are negative, which allows the population to encode negative log-odds. Absence of spiking means that $p(X = 1) = 0.5$.

Assumption 2: Efficiency principle. The network minimizes the distance between L and $\hat{L}$ by minimizing the cost function (note we take ν from Boerlin et al. (2013), to be zero):

$$E = (L - \hat{L})^2 + \mu \|r\|^2 \tag{2.20}$$

(trade-off between minimizing the distance between L and $\hat{L}$ and minimizing the firing rate of the network)

It comes from assumption 1 that if we define the firing rate $\boldsymbol{r}(t)$ (vector) as $\dot{\boldsymbol{r}} = -k\boldsymbol{r} + k\boldsymbol{o}(t)$), then:

$$\hat{L} = \frac{\Gamma}{k} r \equiv D^T r \tag{2.21}$$

Using now assumption 2, we obtain: $E = (L - D^T r)^2 + \mu \|r\|^2$. The minimization of the energy gives:

$$\frac{dE}{dr} = 0 = 2[D(L - D^T r) + \mu r \tag{2.22}$$

which implies

$$r = (DD^T + \mu I)^{-1} DL \equiv FL \tag{2.23}$$

Defining the "instantaneous fMRI activation" as $a = M^T o(t)$ (where M is a vector with positive coefficients though not necessarily identical, e.g., some neurons are located in different columns thus don't have the same weight on the fMRI activation),

$$y = \frac{M^T}{k}(\dot{r} + kr) = \frac{M^T F}{k}(\dot{L} + kL) \tag{2.24}$$

It comes: $a \propto \left|\dot{L} + kL\right|$ (as by construction, $o(t) \leq 0$, implying $\dot{r} + kr \leq 0$). Finally, to be closer to the real measured fMRI activation, a should be averaged/smoothed over time.

2.3.8.4 Definition of graph modules

The structural modules represented in Fig 2.9b, 2.12a, 2.12b, 2.13 were determined using the function community_louvain from Python's Brain Connectivity Toolbox in the case of abstract graphs. The 4 modules of size 8 each were recovered. In the case of realistic graphs (AAL2 atlas), we used a canonical division of nodes in a priori communities (or modules, see Fig 2.13) taken from Bertolero et al. (2018), obtaining 9 modules for the 86 nodes of the network. Modules are used to compute the participation coefficient, the within-community strength, and the intra-inter modular ratio (see below for the definition of these 3 node metrics).

2.3.8.5 Functional connectivity analysis

Computing the correlations The functional correlations (Pearson r correlation) were computed from the activation function. To correct for multiple comparisons, we performed an FDR correction and discarded the connections without a significant correlation.

To select significant connections for the functional graph, we selected all edges whose absolute correlation greater than a given threshold, which corresponded to a p-value higher than 10-9. This threshold was chosen to have a density close to 0.2 as in the structural graph.

Intra-inter modular ratio The intra-inter modular ratio of the functional connectome is defined as the ratio between the number of significant functional connections within a structural module (intramodular) and the number of significant functional connections between nodes from different structural modules (intermodular):

$$\frac{\#\text{intramodular connections}}{\#\text{intermodular connections}} \tag{2.25}$$

We explored the intra-inter modular ratio (Fig 2.12b) for the abstract and brain-based graphs and compared them to BP ($\alpha = 100\% = 1$) using paired t-test with Bonferroni corrections. Decreasing inhibitory control significantly increased the intra-inter modular ratio and thus the graph modularity. This was true for the randomly-generated graphs (t(0.9 vs 1) = 6.04 , p = 4.24 e-07 ; t(0.8 vs 1) = 8.54 , p = 2.39 e-11 ; t(0.7 vs 1) = 1.04 e+01 , p = 1.91 e-14; t(0.6 vs 1) = 1.17 e+01, p = 1.81 e-16) as well as for the real connectome ((0.9 vs 1) = 2.38 , p = 8.97 e-02 ; t(0.8 vs 1) = 4.68 , p= 1.50 e-04 ; t(0.7 vs 1) = 4.58 , p = 2.03 e-04 ; t(0.6 vs 1) = 4.39 , p = 3.61 e-04).

2.3.8.6 Graph metrics

We used 3 different node metrics: degree centrality, participation coefficient, and within-community strength.

Degree centrality The degree centrality of a node i is the fraction of nodes it is connected to. It is equal to the degree (the number of neighbors of the node) divided by the number of nodes of the graph.

Participation coefficient The participation coefficient of a node measures the diversity of intermodular connections (connections with nodes belonging to other modules) of the node in the graph.

$$PC_i = 1 - \sum_{m=1}^{N} \left(\frac{d_{im}}{d_i}\right)^2 \tag{2.26}$$

where d_i is the degree of node i (total number of connections) and d_{im} is the degree of node i inside module m (number of connections between i and nodes of module m). N is the number of modules inside the graph. The participation coefficient (PC_i) is thus a measure of how evenly distributed a node's edges are across modules. The participation coefficient of a node is maximal if the node has the same number of neighbors in each module in the network. On the contrary, the participation coefficient of a node is minimal (equal to zero) if all neighbors of the node belong to the same module.

Within-community strength The within-community strength of a node is a measure of the locality of the node in the graph. It is computed by considering the subgraph corresponding to the module (community) of the node.

$$z_i = \frac{k_i - \text{mean}(\{k_j; j \in m_i\})}{\text{std}(\{k_j; j \in m_i\})} \tag{2.27}$$

where k_i is the number of connections node i has with nodes inside its module m_i. k_i is centered and normalized with respect to the$k_j$of all nodes j belonging to the module m_i (including node i). Within-community strength (z_i) is thus a measure of how much node i is connected to other nodes inside its module compared to other nodes of this module.

2.3.8.7 Definition of hubs

Our networks contain different types of nodes, which we divide into 3 categories: connector hubs, local hubs, and other nodes. Connector hubs are nodes whose connections are diversely distributed across modules (nodes with a *participation coefficient* in the top 20%). Local hubs are nodes that are highly connected within their own module (nodes with a *within-community strength* in the top 20% among the remaining nodes). Other nodes are all the nodes that are not connector hubs nor local hubs. This definition is the same as the one in Bertolero et al. (2018), except that we divide the nodes into 3 categories whereas Bertolero and colleagues have 4 (connector hub or not; local hub or not). Nodes that had both high participation coefficient and high within-community strength were considered, as stated above, as connector hubs. The participation coefficient and the degree centrality strongly correlate in our graphs as it often does.

2.3.8.8 Linear regression analysis

We regressed the amount of overactivation induced by CI versus three graph metrics: the degree centrality, the participation coefficient, and the within-community strength of a node (metrics are defined in the paragraph "Graph Metrics"). Because of the collinearity between the three regressors, we referred to Partial Least Squares Regression. The results of this analysis are presented in Fig 2.14. We can see that overactivations mainly relate to degree centrality and that the value of coefficients increases with the amount of circularity (i.e. while α decreases from 100% to 60%) as overactivation increases with the amount of circularity.

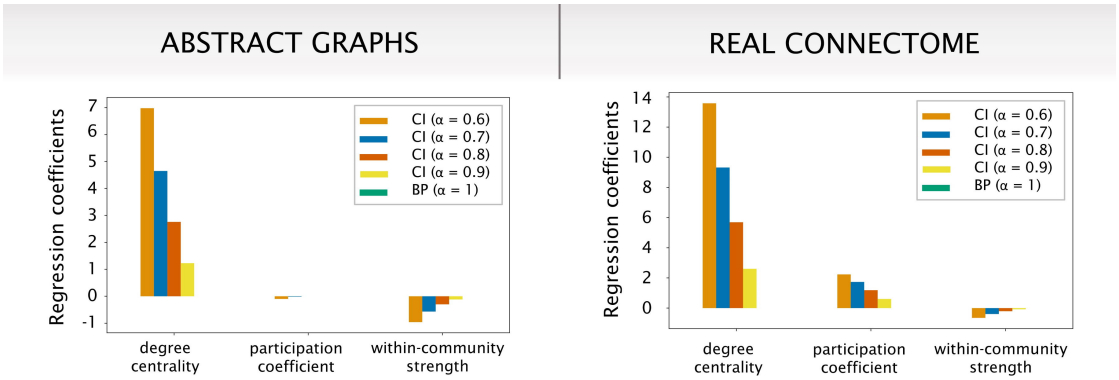

Figure 2.14: **Linear regression analysis** was performed to explain the variation in overconfidence due to circular inference (CI). The normalized regression coefficients of three variables are displayed. The degree centrality of a node explains most of the variation compared to its participation coefficient or within community strength (abstract graphs are on the left; real connectome on the right). Decreasing the level of α (i.e., increasing the level of CI) increases the magnitude of regression coefficients.

2.3.8.9 Supplementary figures

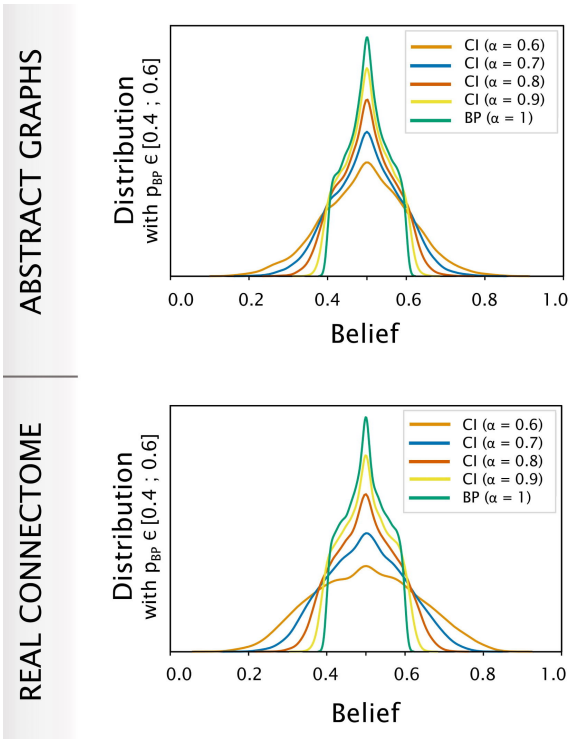

FIGURE 2.15: **Distribution of beliefs** in a single node when varying inhibitory control (α) and bounding the belief (probability estimate) for BP between 0.4 and 0.6, corresponding to relative uncertainty. Increasing the level of CI still causes more extreme beliefs.

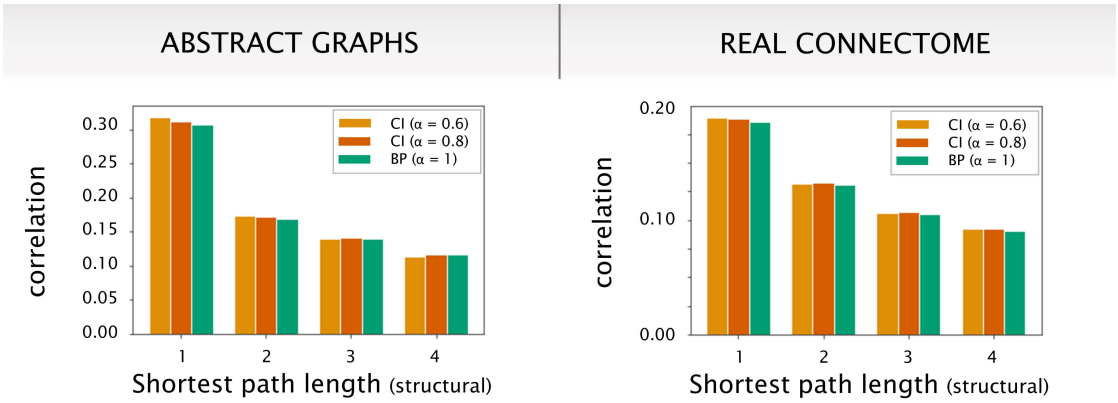

FIGURE 2.16: **Average correlation between nodes as a function of the structural path-length**. Decreasing the level of α (i.e., increasing the level of CI) increases short-range connections relative to long-range connections, compared to belief-propagation ($\alpha = 100\%$).

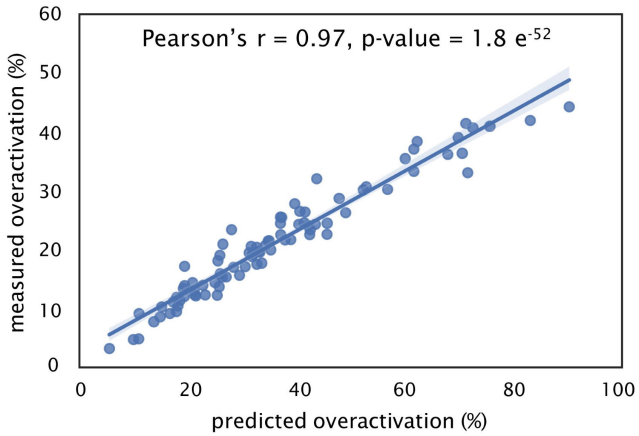

Figure 2.17: **Correlation between predicted and true overactivation**. The predicted overactivation induced by circular inference ($\alpha = 60\%$) in the realistic connectome based on a linear model applied to the randomly generated graphs is strongly correlated with the overactivation observed in the same realistic connectome.

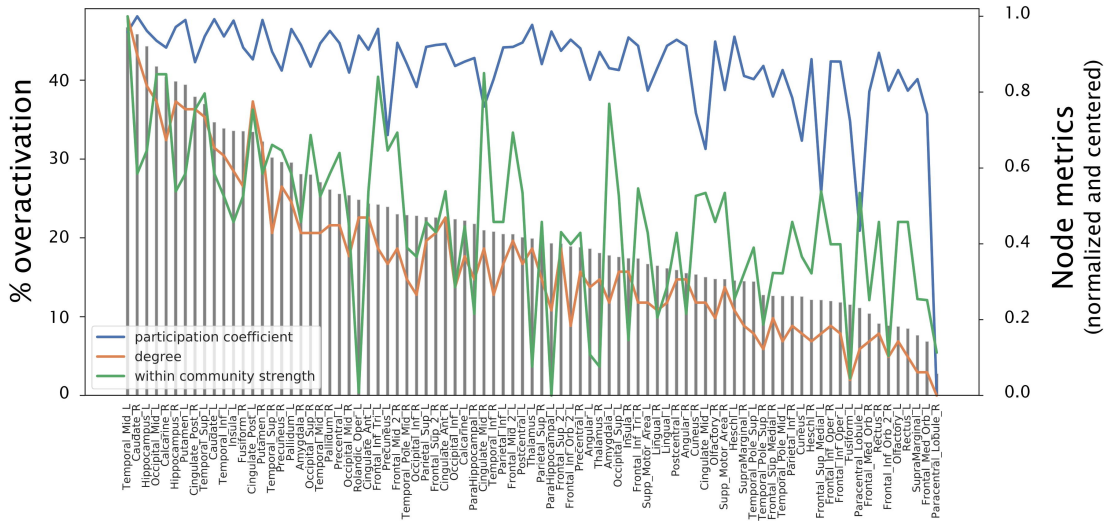

Figure 2.18: **Overactivation relates to nodal properties in the real connectome**. Parcels from the realistic connectome are ranked according to the level of overactivation induced by circular inference ($\alpha = 60\%$). The level of overactivation nicely follows the degree centrality of the presented nodes, but only slightly relates to the within-community strength or the participation coefficient.

2.4 Conclusion

In this chapter, we provided further evidence for Circular BP as a model of suboptimal behavior. We showed that some amounts of circularity were required to account for the phenomenon of bistable perception, by stabilizing percepts thanks to the so-called "descending loops" of the system (reverberation of the prior which corrupts the sensory evidence). Additionally, we showed that the effect of circularity in cyclic graphs is consistent with the specific overactivation patterns and functional connectivity changes observed in schizophrenia.

The Circular Inference model was designed to specifically account for the positive symptoms of schizophrenia. Jardri et al. (2017) later showed that negative symptoms (surprisingly) tighly relate to the model as well. However, the scope of the Circular Inference model goes beyond schizophrenia. As stated previously, the model could be applied to autism, subclinical populations (e.g., conspirationnists), pharcologically-induced psychosis (for instance with the ketamine drug), the NMDA-R encephalitis auto-immune disease, and even the general population, as suggested in the first part of this chapter, as bistable perception does occur for all populations. Let alone this example of bistable perception, which is uncommon in real life and a laboratory-designed task producing high uncertainty, it will be important to understand in future work what the scope of the Circular Inference model is, and more particularly, whether it can also account for other mental disorders than schizophrenia (see a discussion about autism in Chrysaitis et al. (2021)).

Chapter 3

Circular Belief Propagation as model of optimal behavior

Summary of Chapter 3

For the human brain, everyday life requires performing complex probabilistic inference tasks, many of which are mathematically intractable. A crucial question is how the brain can carry out such approximate inference in an efficient manner. A possibility is that the structure of neural networks mirrors probabilistic graphs and that the network activity corresponds to the Belief Propagation algorithm in the underlying graph. However, the poor performance of BP in cyclic graphs casts doubt on this proposition, given the recurrent nature of the brain. In this chapter, we propose Circular BP as model of (nearly) optimal behavior.

First, the Circular BP algorithm is compared to other approximate inference based on BP, in particular Fractional BP and Tree-Reweighted BP, whose update equations resemble the one of Circular BP. This provides a mathematical foundation for Circular BP, and allows for an extension of Circular BP. This extended Circular BP is based on an extension of Fractional BP (Generalized BP) which is linked to a generalization of BP's Bethe Free Energy on factor graphs.

Second, we show that Circular Belief Propagation algorithm with particular parameters performs approximate inference with higher quality than BP, and that these parameters can be learnt using an unsupervised learning rule. This method enables the network to (partly) compensate for the loops of information arising from recurrence. We show that the proposed approach not only improves the quality of probabilistic inferences but also brings better convergence properties to the network.

Finally, we propose a variant of the Circular BP algorithm called "Circular BP with memory", with the idea that the information travelling through cycles should be cancelled at the moment when it gets back to its initial sender. This algorithm extends Circular BP and is hypothesized to perform better approximate inference than Circular BP.

The work included in this Chapter was supervised by S. Denève and R. Jardri. The corresponding manuscript is in preparation.

3.1 Introduction: countering the effect of cycles with Circular BP

Chapter 2 examined how Circular BP could be used to model suboptimal behavior such as the bistable perception phenomenon, and how Circular BP looks compatible with particular disturbances of the functional connectivity in schizophrenia.

All this work relies on the hypothesis that the Belief Propagation algorithm is an accurate model of how the healthy brain performs probabilistic inference. Circular BP would then simply account for the pathology, and Belief Propagation for normal functioning. Indeed, increased circularity in the Circular Belief Propagation algorithm (that is, $\alpha < 1$) brings overconfidence to the network compared to BP, as previously shown in Figure 1.4 for an acyclic network and in Figure 2.10 for a cyclic network.

A crucial note, which motivates this chapter, is that Belief Propagation itself is suboptimal in cyclic graphs, and can even sometimes perform quite poorly. This makes Belief Propagation a questionable candidate for how probabilistic inference is implemented in the healthy brain, given that humans are usually pretty good at inference tasks as seen in introduction chapter of this thesis (Chapter 1).

Double-counting with BP in cyclic graphs The Belief Propagation algorithm is exact in acyclic graphs and used in practice in sparse graphs where it performs very well (see the Introduction chapter). A particular feature of BP in acyclic graphs is that messages going in opposite directions are decorrelated thanks to the removal of redundant information, thus avoiding positive feedback and incorrect computations. In other words, the probabilistic message $m_{i\to j}$ from node i to node j carries different information from $m_{j\to i}$; see Figure 1.1. However, one drawback of Belief Propagation is that it often performs poorly in highly cyclic graphs (Murphy et al., 1999; Weiss, 2000). Cycles are indeed responsible for reverberations of information: messages can pass from node to node and return to the original node, causing the same piece of information to be counted several times. This is known as the "double-counting" problem and is incompatible with highly recurrent brains solving accurately probabilistic tasks. This reverberation of information in BP due to cycles is explained in Figure 3.1. Information sent by node j to i will come back to j and thus i, leading to information being counted multiple times. The correction in Belief Propagation consists of removing the potential reverberation of information on single edges (sender $\to$ target $\to$ sender, that is, $i \to j \to i$), which is enough for acyclic graphs and explains why BP performs exact inference in this case. However, for cyclic graphs, correcting for these loops of length 2 is no longer sufficient.

Interestingly, cycles create what seems to be systematic overconfidence in BP (respectively systematic underconfidence, depending on the node considered); see Figure 3.2. The sign of the bias (overconfidence or underconfidence) can partly be guessed. As a matter of fact, the quantity that determines whether information is amplified or attenuated in a cycle is the product of edge weights over the cycle $\prod_{\text{cycle}} J_{\text{edge}}$. The propagation of messages is indeed determined by function $f_{ij}(x) = \phi^{-1}(\phi(J_{ij})\phi(x))$ which has the same sign as $J_{ij} \times x$. Therefore, a message entering node i travels in the cycle before coming back to the same node i, with a weight whose sign is $\text{sgn}(\prod_{\text{cycle}} J_{\text{edge}})$. Therefore, the belief at node i is overconfident if $\prod_{\text{cycle}} J_{\text{edge}} > 0$, and underconfident otherwise. In fact, function f_{ij} can be linearized with $f_{ij}(x) \approx J_{ij} \times x$. As a consequence, $\prod_{\text{cycle}} J_{\text{edge}}$ not only gives the sign of the amplification, but it also represents the level of amplification of the message by the cycle. Note that this way of predicting the sign of the bias (overconfidence or underconfidence) only works for graphs with a single cycle. With multiple cycles possibly including identical edges, a first approximation would be to sum each contribution of all cycles passing through node i with $\sum_{\text{cycles}} \prod_{\text{cycle}} J_{\text{edge}}$ and consider its sign, but this is sometimes not right because of the non-linearities in the system.

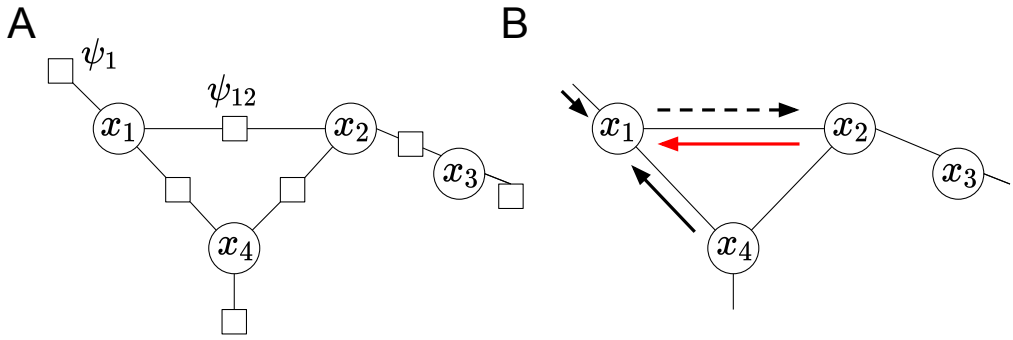

FIGURE 3.1: **Belief Propagation (and Circular Belief Propagation) applied to a cyclic graph**. **(A)** The probability distribution $p(\mathbf{x})$ is represented by a factor graph with pairwise potentials ψ_{ij} and unitary potentials ψ_i. Here $p(\mathbf{x}) = \psi_{12}(x_1, x_2)\psi_{14}(x_1, x_4)\psi_{23}(x_2, x_3)\psi_{24}(x_2, x_4)\psi_1(x_1)\psi_3(x_3)\psi_4(x_4)$. **(B)** Belief Propagation in a cyclic probabilistic graph. The message $m_{1\to2}$ sent by node x_1 to node x_2 (dotted black line) depends, as for acyclic graphs, on three components (see full black lines): the messages received by node x_1 from its neighbors except x_2, the unitary potential ψ_1 at node x_1, and the interaction ψ_{12} between nodes x_1 and x_2. The Belief Propagation is not exact in this case, because messages travel through cycles (for instance, the message sent by node x_1 to node x_2 naturally travels back to node x_1 because of the cycle $x_1 - x_2 - x_4$) and therefore get counted multiple times. Furthermore, $m_{1\to2}$ is correlated to $m_{2\to1}$ as both depend for instance on $m_{3\to2}$ ($m_{1\to2}$ depends on $m_{4\to1}$ which depends on $m_{2\to4}$ which depends on $m_{3\to2}$, and $m_{2\to1}$ depends directly on $m_{3\to2}$). Contrary to BP, Circular BP takes into account a fourth component (full red line) to compute $m_{1\to2}$: the message $m_{2\to1}$, taken with weight $1 - \alpha_{12}$.

As shown previously in Figure 1.4 for acyclic graphs, Circular Belief Propagation also creates overconfidence in all nodes of the graph. More specifically, the choice of the circularity level in Circular BP leads to whether overconfidence (for $\alpha < 1$ in acyclic graphs, which was the initial idea behind explaining circular inferences) or underconfidence (for $\alpha > 1$ in acyclic graphs). What comes from these observations is the idea that in cyclic graphs, the right level of circularity in Circular BP could improve the inference compared to standard BP, by counteracting the effect of cycles.

Improving the quality of inference with Circular BP For instance, if all graph weights are positive, BP is overconfident on all nodes, and "anti-Circular BP" (i.e., Circular BP with $\alpha > 1$) causes some underconfidence, (partly) compensating for the overconfidence created by cycles; see Figure 3.3. Figure 3.4B shows that the estimation of the marginal probability by Circular BP is best around $\alpha = 1.2$ (a second example is also shown in the same plot for a different random graph). In graphs with only positive weights, the optimal uniform α is always > 1, because all cycles contribute to overcounting the same information several times.

In graphs with only negative weights, cycles with length 3 (which are the smallest possible cycles and have more impact in general compared to cycles of higher length) contribute to undercounting information as the product of the weights J_{ij} over such cycles is negative. For this reason, the optimal α is (nearly) always < 1 for such graphs, as shown in Figure 3.4C.

Finally, and most interestingly, for graphs with both positive and negative weights chosen randomly, the optimal level of circularity α in the graph is often close to 1 (where $\boldsymbol{\alpha}$ is as before

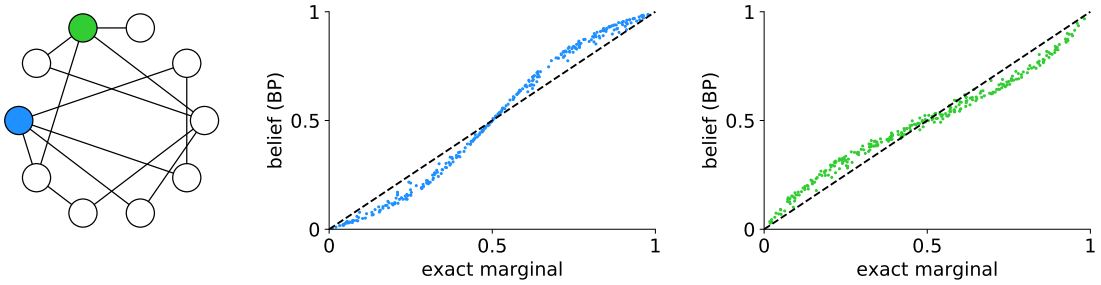

Figure 3.2: **Systematic bias (overconfidence or underconfidence, depending on the node) is observed when applying the Belief Propagation algorithm to a cyclic graph**. The direction of the bias is determined by the local structure around the node. For instance, the blue node belongs to a *positive cycle* (that is, a cycle for which the product of the weights $J_{ij} > 0$) and therefore produces overconfident beliefs due to double counting, and the green node belongs to a *negative cycle* and therefore produces underconfident beliefs.

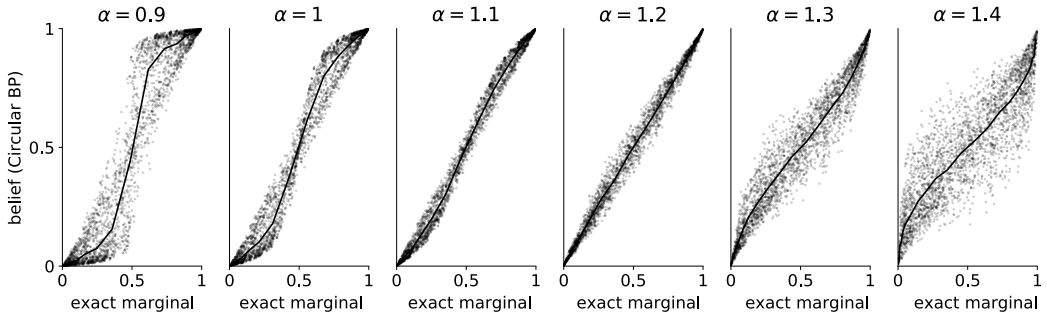

Figure 3.3: **How Circular BP could fight the effect of loops**. In this example, weights of the (cyclic) graph J_{ij} are taken all positive, and Circular BP is applied to the graph with $\boldsymbol{\alpha}$ taken uniformly over the edges. Because of the positive weights, the marginals approximated by BP (case $\alpha = 1$) are overconfident. However, underconfidence brought by (anti-)Circular BP partly compensates for the overconfidence naturally originating from the graph cycles. There is an "optimal" value for α (here $\alpha_{\text{opt}} \approx 1.2$, for which beliefs are not overconfident ($\alpha < \alpha_{\text{opt}}$) nor underconfident ($\alpha > \alpha_{\text{opt}}$). The graph was randomly generated and randomly weighted. 300 examples $\boldsymbol{M}_{\text{ext}}$ were randomly generated. One point represents the approximate marginal under Circular BP versus exact marginal, for each node and each example. The full line represents the average of all points. See Figure 1.4 for acyclic graphs instead.

taken uniformly); see Figure 3.4D. This comes from having both *negative cycles* and *positive cycles*, where the sign of a cycle is by definition the sign of the product of the weights J_{ij} over this cycle. Positive and negative cycles have different effects. Negative cycles may lead to oscillations while running the BP algorithm. Positive cycles may amplify noisy or random initial messages and therefore have incorrect convergence values of the beliefs, which depend on the sign of the initial conditions, and on the network structure rather than on the external input value itself.

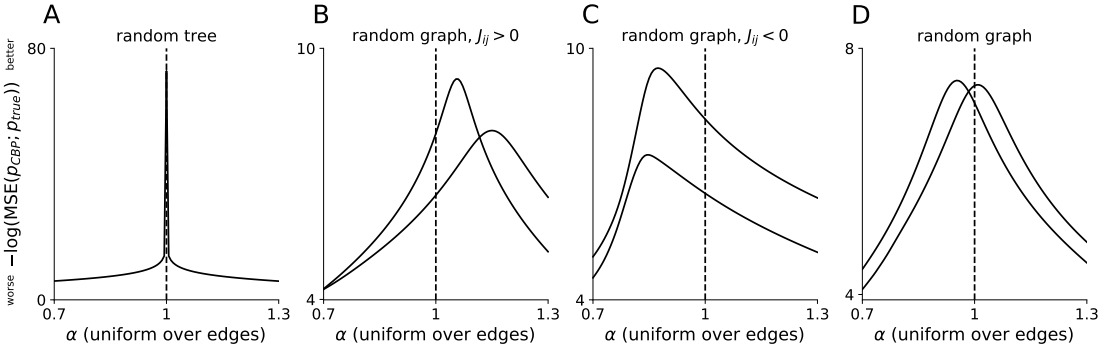

FIGURE 3.4: **How to choose $\boldsymbol{\alpha}$ in Circular BP to fight the effect of loops**. The log-error between the exact probability and the belief obtained from Circular BP are computed depending on parameter $\boldsymbol{\alpha}$ used in Circular BP (here $\boldsymbol{\alpha}$ is taken uniformly over the edges, equal to α). Two random graphs are considered, except for panel A (one random graph). **(A)** In random trees, BP is an exact algorithm, thus the optimal value for α is 1. **(B)** In random graphs with positive weights, BP is overconfident as all cycles are positive, thus the optimal value for α is > 1. The curve peaking around $\alpha = 1.2$ represents the graph of Figure 3.3. **(C)** In random graphs with negative weights and high enough density, BP is underconfident as the smallest cycles (length 3) are always negative, thus the optimal value for α is < 1. **(D)** In random graphs with both positive and negative weights, the existence of both positive and negative cycles makes the optimal α be < 1 or > 1, depending on the random graph.

Non-uniform $\boldsymbol{\alpha}$ can even improve the quality of inference more than for uniform $\boldsymbol{\alpha}$. The optimal amount of *local* circularity α_{ij} in Circular BP (associated to edge (x_i, x_j) in the probabilistic graph) depends on the local structure around the edge. If the edge (x_i, x_j) belongs to a positive cycle then marginals estimated by BP for nodes x_i and x_j are overconfident and α_{ij} should be > 1 to compensate for it. Conversely, if the edge belongs to a negative cycle then marginals estimated by BP for x_i and x_j are underconfident and α_{ij} should be < 1. Of course, in practice edges belong to several cycles of different signs and strengths (product of weights) thus it can be complicated to predict on which side of 1 parameter α_{ij} should be.

Goal of this chapter In this chapter, we use Circular BP in order to counterbalance the amount of "natural" overcounting in cyclic graphs. More specifically, we show that appropriate amounts of *local* circularity α_{ij} in Circular BP (where α_{ij} is associated to edge (x_i, x_j) in the probabilistic graph) can improve the quality of inference. In this case, we talk about *Balanced* Circular BP to describe a state in which overall, information is transmitted accurately, without "explosion" of the system (overamplification, seen as a lack of inhibition with respect to excitation) nor "shutting down" of the system (overdampening, seen as an excess of inhibition). Consequently, the Circular BP algorithm can not only describe disturbed inference and the modelling of psychosis (in acyclic graphs or cyclic ones), but can also model near-optimal inference (in cyclic graphs) and constitutes a particularly interesting approach to account for a large range of normal behavior in non-pathological brains.

Organization of this chapter This chapter is organized as follows. In section 3.2, we provide a theoretical justification to Circular BP (which was initially defined through intuition rather than from normative principles): Circular BP is an approximation to Fractional BP, which was

proposed in Wiegerinck and Heskes (2002) as an approximate inference method extending BP. In section 3.3, we propose a generalization of Circular BP (the "extended Circular BP" algorithm) based on an extension of Fractional BP. After showing that it is possible to guarantee the convergence of the new algorithm in section 3.4, we show through simulations in section 3.5 that it is possible to learn the parameters of extended Circular BP in a supervised manner, so that it performs approximate inference accurately. Finally, in section 3.6, we propose an algorithm, "Circular BP with memory", similar to Circular BP, but which more closely relates to the initial idea of cancelling messages being reverberated through the network cycles.

3.2 Theoretical foundation for Circular BP

The modification of BP defining Circular BP was initially motivated at intuitive point of view only (Denève, 2005, 2008; Jardri and Denève, 2013a); see also section 3.6. Here we provide a theoretical justification to the Circular BP algorithm. We do so by relating Circular BP to Fractional BP, a very similar approximate inference algorithm derived from a parametric approximation of the entropy of the approximating distribution $b(\mathbf{x})$.

3.2.1 Fractional Belief Propagation

3.2.1.1 Definition

Fractional Belief Propagation, originally proposed in Wiegerinck and Heskes (2002), is an approximative inference algorithm which extends the standard Belief Propagation algorithm. We provide here a brief introduction to Fractional BP.

Fractional BP extends Belief Propagation based on a parametric approximation of the entropy of the approximating distribution $b(\mathbf{x})$. This parametric approximation consists of introducing parameter $\boldsymbol{\alpha}$, where α_{ij} is assigned to the undirected edge (i,j):

$$b(\mathbf{x}) \approx \prod_{i,j} \left(\frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)} \right)^{1/\alpha_{ij}} \prod_i b_i(x_i) \tag{3.1}$$

where parameter $\boldsymbol{\alpha}$ would need to be adapted to the true probability distribution $p(\mathbf{x})$ (see learning in section 3.5). On the contrary, BP associates with $\boldsymbol{\alpha} = \mathbf{1}$.

As shown in Appendix A, and similarly to BP as described in section 1.4.2.2, Equation (3.1) is equivalent to making a parametric approximation of the Gibbs free energy: $G \approx G_{\text{approx}}$ where G_{approx} is given below.

$$\begin{aligned} G_{\text{approx}} = &\sum_{(i,j)} \frac{1}{\alpha_{ij}} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log \left(\frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)} \right) - \sum_{(i,j)} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log \Big(\psi_{ij}(x_i,x_j) \Big) \\ &+ \sum_i \sum_{x_i} b_i(x_i) \log \Big(b_i(x_i) \Big) - \sum_i \sum_{x_i} b_i(x_i) \log \Big(\psi_i(x_i) \Big) \end{aligned} \tag{3.2}$$

This parametric approximation generalizes the Bethe approximation using in BP. G_{approx} is indeed a parametric generalization of the Bethe free energy (we recover the Bethe free energy of BP with $\boldsymbol{\alpha} = \mathbf{1}$). We obtain the following modified update equation for the messages (see demonstration in Appendix A):

$$m_{i\to j}^{\text{new}}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i,x_j)^{\alpha_{ij}} \psi_i(x_i) \prod_{k \in \mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}} \right)^{1/\alpha_{ij}} \tag{3.3}$$

and beliefs are computed using:

$$b_i(x_i) \propto \psi_i(x_i) \prod_{k \in \mathcal{N}(i)} m_{k \to i}(x_i) \tag{3.4}$$

Note that Equation (3.3) is not identical to the one given in the original Fractional BP paper (Wiegerinck and Heskes, 2002), but is the same up the amount of damping in the algorithm; see section 3.2.1.2.

In the special case of the probability distribution $p(x)$ over binary variables, the resulting system of equations defining Fractional BP in the log-domain is:

$$\begin{cases} M_{i \to j}^{\text{new}} = g_{ij}\Big(B_i - \alpha_{ij} M_{j \to i}\Big) & \text{(3.5a)} \\ B_i = \sum_{j \in \mathcal{N}(i)} M_{j \to i} + M_{\text{ext} \to i} & \text{(3.5b)} \end{cases}$$

Quantities $\boldsymbol{M}$ and $\boldsymbol{B}$ are given by: $B_i \equiv \frac{1}{2} \log\left(\frac{b_i(x_i=+1)}{b_i(x_i=-1)}\right)$, $M_{i \to j} \equiv \frac{1}{2} \log\left(\frac{m_{i \to j}(x_j=+1)}{m_{i \to j}(x_j=-1)}\right)$, and $M_{\text{ext} \to i} \equiv \frac{1}{2} \log\left(\frac{\psi_i(x_i=+1)}{\psi_i(x_i=-1)}\right)$. Function g_{ij} is a sigmoidal function defined by:

$$g_{ij}(x) = \frac{1}{2\alpha_{ij}} \log\left(\frac{\left(\psi_{ij}^{1,1}\right)^{\alpha_{ij}} e^{2x} + \left(\psi_{ij}^{1,0}\right)^{\alpha_{ij}}}{\left(\psi_{ij}^{0,1}\right)^{\alpha_{ij}} e^{2x} + \left(\psi_{ij}^{0,0}\right)^{\alpha_{ij}}}\right) \tag{3.6}$$

in the general case, and

$$g_{ij}(x) = \frac{1}{\alpha_{ij}} \phi^{-1}\Big(\phi\Big(\alpha_{ij} J_{ij}\Big)\phi(x)\Big) \tag{3.7}$$

for an Ising model, for which $\psi_{ij}(x_i, x_j) \propto \exp(J_{ij} x_i x_j)$. The algorithm eventually computes the approximate marginals (or beliefs), given by $b_i(x_i = \pm 1) = \sigma(\pm 2B_i)$, i.e., $b_i(x_i) \propto \exp(B_i x_i)$.

The fact that α_{ij} appears in three places (twice in Equation (3.7), once in Equation (3.5a)) and thus represent three biological quantities with identical values, makes it highly implausible that such an algorithm is implemented in the brain as such.[1]

Equation (3.5a) means that node i, encoding for variable x_i, sends to j everything it knows (B_i) except a *rescaled* version of what j communicated to i ($M_{j \to i}$). The corrective multiplicative factor α_{ij} helps to compensate, in a linear manner, for the effect of reverberation of information in cyclic graphs. Information is still being propagated, but redundant information brought by cycles is eventually being cancelled by the control units once messages return to the sender. This illustrates the fact that when there are cycles, messages are wrongly estimated and need rescaling. For instance, if all interaction weights J_{ij} are positive, messages are overcounted in BP, therefore, to compensate, control units need to remove more evidence than if there were no cycles: $\alpha_{ij} > 1$ is required in this case.

3.2.1.2 Related algorithms

The Fractional Belief Propagation algorithm is not only closely related to BP but also to other well-known approximate inference algorithms such as Power EP (Minka and Lafferty, 2002; Minka, 2004), α-BP (Liu et al., 2019, 2020), Tree-reweighted BP (Wainwright. et al., 2002, 2003; Wainwright et al., 2005), and Variational message-passing (Winn and Bishop, 2005).

[1] This is contrary to Circular BP, for which α_{ij} does not appear in the definition of the update function, as $f_{ij}(x) = \phi^{-1}(\phi(J_{ij})\phi(x))$.

Equivalence between damped Fractional BP, Power EP, and α-BP Fractional BP as defined by Equations (3.5a) and (3.5b) can be related to several models previously proposed that themselves extend BP: Power EP from Minka and Lafferty (2002); Minka (2004) and α-BP from Liu et al. (2019, 2020). These models are conceptually very similar to each other (see Minka (2005)). In fact, the three algorithms are identical up to the amount of damping, where damping consists of taking partial message update steps (Murphy et al., 1999). Indeed, damped Fractional BP, defined similarly to damped BP (see section 4.2.1) from the undamped Fractional BP message update equation in Equation (3.3), is written:

$$m_{i\to j}^{\text{new}}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\alpha_{ij}} \psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}} \right)^{(1-\epsilon_{i\to j})/\alpha_{ij}} \times m_{i\to j}(x_j)^{\epsilon_{i\to j}} \tag{3.8}$$

becomes for the particular damping value $\epsilon_{i\to j} = 1 - \alpha_{ij}$, :

$$m_{i\to j}^{\text{new}}(x_j) \propto \Big(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\alpha_{ij}} \psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}} \Big) m_{i\to j}(x_j)^{1-\alpha_{ij}} \tag{3.9}$$

which is exactly the message update equation of Fractional BP (see Equation (17) of Wiegerinck and Heskes (2002)) and α-BP (see Equation (17) of Liu et al. (2020)). When it comes to Power EP (Minka, 2004), the damping does not appear from the derivations of the algorithm but Power EP is still presented with the possibility of having (any) damping $\boldsymbol{\epsilon}$ (see Equations (22) and (23) of Minka (2004)). It is stated in the paper that with the particular value of damping $\epsilon_{i\to j} = 1 - \alpha_{ij}$, the algorithm is "convenient computationally and tends to have good convergence properties". Interestingly, with this particular value of damping, it is shown that the Power EP (and equivalently, α-BP and Fractional BP) proposed implements a minimization of the α-divergence (see also the work on α-BP (Liu et al., 2020) as well as Minka (2005)).

Again, the damping values only alters the convergence properties of the algorithm but not the fixed points. A consequence is that if the system has a single fixed point and the undamped algorithm converges (respectively damped), then the damped (respectively undamped) algorithm converges and the convergence value is identical between the damped and undamped algorithms. As previously stated in Chapter 4, we eventually consider a continuous-time system as model of the brain (which deals with continuous signals) instead of a discrete system. The slight difference between the three algorithms thus disappears, as all amounts of damping lead to the same continuous system. In the rest of this chapter, we will therefore consider these algorithms as one. Note that all algorithm implemented in the simulations, including Fractional BP, do not have damping in order not to confound the effect of having different algorithms (for instance, Circular versus Fractional BP) with the effects coming from using different damping values.

Tree-reweighted BP as special case of Fractional BP Tree-reweighted BP (Wainwright. et al., 2002, 2003; Wainwright et al., 2005) is a particular case of the algorithms mentioned above. The message update equation of Tree-reweighted BP is identical to Equation (3.3). The only difference is that in Tree-reweighted BP, α_{ij} symbolizes the inverse appearance probability of edge (i, j) in the set of spanning trees and therefore it imposes the constraint $\alpha_{ij} \geq 1$ (Minka, 2005). Note that Fractional BP does not impose any constraints on the value of α_{ij}, which even could be negative.

Variational message-passing as special case of Fractional BP Variational message-passing (Winn and Bishop, 2005), which is the message-passing version of the mean-field method

(Peterson and Anderson, 1987), is a particular case of Fractional BP as well. Indeed, Variational message-passing is the same as Fractional BP for $\boldsymbol{\alpha} = \mathbf{0}$ (Wiegerinck and Heskes, 2002; Minka, 2005). Mean-field is known to be overconfident and perform poorly compared to BP (Weiss, 2001; Mooij and Kappen, 2004). Note that mean-field inference or variational message-passing is different from BP without subtraction or equivalently, full Circular BP (Circular BP with $\boldsymbol{\alpha} = \mathbf{0}$); see also section 4.4.

BP as special case of Fractional BP Finally, BP itself is a special case of Fractional BP, as it corresponds to $\boldsymbol{\alpha} = \mathbf{1}$ in Equation (3.3) as its update equation is:

$$m_{i\to j}^{\text{new}}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) \tag{3.10}$$

3.2.2 Approximating Fractional BP as Circular BP

Message-update equations for the Fractional BP algorithm (Equation (3.3)) and the Circular BP algorithm (Equation (1.12)) are very similar. It becomes even more obvious when considering an Ising model. As a reminder of section 1.4.4, Circular BP (Jardri and Denève, 2013a) is defined in this case as:

$$\begin{cases} M_{i\to j}^{\text{new}} = f_{ij}(B_i - \alpha_{ij} M_{j\to i}) & \text{(3.11a)} \\ B_i = \sum_{j\in\mathcal{N}(i)} M_{j\to i} + M_{\text{ext}\to i} & \text{(3.11b)} \end{cases}$$

where $f_{ij}(x) = \phi^{-1}(\phi(J_{ij})\phi(x))$. This is the same exact system as for standard Fractional BP, except Fractional BP uses function g_{ij} (which depends on α_{ij} and J_{ij}, see its formula in Equation (3.7)) and Circular BP uses f_{ij} (which depends solely on J_{ij}). Nevertheless, functions g_{ij} and f_{ij} are close to each other:

$$g_{ij}(x) \equiv \frac{1}{\alpha_{ij}}\phi^{-1}\Big(\phi\big(\alpha_{ij}J_{ij}\big)\phi\big(x\big)\Big) \approx \phi^{-1}\Big(\phi\big(J_{ij}\big)\phi\big(x\big)\Big) \equiv f_{ij}(x) \tag{3.12}$$

which can be easily justified mathematically for α_{ij} or J_{ij} small enough, and can be seen in practice for various J_{ij} and α_{ij} in Figure 3.5A. This means that the message update equation of Fractional BP approximates the update equation of Circular BP.

As shown in Figure 3.5B, as consequence of the similarity between their own update equations, the Circular BP algorithm is a good approximation of Fractional BP.

Note that Circular BP was initially defined in Jardri and Denève (2013a) (see Equation (1.12)) without any constraints on parameter $\boldsymbol{\alpha}$. In the present thesis (with the exception of section 2.2), the distinction between descending loops and ascending loops is not made as we impose $\boldsymbol{\alpha}$ to be a symmetric matrix: $\alpha_{i\to j} = \alpha_{j\to i} \equiv \alpha_{ij}$. The reason for that is purely mathematical as shown above in this section: α_{ij} is associated to the *undirected* edge (i, j) in Fractional BP and is therefore adirectional: the same parameter α_{ij} is used for the computation of $M_{i\to j}$ and $M_{j\to i}$. Circular BP, seen as an approximation to Fractional BP, thus must respect this constraint on $\boldsymbol{\alpha}$ as well. Relaxation of the constraint (that is, allowing $\alpha_{i\to j} \neq \alpha_{j\to i}$) is considered in paragraph "Breaking symmetry" of section 3.5.4, but hierarchies in the probabilistic graph are not considered here, making it impossible to even talk about ascending or descending loops. Indeed, probabilities are randomly generated, without any notion of hierarchy (see paragraph "Types of graphs used" in section 3.5.1.

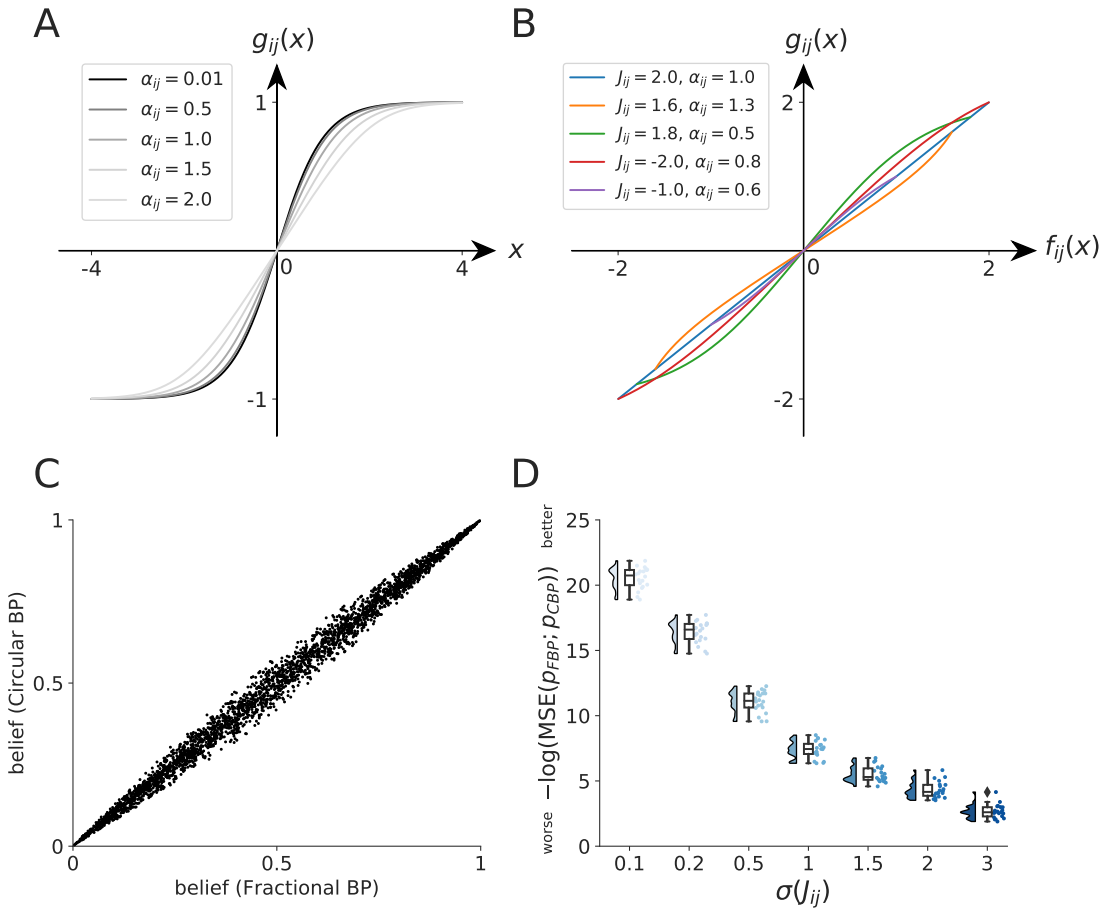

Figure 3.5: **Comparison between Fractional BP and its approximation Circular BP**. **(A)** Visual comparison between the update functions for Fractional BP (function g_{ij}, which depends on the strength of the relationship J_{ij} and α_{ij}) and Circular BP (function f_{ij} which depends only on J_{ij} and corresponds to g_{ij} with $\alpha_{ij} = 1$) for various α_{ij} and given $J_{ij} = 1$. All functions have the identical sigmoidal shape. **(B)** The update functions for Circular BP (f_{ij}) and for Fractional BP (g_{ij}) are approximately identical, for various J_{ij} and α_{ij} **(C)** Example of marginals produced by the Fractional BP algorithm versus produced by the Circular BP algorithm for a given random graph with randomly generated weights and random $\boldsymbol{\alpha}$. **(D)** Influence of the interaction strength $\sigma(J_{ij})$ on the approximation: a lower J_{ij} leads to a better fit between Circular BP and Fractional BP. One point represents different randomly generated generated $\boldsymbol{\alpha}$ for the same graph.

3.3 Extended Circular BP

In this section, an extension to the Circular BP algorithm, called subsequently *extended Circular BP*, is proposed, with additional parameters ($\boldsymbol{\kappa}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$) on top of existing ones ($\boldsymbol{\alpha}$). This extension builds up on an extension to Fractional BP known as Generalized Belief Propagation (Yedidia et al., 2005). This extended Fractional BP algorithm is approximated similarly to section 3.2.2, (where Circular BP is seen as an approximation to Fractional BP) to define extended Circular BP.

3.3.1 Extended Fractional BP algorithm (Generalized BP)

Definition As seen above, the Fractional BP algorithm is defined as a result of approximating the entropy of the variational distribution $b(\mathbf{x})$. This is a particular case of approximating the Gibbs free energy - the difference between the average entropy and the variational entropy - as follows:

$$\begin{aligned} G_{\text{approx}} = &\sum_{(i,j)} \frac{1}{\alpha_{ij}} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log\Big(\frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)}\Big) - \sum_{(i,j)} \beta_{ij} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log\Big(\psi_{ij}(x_i,x_j)\Big) \\ &+ \sum_i \frac{1}{\kappa_i} \sum_{x_i} b_i(x_i) \log\Big(b_i(x_i)\Big) - \sum_i \gamma_i \sum_{x_i} b_i(x_i) \log\Big(\psi_i(x_i)\Big) \end{aligned} \tag{3.13}$$

where $(\mathbf{1}/\boldsymbol{\alpha}; \boldsymbol{\beta}; \mathbf{1}/\boldsymbol{\kappa}; \boldsymbol{\gamma})$ are called *counting numbers*, with *entropic counting numbers* $(\mathbf{1}/\boldsymbol{\alpha}; \mathbf{1}/\boldsymbol{\kappa})$ and *average energy counting numbers* $(\boldsymbol{\beta}; \boldsymbol{\gamma})$. Fractional BP corresponds to $(\boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$, and BP is associated with to $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$. This approximation of the Gibbs free energy can be justified by approximating both the variational entropy and the average energy. This is a special case of Yedidia et al. (2005) which presents a method to build generalized BP algorithms based on such an approximation. We consider in this section the situation where the regions on which to make the approximation of the Gibbs free energy are simply the set of graph edges and individual nodes, as in the Bethe approximation. Another difference is that the work of Yedidia and colleagues focuses on so-called "*valid* region-based approximations" which imposes constraints between the counting numbers. These constraints are not considered in the present work.

We approximate the entropy of the variational distribution $b(x)$ as if it could be written as function of its unitary and pairwise marginals $b_i(x_i)$ and $b_{ij}(x_i, x_j)$ as:

$$b(\mathbf{x}) \approx \prod_{i,j} \Big(\frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)}\Big)^{1/\alpha_{ij}} \prod_i (b_i(x_i))^{1/\kappa_i} \tag{3.14}$$

and if $b(x)$ could be written as function of its unitary and pairwise potentials $\psi_i(x_i)$ and $\psi_{ij}(x_i, x_j)$ as:

$$b(\mathbf{x}) \approx \prod_{i,j} \big(\psi_{ij}(x_i,x_j)\big)^{\beta_{ij}} \prod_i \big(\psi_{ij}(x_i,x_j)\big)^{\gamma_i} \tag{3.15}$$

Note that parameters $(\boldsymbol{\kappa}, \boldsymbol{\gamma})$ are nodal terms (κ_i is assigned to variable node x_i). On the other hand, ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) are related to edges of the graph (α_{ij} is assigned to the undirected edge node (x_i, x_j)).

This gives the following parametric approximation of the entropy of $b(\mathbf{x})$:

$$-S_b \approx \sum_{(i,j)} \frac{1}{\alpha_{ij}} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log\Big(\frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)}\Big) + \sum_i \frac{1}{\kappa_i} \sum_{x_i} b_i(x_i) \log(b_i(x_i)) \tag{3.16}$$

with entropy counting numbers $(\mathbf{1}/\boldsymbol{\alpha}, \mathbf{1}/\boldsymbol{\kappa})$.

The following parametric approximation of the average energy is written:

$$U_b \approx -\sum_{(i,j)} \beta_{ij} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log\Big(\psi_{ij}(x_i,x_j)\Big) - \sum_i \gamma_i \sum_{x_i} b_i(x_i) \log\Big(\psi_i(x_i)\Big) \tag{3.17}$$

with average energy counting numbers $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

These approximations of the variational entropy (Equation (3.16)) and average energy (Equation (3.17)) lead to the approximated Gibbs free energy $G_{\text{approx}} \equiv U_{\text{approx}} - S_{\text{approx}}$ given in Equation (3.13). This parametric approximation of the Gibbs free energy generalizes the Bethe free energy used in BP.

This eventually leads (see Appendix A) to a generalized BP or extended Fractional BP (eFBP) defined by the following update equation:

$$m_{i\to j}^{\text{new}}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i,x_j)^{\alpha_{ij}\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i} \right)^{\kappa_i} \right)^{1/\alpha_{ij}} \tag{3.18}$$

and beliefs are computed using:

$$b_i(x_i) \propto \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i) \right)^{\kappa_i} \tag{3.19}$$

eFBP in the binary case This class of generalized BP algorithms still takes a very simply form in the log-domain when applied to probability distributions over binary variables:

$$\begin{cases} M_{i\to j} = g_{ij}\Big(B_i - \alpha_{ij} M_{j\to i}\Big) & \text{(3.20a)} \\ B_i = \kappa_i \Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i} \Big) & \text{(3.20b)} \end{cases}$$

where g_{ij} is a sigmoidal function given by:

$$g_{ij}(x) = \frac{1}{2\alpha_{ij}} \log\left(\frac{\left(\psi_{ij}^{1,1}\right)^{\alpha_{ij}\beta_{ij}} e^{2x} + \left(\psi_{ij}^{1,0}\right)^{\alpha_{ij}\beta_{ij}}}{\left(\psi_{ij}^{0,1}\right)^{\alpha_{ij}\beta_{ij}} e^{2x} + \left(\psi_{ij}^{0,0}\right)^{\alpha_{ij}\beta_{ij}}} \right) \tag{3.21}$$

in the general case, and

$$g_{ij}(x) = \frac{1}{\alpha_{ij}} \phi^{-1}\Big(\phi\Big(\alpha_{ij}\beta_{ij} J_{ij}\Big) \phi(x) \Big) \tag{3.22}$$

for an Ising model, where $\psi_{ij}(x_i,x_j) \propto \exp(J_{ij} x_i x_j)$. These expressions of g_{ij} generalize the ones used in Fractional BP (see Equations (3.6) and (3.7)).

A neural implementation of eFBP Note the presence of parameter α_{ij} at several places in the algorithm (multiplicative factor to the message in Equation (3.20a), multiplicative factor to the weights in Equation (3.22), and divisive factor in the non-linearity in Equation (3.22)) does not prevent eFBP from being potentially implemented in the brain as such, contrary to FBP.

Indeed, extended FBP in Ising models can be rewritten:

$$\begin{cases} B_j^i = \phi^{-1}\Big(\phi\big(\tilde{\beta}_{ij}J_{ij}\big)\phi\big(B_i - B_i^j\big)\Big) & \text{(3.23a)} \\ B_i = \kappa_i\Big(\sum\limits_{j\in\mathcal{N}(i)} \dfrac{1}{\alpha_{ij}}\phi^{-1}\Big(\phi\big(\tilde{\beta}_{ij}J_{ij}\big)\phi\big(B_i - B_i^j\big)\Big) + \gamma_i M_{\text{ext}\to i}\Big) & \text{(3.23b)} \end{cases}$$

with $B_i^j = \alpha_{ij}M_{j\to i}$ and $\tilde{\beta}_{ij} \equiv \alpha_{ij}\beta_{ij}$ which incorporates the dependence in α_{ij}. In this case, κ_i represents, as for extended Circular BP, the synaptic scaling factor associated to the unit encoding for B_i. However, α_{ij} is not the synaptic scaling factor associated to the unit encoding for B_i^j as for eCBP but instead, (the inverse of) a connection weight between the unit encoding for B_i^j and the unit encoding for B_i^j; see system of equations (4.13) for comparison with eCBP.

Relation to other algorithms The extended Fractional BP algorithm generalizes Fractional BP (Wiegerinck and Heskes, 2002), Power EP (Minka, 2004) and α-BP (Liu et al., 2019) which all correspond to $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$, and more particularly use the damped message update equation (A.8) rather than its undamped version (A.9) (see damping in section 4.2.1).

Of course, extended Fractional BP extends particular cases of the Fractional BP algorithm, including the Belief Propagation algorithm (recovered for $(\boldsymbol{\kappa}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$).

Lastly, Circular BP (Jardri and Denève, 2013a) is not associated to any choice of $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ in eFBP, thus does not come from any modification of the Bethe Free Energy. However, as previously explained in section 3.2.2, Circular BP can be seen as an approximation of Fractional BP. Furthermore, and contrary to the models mentioned above, Circular BP was proposed as a way of disturbing inference capabilities of the network rather than improving them, in order to capture the effects of excitation-inhibition imbalance on mental states and belief formation.

3.3.2 Extended Circular BP as approximation of extended Fractional BP

We introduce in this section a generalization of Circular BP called *extended Circular BP*. This algorithm has additional parameters $\boldsymbol{\kappa}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ compared to BP. Extended Circular BP (eCBP) is defined by the following message update equation (which generalizes the one for Circular BP in Equation (1.12)):

$$m_{i\to j}^{\text{new}}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)^{\beta_{ij}} \Big(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i}\Big)^{\kappa_i} \tag{3.24}$$

and the beliefs are computed using:

$$b_i(x_i) \propto \Big(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i)\Big)^{\kappa_i} \tag{3.25}$$

This translates, for probability distributions over binary variables, into the following equation in the log-domain:

$$\begin{cases} M_{i\to j}^{\text{new}} = f_{ij}\Big(B_i - \alpha_{ij}M_{j\to i}\Big) & \text{(3.26a)} \\ B_i = \kappa_i\Big(\sum\limits_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big) & \text{(3.26b)} \end{cases}$$

Algorithm 3 Extended Circular Belief Propagation algorithm in a pairwise factor graph

for all directed edges $i \to j$ **do**
  $m_{i\to j}(x_j) \leftarrow$ some distribution {Initialize the messages}
end for
repeat
  for all directed edges $x_i \to x_j$ **do**
    $m^{\text{new}}_{i\to j}(x_j) \leftarrow \sum_{x_i} \psi_{ij}(x_i, x_j)^{\beta_{ij}} \Big(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i}\Big)^{\kappa_i}$
    {Update the messages}
  end for
  $m \leftarrow m^{\text{new}}$
until convergence
for all nodes x_i **do**
  $b_i(x_i) \leftarrow \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i)\right)^{\kappa_i}$ {Compute the beliefs}
end for

where

$$f_{ij}(x) = \frac{1}{2}\log\left(\frac{\left(\psi^{1,1}_{ij}\right)^{\beta_{ij}} e^{2x} + \left(\psi^{1,0}_{ij}\right)^{\beta_{ij}}}{\left(\psi^{0,1}_{ij}\right)^{\beta_{ij}} e^{2x} + \left(\psi^{0,0}_{ij}\right)^{\beta_{ij}}}\right) \tag{3.27}$$

in the general case, and

$$f_{ij}(x) = \phi^{-1}\Big(\phi(\beta_{ij}J_{ij})\phi(x)\Big) \tag{3.28}$$

in the specific case of Ising models. Note that the expression of f_{ij} in extended Circular BP generalizes the formula for f_{ij} in Circular BP (Equation (1.21)) which corresponds to the case $\boldsymbol{\beta} = \mathbf{1}$.

In the remainder of this thesis, unless explicitly stated otherwise, all algorithms (e.g. Circular BP) will be considered in the binary case described above.

Algorithm 4 Binary extended Circular BP algorithm in a pairwise factor graph

for all directed edges $i \to j$ **do**
  $M_{i\to j} \leftarrow$ random value {Initialize the messages' log-odds}
end for
repeat
  for all nodes x_i **do**
    $B_i = \kappa_i\Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big)$ {Compute the beliefs' log-odds}
  end for
  for all directed edges $x_i \to x_j$ **do**
    $M^{\text{new}}_{i\to j} = f_{ij}(B_i - \alpha_{ij}M_{j\to i})$ {Update the messages' log-odds}
  end for
  $M \leftarrow M^{\text{new}}$
until convergence
for all nodes x_i **do**
  $b_i(x_i = \pm 1) \leftarrow \sigma(\pm 2B_i)$ {Compute the beliefs}
end for

This generalization of Circular BP is connected to the extended Fractional BP algorithm (or Generalized BP) previously described. More specifically, eCBP is built as an approximation to eFBP, using the fact that the update functions f_{ij} and g_{ij}, which are the only thing that differenciates eFBP from eCBP in Ising models, are approximately equal:

$$g_{ij}(x) \equiv \frac{1}{\alpha_{ij}}\phi^{-1}\Big(\phi\Big(\alpha_{ij}\beta_{ij}J_{ij}\Big)\phi\big(x\big)\Big) \approx \phi^{-1}\Big(\phi\Big(\beta_{ij}J_{ij}\Big)\phi\big(x\big)\Big) \equiv f_{ij}(x) \tag{3.29}$$

3.4 Convergence of the proposed algorithm

As seen in Figures 3.6 and 3.7, convergence of the algorithm is crucial to carry out approximate inference: it can be observed a sharp transition between having beliefs very close to the correct marginals and very far from them. When one of the algorithms does not converge, beliefs have very little to do with the correct marginals (this observation was already made for BP in Murphy et al. (1999)). It is therefore important to control the convergence properties, and if possible, to ensure the convergence of our proposed algorithm.

Importantly, for all the variants of BP cited previously (Fractional BP, Circular BP, Variational message-passing, ...), there is no theoretical result about the existence of parameters that guarantee convergence. Absence of convergence of an algorithm means that it gets trapped in a limit cycle (*frustration* phenomenon) or alternatively wanders around chaotically without being able to produce approximate marginals. The typical behavior is a strong oscillation of the beliefs from one extreme ($b_i(x_i = 1) \approx 100\%$) to the other ($b_i(x_i = 1) \approx 0\%$). This is an serious obstacle to the use of such algorithms to perform approximate inference, let alone be a candidate for neural implementation of inference.

On the contrary, we prove here that in the case of Ising models, there exist parameters for which extended Circular BP is stable. More precisely, for any set of parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, one can choose parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa})$ such that the algorithm converges (see Theorem 3). Note that a very similar demonstration applies to extended Fractional BP as well.

Having a stable algorithm is important, for two reasons. First, it helps performing reasonably good approximate inference with the algorithm for *any* probability distribution. Indeed, whatever the probability distribution is (i.e., the weights $\boldsymbol{J}$ and external inputs $\boldsymbol{M}_{\text{ext}}$ are), it is possible to select parameters for which the algorithm will produce approximate marginals. Second, and most importantly, convergence is important for the biological plausibility of the algorithm. Not only the brain should remain in a stable regime while performing computations (the contrary can be seen as epilepsy), estimations of probabilities should also not depend on the past state of the system (meaning that the system should have only one fixed point for any probability distribution, and the system should converge to this fixed point).

3.4.1 Convergence results

Here we state sufficient conditions for the convergence of extended Circular BP in an Ising model.

We start by defining matrix $\boldsymbol{A}$ whose coefficients are:

$$A_{i\to j,k\to l} = |\kappa_i| \tanh\big|\beta_{ij}J_{ij}\big| \, \delta_{il} \mathbb{1}_{\mathcal{N}(i)}(k) \Big|1 - \frac{\alpha_{ij}}{\kappa_i}\Big|_{j=k} \tag{3.30}$$

where $\mathbb{1}_{\mathcal{N}(i)}(k) = 1$ if $k \in \mathcal{N}(i)$, otherwise $= 0$, $\delta_{il} = 1$ if $i = l$, otherwise $= 0$, and $\Big|1 - \frac{\alpha_{ij}}{\kappa_i}\Big|_{j=k} = 1 - \frac{\alpha_{ij}}{\kappa_i}$ if $j = k$, otherwise 1.

Theorem 1 *If for any induced operator norm $\|\cdot\|$ (sometimes called natural matrix norm), $\|A\| < 1$, then eCBP converges to a unique fixed point and the rate of convergence is at least linear.*

Proof 3.4.1 *The proof of Theorem 1 follows closely the one of Lemma 2 of Mooij and Kappen (2007b) for Belief Propagation, that is, the special case $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$. The message update equation for extended Circular BP, given by Equations (3.26a) and (3.26b), is:*

$$M_{i\to j}^{new} = \tanh^{-1}\left[\tanh\left(\beta_{ij}J_{ij}\right)\tanh\left(\kappa_i\gamma_i M_{ext\to i} + \kappa_i \sum_{k\in\mathcal{N}(i)\setminus j} M_{k\to i} + (\kappa_i - \alpha_{ij})M_{j\to i}\right)\right] \tag{3.31}$$

Let F be the update function for the messages: $\mathbf{M}^{new} = F(\mathbf{M}^{old})$ (in what follows, we drop the superscript old for clarity). The derivative of F is:

$$F'(\mathbf{M})_{i\to j,k\to l} = \frac{\partial M_{i\to j}^{new}}{M_{k\to l}} \tag{3.32}$$

$$= \tilde{A}_{i\to j,k\to l} B_{i\to j}(\mathbf{M}) \tag{3.33}$$

where

$$\tilde{A}_{i\to j,k\to l} = |\kappa_i| \tanh\left|\beta_{ij}J_{ij}\right| \delta_{il} \mathbb{1}_{\mathcal{N}(i)}(k)\left(1 - \frac{\alpha_{ij}}{\kappa_i}\right)_{j=k} \tag{3.34}$$

$$B_{i\to j}(\mathbf{M}) = \operatorname{sgn}(\kappa_i\beta_{ij}J_{ij}) \frac{1 - \tanh^2\left(\kappa_i\gamma_i M_{ext\to i} + \kappa_i \sum_{k\in\mathcal{N}(i)\setminus j} M_{k\to i} + (\kappa_i - \alpha_{ij})M_{j\to i}\right)}{1 - \tanh^2\left(M_{i\to j}^{new}\right)} \tag{3.35}$$

Note that $\sup_{\mathbf{M}} B_{i\to j}(\mathbf{M}) = 1$, as $\sup_x \frac{1 - \tanh^2(x)}{1 - \tanh^2(K)\tanh^2(x)} = 1$.
It comes that $\sup_{\mathbf{M}} F'(\mathbf{M})_{i\to j,k\to l} \leq A_{i\to j,k\to l}$ where we defined A as $A = \left|\tilde{A}\right|$. A does not depend on the external messages $\mathbf{M}_{ext}$, nor does it depend on parameter $\boldsymbol{\gamma}$.
If for any norm $\|\cdot\|$ on matrices, $\|A\| < 1$, $\sup_{\mathbf{M}} F'(\mathbf{M})_{i\to j,k\to l} < 1$, then function F is a $\|\cdot\|$ - contraction (see Lemma 2 of Mooij and Kappen (2007b)) and the sequence $M, F(M), F \circ F(M), \ldots$, that is, the Circular BP algorithm, converges to a unique fixed point with at least a linear rate. □

Hence choosing in Theorem 1 the spectral norm (induced by the l_2-norm), it comes straightforwardly:

Corollary 1.1 *If the largest singular value of A, $\sigma_{\max}(A) < 1$, then Circular BP converges.*

Theorem 2 *If $\forall(i,j)$, $\alpha_{ij}/\kappa_i \leq 1$ and the spectral radius of A, $\rho(A) < 1$, then Circular BP converges to the unique fixed point.*

Proof 3.4.2 *The proof of Theorem 2 follows closely the one of Corollary 3 of Mooij and Kappen (2007b). According to Equation (3.33),*

$$F'(\mathbf{M})_{i\to j,k\to l} = \tilde{A}_{i\to j,k\to l} B_{i\to j}(\mathbf{M}) \tag{3.36}$$

Note that $A \equiv \left|\tilde{A}\right| = \tilde{A}$ *as* $\tilde{A}$ *is non-negative (because of the hypothesis* $\alpha_{ij}/\kappa_i \leq 1$*). We directly conclude by using Theorem 2 of Mooij and Kappen (2007b):* $F'(\mathbf{M})$ *is the product of* A*, a constant non-negative matrix, with* $B(\mathbf{M})$ *whose coefficients are bounded by 1 in absolute value, thus the sequence* $M, F(M), F \circ F(M), \ldots$*, meaning, the Balanced Circular BP algorithm, converges to a fixed point and this fixed point does not depend on the initial condition* M*.* □

Theorem 2 has the strong constraint that $\alpha_{ij}/\kappa_i \leq 1$ for all (i, j). However, when this constraint is verified, the spectral radius criterion is sharper than the norm criterion of Theorem 1 because $\rho(A) \leq \|A\|$ for all induced operator norms.

Finally, a consequence of Theorem 2 is the following fundamental result, which distinguishes extended Circular BP from related approaches like Power EP, Fractional BP and α-BP (see section 3.2.1.2):

Theorem 3 *For a given weighted graph (defined by its weights* J_{ij}*), it is always possible to find parameters* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\kappa}$ *such that extended Circular BP converges for any external input* $\boldsymbol{M}_{ext}$ *and any choice of parameters* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$*.*

Proof 3.4.3 *Here we show Theorem 3. Let us take* $\alpha_{ij} = \kappa_i \equiv p \in \mathbb{R}_+$*. In this case,* $A_{i\to j,k\to l} = p \tanh\left|\beta_{ij} J_{ij}\right| \delta_{il} \mathbb{1}_{\mathcal{N}(i)\setminus j}(k)$*. When* $p \to 0$*, all coefficients of* A *go to zero. The spectral radius is a continuous application, and the null matrix has a spectral radius of zero, thus the spectral radius of* A *goes to zero when* $p \to 0$*. We conclude by using Theorem 2 as* $\alpha_{ij}/\kappa_i = 1 \leq 1$*: there exists* $p^\star$ *such that for all* $p < p^\star$*, eCBP converges.* □

Notably, proof of Theorem 3 shows that choosing $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ uniformly, equal and large enough guarantees the convergence of extended CBP.

We use this result to initialize parameters in the supervised fitting procedure (gradient-descent based). $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are first set at the BP value of $p = 1$, and we decrease p until the spectral radius of matrix $\boldsymbol{A}$ goes below 1, which ensures that extended Circular BP converges according to the theory. In practice, we decrease p by incrementing $1/p$ by steps of 1. Note that we are using a sufficient condition, thus extended Circular BP could converge for higher values of p than the one chosen.

3.4.2 Extension of the convergence results to other related algorithms

As shown above, extended Circular BP (with parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$) converges to the unique fixed point, whatever the probability distribution and parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are, given the right choice of parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa})$.

In particular, for $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1})$, eCBP with parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa})$ has the same convergence properties as the general eCBP. In fact, it is the combination of these two parameters that makes the convergence possible. Without parameter $\boldsymbol{\kappa}$ (i.e., with $\boldsymbol{\kappa} = \mathbf{1}$), it is not possible to guarantee that there exists $\boldsymbol{\alpha}$ such that the algorithm converges. We go even further by stating the following conjecture: there exist weighted graphs (typically, with strong enough weights) for which, for any choice of parameter $\boldsymbol{\alpha}$, Circular BP (and similarly, Fractional BP, both having $\boldsymbol{\kappa} = \mathbf{1}$) does not converge. As a reminder, in the implementation of eCBP in chapter 4, parameter κ_i corresponds to the synaptic scaling factor at the unit encoding the marginal probability of node x_i, and α_{ij} is the synaptic scaling factor at the unit encoding for the prediction of node x_i by node x_j. It is rather intuitive that controlling the scaling factors of all units through $\boldsymbol{\kappa}$ and $\boldsymbol{\alpha}$) is necessary to ensure the stability of the system.

Very similarly, eCBP converges to the unique fixed point, whatever the probability distribution and parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\gamma})$ are, given the right choice of parameter $\boldsymbol{\beta}$. This can be seen easily: in the demonstration of Theorem 3, coefficients of A go to zero as well if $\boldsymbol{\beta} \to \mathbf{0}$, thus for $\boldsymbol{\beta}$ uniform and sufficiently small, the algorithm converges to the unique fixed point. One way to interpret this is that the weak connections in the forming brain allow him to remain in a stable regime and perform inference.This convergence condition remains valid for all extended BP variants (not only eCBP but also eFBP, eBP, and eCBP with $\boldsymbol{\alpha} = \mathbf{0}$ for instance).

All the convergence results stated for eCBP remain valid for eFBP (the demonstrations are similar to the ones of eCBP).

As stated above, the convergence result for eCBP (and eFBP) distinguishes extended Circular BP from related approaches like Circular BP, Power EP, Fractional BP and α-BP (see section 3.2.1.2). Indeed, as seen above, $\boldsymbol{\alpha}$ alone is not enough to ensure convergence of these algorithms: an additional parameter $\boldsymbol{\kappa}$ would be needed but is not present in these algorithms. However, controlling parameter $\boldsymbol{\kappa}$ alone (i.e., taking fixed $\boldsymbol{\alpha}$) is not sufficient to guarantee the convergence of extended CBP, nor extended BP (for which $\boldsymbol{\alpha} = \mathbf{1}$), but is enough for extended CBP with fixed $\boldsymbol{\alpha} = \mathbf{0}$. Note that not being able to guarantee the convergence of an algorithm does not mean that the algorithm does not converge. Simulations seem to indicate that for strong enough weights and strong enough α values, $\kappa \to 0$ is not enough to make eCBP converge. However, such an approach is successful on all the graphs used for which the weights are moderately strong $(J_{ij} \sim \mathcal{N}(0, 1))$.

3.5 Numerical experiments - Learning to outperform BP

The goal is to use the extended Circular BP algorithm in order to perform approximate inference for any given external input $\boldsymbol{M}_{\text{ext}} = \{M_{\text{ext},i}\}$ given the interactions $\boldsymbol{J} = \{J_{ij}\}$ between variables of the probability distribution. To achieve this goal, parameters of the model $\boldsymbol{\alpha} = \{\alpha_{ij}\}$, $\boldsymbol{\kappa} = \{\kappa_i\}$, $\boldsymbol{\beta} = \{\beta_{ij}\}$ and $\boldsymbol{\gamma} = \{\gamma_i\}$ are learnt so that the approximate marginals or beliefs $\{b_i(x_i)\}$ are as close as possible to the true marginals $\{p_i(x_i)\}$, for all possible external inputs $\mathbf{M}_{\text{ext}}$ (meaning, unitary factors $\psi_i(x_i)$) given a graph with interactions $\boldsymbol{J}$ (i.e., pairwise factors $\psi_{ij}(x_i, x_j) = \exp(J_{ij} x_i x_j)$). This would mean that one can learn the appropriate parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ so that inference is possible on all probability distributions written as $p(\mathbf{x}) \propto \prod_{(i,j)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i)$ where $\{\psi_{ij}\}$ are fixed but $\{\psi_i\}$ can vary. According to the implementation of chapter 4, that corresponds to the brain having relations between variables $(\boldsymbol{J})$ as well as parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ encoded in connections between units such that the activity of unit i reflects the marginal probability of variable x_i, whatever the sensory information are.

3.5.1 Experimental setting

Types of graphs used We consider Ising models for the numerical simulations, meaning that the probability distribution can be factorized into a product of pairwise potentials $\phi_{ij}(x_i, x_j) \propto \exp(J_{ij} x_i x_j)$, and unitary potentials $\phi_i(x_i)$, where variable x_i is a binary variable. Several graph topologies are considered: structured graphs and Erdős-Rényi random graphs. Erdős-Renyi graphs are generated with various connection probabilities p (ranging from 0.2 to 1); see Figure 3.6. Structured graphs are generated according to predefined structures (grid, ladder, bipartite graph, ...) identical to Yoon et al. (2018); see Figure 3.7. All graphs have 9 nodes.

Artificial data generated For each graph topology, 30 graphs are generated: an absence of edge means $J_{ij} = 0$ while existing edges had their weights sampled randomly according

to $J_{ij} \sim \mathcal{N}(0,1)$ (spin-glass). As a reminder, J_{ij} is associated to the *unoriented* edge (i,j): $J_{i\to j} = J_{j\to i} \equiv J_{ij}$. For each weighted graph, 200 training examples, 100 validation examples and 100 test examples are generated, where an example is a vector of external evidences $\boldsymbol{M}_{\text{ext}}$ generated according to $M_{\text{ext},i} \sim \mathcal{N}(0,1)$.

3.5.2 Learning procedure

3.5.2.1 Description of the learning procedure

We use supervised learning to fit the $2n_{\text{nodes}} + 2n_{\text{edges}}$ parameters of eCBP $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$: matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have n_{edges} degrees of freedom and vectors $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ have n_{nodes} degrees of freedom. To do so, we want to minimize the difference between beliefs obtained with eCBP, and the true marginals. More specifically, the loss being minimized is the MSE loss or squared L2 norm:

$$\mathcal{L}(b,p) = \frac{1}{n_{\text{nodes}}} \sum_{i=1}^{n_{\text{nodes}}} \big(b_i(x_i = +1) - p_i(x_i = +1)\big)^2 \tag{3.37}$$

between the true marginals $p_i(x_i)$ and the approximate ones $b_i(x_i)$, obtained by propagating messages for $T = 100$ time steps. True marginals $p_i(x_i)$ are computed using an exact inference algorithm, the Junction Tree algorithm, via the pgmpy library (Ankan and Panda, 2015). The model is trained on PyTorch (Paszke et al., 2019) using backpropagation through time with a gradient-descent based method, RPROP (Riedmiller and Braun, 1993). We use a learning rate of 0.001 and stop the learning once the validation loss saturates, which usually takes around 50 epochs. Parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ of the model are initialized such that eCBP converges with this choice of parameters (see section 3.4). One initialization tested, among others, is to initialize $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ at the value $\mathbf{1}$, while $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are initialized at the same small enough value; see Theorem 3 and its proof. The model is also trained using function *least squares* of SciPy (Virtanen et al., 2020). We take the best fitting parameters between PyTorch and SciPy depending on their performance on the validation set. The source code is publicly available on GitHub.

Marginals obtained after fitting the eCBP algorithm can be seen in Figure 3.6 for Erdos-Renyi graphs, and Figure 3.7 for structured graphs.

Many other models than eCBP are fitted, using the same learning procedure. This includes special cases of eCBP (with one or more of the parameters being fixed) but different classes of algorithms as well, like eFBP, the classical rate network, etc. Each model is initialized when possible with parameters which guarantee convergence of the algorithm (otherwise the learning procedure sometimes does not find a region of convergence). This can be done through the combination of low and positive $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ as for eCBP, but also with low enough $\boldsymbol{\beta}$ (see section 3.4.2) and lastly, with low enough $\boldsymbol{\kappa}$ (even if there is no theoretical result about the spectral radius of A being lower than one if $\boldsymbol{\kappa}$ is sufficiently low, it is the case in practice with the generated graphs and thus guarantees convergence).

3.5.2.2 Generalization ability

Here we show that the extended Circular BP algorithm, using the learning procedure described above, is able to generalize to new data. This shows the goodness of the learning procedure.

Within-set and out-of-set generalization A first necessary check for the supervised learning procedure is to make sure that the proposed model is able to learn properly the training data and to generalize to new inputs.

The training indeed learns to represent the training data and allows the model to generalize well to unseen inputs with identical statistics as in the training set; see Figure 3.8A. In other words, the extended Circular BP algorithm is able to generalize to unseen situations, where a "situation" is a vector $\boldsymbol{M}_{\text{ext}}$ of size n_{nodes}. *Unseen* refers to data not included in the training examples. Note that only the external inputs $\boldsymbol{M}_{\text{ext}}$ change in the test set: the graph weights $\boldsymbol{J}$ are fixed. In other words, parameters $\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are learnt given the weights $\boldsymbol{J}$, although for any external inputs $\boldsymbol{M}_{\text{ext}}$.

The model also generalizes well to out-of-set inputs, that is, to external inputs $\boldsymbol{M}_{\text{ext}}$ with different statistics from the ones used for training; see Figure 3.8B.

Extended Circular BP outperforms its special cases The extended Circular BP algorithm is expected to outperform all its special cases, including Circular BP (i.e., $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$), extended BP (i.e., $\boldsymbol{\alpha} = \mathbf{1}$) and BP (i.e., $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$). This is indeed the case, as shown in Figure 3.9 where models are ordered according to the following score (average over graphs of the log of the MSE averaged over examples) written "$-\log_{10}$(avg MSE)" in figures:

$$\text{score} = -\frac{1}{N_{\text{graphs}}} \sum_{\text{graphs}} \log_{10} \left(\frac{1}{n_{\text{examples}}} \sum_{\text{examples}} \text{MSE}_{\text{graph, example}} \right) \tag{3.38}$$

where the MSE loss is defined in Equation (3.37). The performance of all models decreases with the complexity of graphs. Several models like BP perform very poorly for dense graphs because the network becomes frustrated (see Figure 3.6). It is the true for Circular BP as well, which shows that parameter $\boldsymbol{\alpha}$ alone is not enough to make the algorithm converge. On the contrary, other models (all models outperforming or equal to "CBP nodal") show no sign of frustration thanks to additional parameters, as expected by Theorem 3. This allows all these models, including extended CBP, to keep a rather good level of performance in the approximate inference task even for complete graphs.

The only exception to the statement that more complex models perform better on the test set is the fact that the model "CBP + weights" (i.e., eCBP with $(\boldsymbol{\kappa}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1})$) perfoms worse than the model "BP + weights" (i.e., eCBP with $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$) for highly dense graphs. The reason for that is not overfitting as "CBP + weights" also performs worse than "BP + weights" on the training set. Instead, the supervised learning algorithm does not manage in this case, because it has to explore over a larger parameter space, to find a better solution than if it has less parameters to fit.

More generally, and quite logically, models which were special cases of others performed comparatively worse on the test set. For instance, as shown in Figure 3.10, extended Fractional BP outperforms, among others, Fractional BP, Generalized BP (Yedidia et al., 2001), Tree-Reweighted BP (Wainwright. et al., 2002) and Variational message-passing (Winn and Bishop, 2005), that is, mean-field, as all these algorithms are special cases of extended Fractional BP: see section 3.2.1.2 for more details. We used the libDAI (Mooij, 2010) implementation of Generalized BP ("GBP_MIN" algorithm, i.e., using minimal clusters: one outer region for each maximal factor), Tree-Reweighted BP (with 10000 sampled trees) and Variational message-passing, to compute the associated approximate marginals.

Among all tested algorithms, it is noteworthy that neither Circular BP nor Fractional BP, for which the only only parameter fitted is $\boldsymbol{\alpha}$, provides good results for moderately dense or highly dense graphs. By looking at the marginals produced by these algorithms for Erdos-Renyi graphs (see Figure 3.6), the reason why they do not perform well is the frustration of the network caused by cycles (strong oscillations between the two extremes $b_i(x_i) \approx 0$ and $b_i(x_i) \approx 1$ without convergence of the algorithm). As observed by Murphy et al. (1999), convergence of BP implies

a good approximation of the correct marginals by the beliefs, and convergence can be forced by using a damped algorithm. Similarly, here, when Circular and Fractional BP converge (for graph with low density), they produce beliefs close to the correct marginals. However, for denser graphs, the algorithms do not converge and in this case beliefs have very little to do with the correct marginals, which Murphy pointed as well. As in Murphy et al. (1999), introducing damping to CBP and FBP would help the algorithms to converge. However, the damped algorithm would become slower to converge in cases where the undamped algorithm converged. This does not seem like an optimal solution to the frustration problem, as the inference performed by the brain should be as fast as possible in particular to react quickly in situations of danger. The question whether the algorithms can converge only by choosing the right $\boldsymbol{\alpha}$, without damping, has not been answered yet. It is probable that the $\boldsymbol{\alpha}$ parameter is not enough to prevent the frustration behavior from occurring. Alternately, it is possible that the learning procedure does not find the optimal $\boldsymbol{\alpha}$. This second possibility as yet not been ruled out.

Influence of the interaction strength Figure 3.11 shows the effect of increasing the interactions weights $\boldsymbol{J} = \{J_{ij}\}$. Both BP and eCBP show a performance decrease with increased interaction strength, but at a much higher rate for BP than for eCBP.

Overconfidence Overconfidence is defined as having a stronger certainty about one's belief compared to the evidence at hand. Overconfidence can be explained by the effect of cycles: each piece of evidence is being counted multiple times as it is reverberated in the cyclic graph. This situation can be observed for BP in Figure 3.11. More generally, we observed systematic overconfidence in BP (or alternatively, systematic underconfidence, depending on the weights of the graph); see also Weiss (2000). With extended Circular BP, this systematic disturbance disappeared thanks to training, as shown in Figure 3.11. In other words, the eCBP algorithm manages to cancel the effects of information amplification (respectively, dampening) by successive passes in the cycles.

3.5.3 Comparison between Fractional and Circular BP

Not only (extended) Circular BP algorithms outperform their (extended) BP equivalents as previously shown in Figure 3.9, they also are close - but inferior - to (extended) Fractional BP algorithms in terms of performance as shown in Figure 3.12. Logically, all approximate inference algorithms show a decreased performance for increased network density, both for structured and Erdos-Renyi graphs, but at a higher rate for BP models than for Circular or Fractional BP models.

3.5.4 Additional analyses

Breaking symmetry In the algorithms, parameters $\boldsymbol{\alpha} = \{\alpha_{ij}\}$ and $\boldsymbol{\beta} = \{\beta_{ij}\}$ are symmetric matrices as elements of the matrix depend on the *unoriented* edge (for instance, for $\boldsymbol{\alpha}$, $\alpha_{i\to j} = \alpha_{j\to i}$). We tested whether removing this symmetry constraint would improve the inference algorithm. In addition, $\alpha_{i\to j}$, $\alpha_{j\to i}$, $\beta_{i\to j}$ and $\beta_{j\to i}$ do not represent inverse overcounting numbers anymore, as these parameters weigh quantities which depends only on the pair (i, j) but not the direction.

As Figure 3.13 shows, removing the symmetry constraint indeed generally allows for better approximate inference, quite logically as the newly created algorithms are generalizations of the previous ones. However, the improvement is only marginal, and requires to learn n_{edges} additional parameters for $\boldsymbol{\alpha}$ and n_{edges} additional parameters for $\boldsymbol{\beta}$. This can lead to a slight overfitting,

visible for eFBP, which usually performs better on the test set than its generalization (model "eFBP + non-symmetric", see Figure) but not the training set (not shown). In addition to a slight overfitting with these additional parameters, the non-symmetric model does not systematically outperform its symmetric counterpart on the training set, meaning that the supervised learning algorithm has difficulties dealing with such a huge amount parameters (an explanation among others could be the low amount of training data).

Comparing extended CBP to other more complex algorithms BP, or sum-product algorithm, is the most elementary one in a family consisting not only of Generalized BP (Yedidia et al., 2001, 2005), Expectation Propagation (Minka, 2001b), Fractional BP (Wiegerinck and Heskes, 2002), which were cited previously in this thesis, but also algorithms such as Survey Propagation (Braunstein and Zecchina, 2004), Expectation Consistence (Opper and Winther, 2005) and double-loop algorithms (Heskes et al., 2002). We compared extended Circular BP to some of these complex approximate inference algorithms. More precisely, we considered two algorithms: Loop-corrected BP (Mooij and Kappen, 2007a) and Double-Loop Generalized BP (Heskes et al., 2002) thanks to the libDAI implementation of these algorithms (Mooij, 2010) (respectively "LCBP_FULLCAVin_SEQRND" and "HAK_MIN" options in libDAI). Loop-corrected BP approximates the cavity distributions for each variable in a two-step way, and Double-Loop Generalized BP uses a double-loop procedure to guarantee the convergence of the Kikuchi free energy (whose extrema correspond to the fixed points of Generalized BP).

Figure 3.14 shows that eCBP and eFBP beat the Double-Loop GBP algorithm (similar to BP in performance), but gets significantly outperformed by the Loop-corrected BP algorithm, especially for structured graphs. However, Figure 3.6 complements the score ranking with an interesting visual perspective. Although extended Circular BP gets outperformed by Loop-corrected BP with the score of Equation (3.37), it looks like Loop-corrected BP makes stronger errors than extended Circular BP (although much more rarely). Humans, on the contrary, would intuitively benefit from never making strong mistakes, for instance in order to survive, and instead making tiny mistakes (possibly rather often). Furthermore, the complexity of Loop-corrected BP (Mooij and Kappen, 2007a) makes it quite implausible for the algorithm to be implemented in the brain as such, contrary to extended Circular BP as seen in chapter 4.

Comparison with the classical rate network A growing field of theoretical neuroscience research consists of training a particular type of recurrent neural network to solve behavioral tasks mastered by animals and humans. This recurrent neural network is described by the following equation:

$$B_i = \sum_{j \in \mathcal{N}(i)} W_{ij}^{\text{rec}} \phi(B_j) + W_i^{\text{in}} M_{\text{ext} \to i} \tag{3.39}$$

(note that the input connectivity matrix is taken diagonal here). This model differs from eCBP by the absence of additional $\phi^{-1} = \text{artanh}$ non-linearity and the absence of subtraction inside function ϕ (scaled in eCBP with parameter $\boldsymbol{\alpha}$); see also Equation (4.14) and section 4.4. On top of these differences, eCBP also has constraints on the recurrent connectivity as $W_{ij}^{\text{rec}} = \kappa_i \phi(\beta_{ij} J_{ij})$, and on the input connectivity as $W_i^{\text{in}} = \kappa_i \gamma_i$.

Figure 3.15 shows that learning parameters $\boldsymbol{W}^{\text{rec}}$ and $\boldsymbol{W}^{\text{in}}$ of the rate network (overall $2n_{\text{edges}} + n_{\text{nodes}}$ weights without imposing symmetry on $\boldsymbol{W}^{\text{rec}}$) produces beliefs of worse quality than the simple model "BP + weights" which consists of fitting parameter $\boldsymbol{\beta}$ (with the artanh non-linearity and the subtraction with scaling $\boldsymbol{\alpha} = \mathbf{1}$). This model has only n_{edges} parameters.

Overall, Figure 3.15 shows that two features are important to carry out approximate inference with good enough quality. First, the artanh non-linearity in function f_{ij} (see Equation (3.28)) is

crucial: none of the approximate models ignoring the artanh function ("rate network" models and "eCBP tanh" models) significantly outperforms its counterparts with the artanh non-linearity. Second, and most importantly, Figure 3.15 shows the importance of the subtraction by the message going in the opposite direction (the "$-\alpha_{ij} M_{j\to i}$" in message update equations) to avoid overcounting information. Indeed, the approach consisting in removing the opposite message as if there were no cycles (eBP, that is, eCBP with $\boldsymbol{\alpha} = \mathbf{1}$) is significantly better than not removing anything (full eCBP, i.e., eCBP with $\boldsymbol{\alpha} = \mathbf{0}$). Furthermore, taking into account cycles to remove the appropriate amount of the opposite message (eCBP) is significantly better than ignoring cycles (eBP). Related to these observations is the fact that full Circular BP (Circular BP with $\boldsymbol{\alpha} = \mathbf{0}$) often performs worse than more naive algorithms like mean-field inference (Raju and Pitkow, 2016), confirming that the subtraction of BP/CBP is important for accurate inferences. Some neural implementations of BP (Ott and Stoop, 2006; Litvak and Ullman, 2009) actually ignore the message exclusion of BP and therefore correspond to full Circular BP; but the poor performance of this algorithm makes it unrealistic that such implementations are used by the brain.

3.5.5 Comparison between all algorithms

To complete the figures, which only illustrate specific points from the text by showing algorithms of the same class, we provide in Table 3.1 a ranked list containing algorithms of all types. The rank is defined by the performance of each algorithm for Erdos-Renyi graphs with connection probability $p = 0.6$. For each connection probability, 30 randomly weighted graphs of 9 nodes were generated, each with 100 examples (vectors M_{ext}) in the test set. More precisely, the performance is measured identically to previously, i.e., by computing the average over graphs of $-\log_{10}(\text{avg MSE})$ between the true marginals and the estimated marginals where "avg" of "avg MSE" is taken over examples.

Overall, among fitted models, the best algorithm was extended Fractional BP, followed by extended Circular BP. Both significantly outperformed BP. Loop-Corrected BP is the best algorithm overall, although much more complex than extended Fractional or Circular BP.

Figure 3.6 provides visual comparison for Erdos-Renyi graphs (including the case $p = 0.6$ considered in the above table) for most models listed in Table 3.1.

Name	Score	# parameters	Reference
Loop-corrected BP	6.38	-	Mooij and Kappen (2007a)
Extended Fractional BP	5.05	$2n_{\text{edges}} + 2n_{\text{nodes}}$	here, Yedidia et al. (2001)
Extended Circular BP	4.83	$2n_{\text{edges}} + 2n_{\text{nodes}}$	here
BP + weights	3.96	n_{edges}	-
Full eCBP	3.26	$n_{\text{edges}} + 2n_{\text{nodes}}$	-
Classical rate network	2.21	$2n_{\text{edges}} + n_{\text{nodes}}$	Wilson and Cowan (1972)
Fractional BP	2.07	n_{edges}	Wiegerinck and Heskes (2002)
Double-loop GBP	1.92	-	Heskes et al. (2002)
Circular BP	1.88	n_{edges}	Jardri and Denève (2013a)
Generalized BP	1.53	-	Yedidia et al. (2001)
BP	1.49	-	Pearl (1988)
Tree-Reweighted BP	1.22	-	Wainwright. et al. (2003)
Mean Field / Variational MP	0.97	-	Winn and Bishop (2005)

Table 3.1: **Comparison between various classes of algorithms**. Graphs considered here are Erdos-Renyi graphs with connection probability $p = 0.6$. *full* versions of algorithms means that $\boldsymbol{\alpha} = \mathbf{0}$.

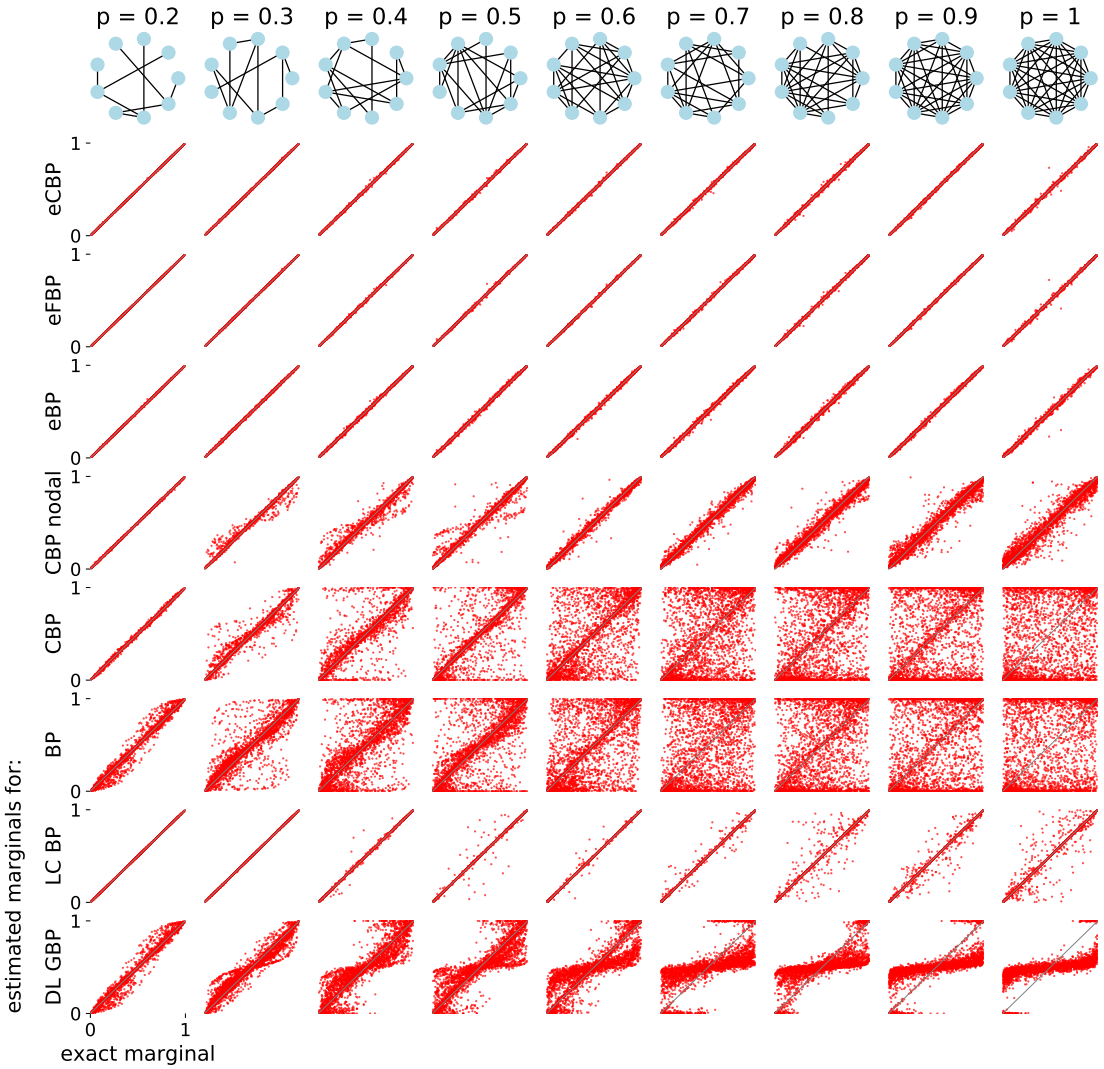

FIGURE 3.6: **Estimated marginals on the test set, for various trained algorithms**. Probability distributions are represented by randomly generated Erdos-Renyi graphs with connection probabilities ranging from 0.2 to 1. For each connection probability, 30 randomly weighted graphs are considered (one example is shown on the first line of the figure), each with 200 external input examples in the training set and 100 in the test set. Each point represents one node. For clarity, only one random node per graph is shown. In the list of such trained models, *CBP* or *FBP* refers to fitting $\boldsymbol{\alpha}$, *extended* ("e") refers to fitting $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, *nodal* refers to fitting $\boldsymbol{\kappa}$, and *weights* refers to fitting $\boldsymbol{\beta}$. BP is the most basic model, seen as a baseline. These models were also compared to more complex algorithms: Loop-Corrected BP and Double-Loop Generalized BP; see section 3.5.4. For CBP model, there is a sharp transition from carrying out excellent approximate inference for $p = 0.2$ to performing really poorly for $p = 0.4$ (at least on one of the 30 graphs) because the network becomes frustrated. eCBP and eFBP are visually the best fitting model, but are in fact outperformed by Loop-Corrected BP (LC BP) as shown below.

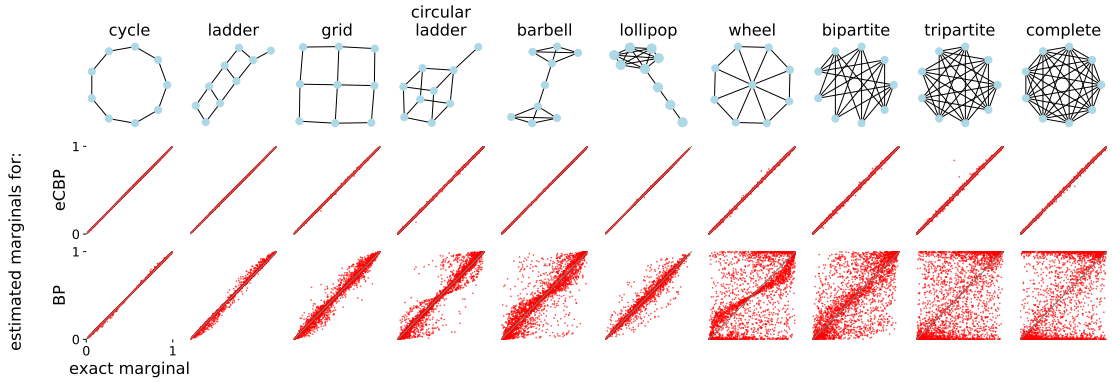

Figure 3.7: **Estimated marginals of extended Circular BP and BP for graphs of various topologies**, on the test set. For each graph topology, 30 randomly weighted graphs are considered, each with 200 external input examples in the training set and 100 in the test set. Results are qualitatively the same as in Figure 3.6 (Erdos-Renyi graphs) including for algorithms not shown here. Graph structures are the same as in Yoon et al. (2018).

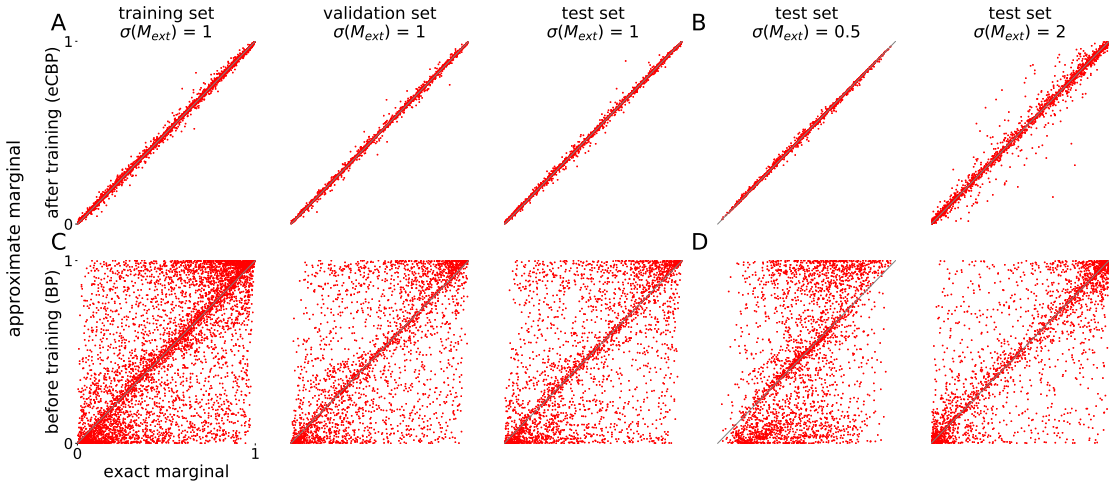

Figure 3.8: **Extended Circular BP generalizes to new data after training**. The topology considered here is the Erdos-Renyi model with $p = 0.6$ and 9 nodes, with 30 random weighted graphs generated, each with 100 test examples. **(A)** The model learns and generalizes well to the test set (within-set generalization). **(B)** Generalization to examples with different statistics (out-of-set generalization). While $\sigma(\boldsymbol{M}_{\text{ext}}) = 1$ on the training set, the model still performs well for lower and higher standard deviations of the input. The performance goes down for increased $\sigma(\boldsymbol{M}_{\text{ext}})$ as expected (the system becomes highly non-linear, and the correction brought by eCBP is linear) but only slightly. Inferences remain relatively good on such examples for instance compared to BP. **(C, D)** Same as above but for BP, which shows frustration in (at least) some of the randomly weighted graphs, that is, absence of convergence of the algorithm, in which case the beliefs have very little to do with the correct marginals.

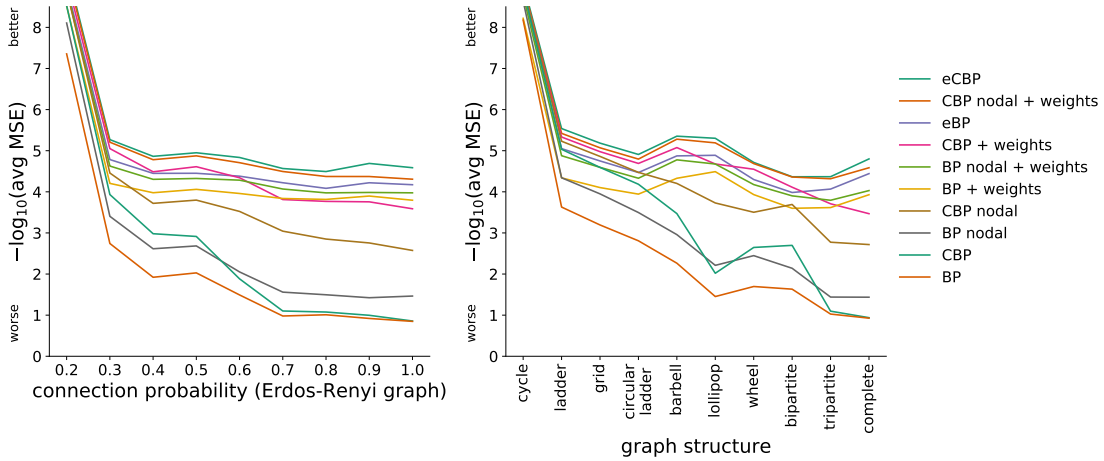

FIGURE 3.9: **From BP to extended Circular BP: additional parameters help generalize better**. Adding parameters to the algorithms from BP (no parameter) to extended Circular BP (parameters $\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}$) increasingly helps generalize better to the test set. More generally, all models which are special cases of others perform worse comparatively on the test set (with the exception that "BP weights" > "CBP + weights" for highly dense graphs, see main text). This indicates an absence of overfitting: parameters of eCBP can be learnt despite their consequent number w.r.t. the amount of training data. In the list of models, *CBP* or *FBP* refers to fitting $\boldsymbol{\alpha}$, *extended* ("e") refers to fitting $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, *nodal* refers to fitting $\boldsymbol{\kappa}$, and *weights* refers to fitting $\boldsymbol{\beta}$. Models are ordered w.r.t. their performance on Erdos-Renyi graphs with $p = 0.6$. Each point represents the log-MSE score on the test set, averaged over 30 weighted graphs (where the MSE is averaged over examples). Weighted graphs are randomly generated from a graph topology, with normally generated weights $J_{ij} \sim \mathcal{N}(0, 1)$. Each weighted graph has 100 test examples.

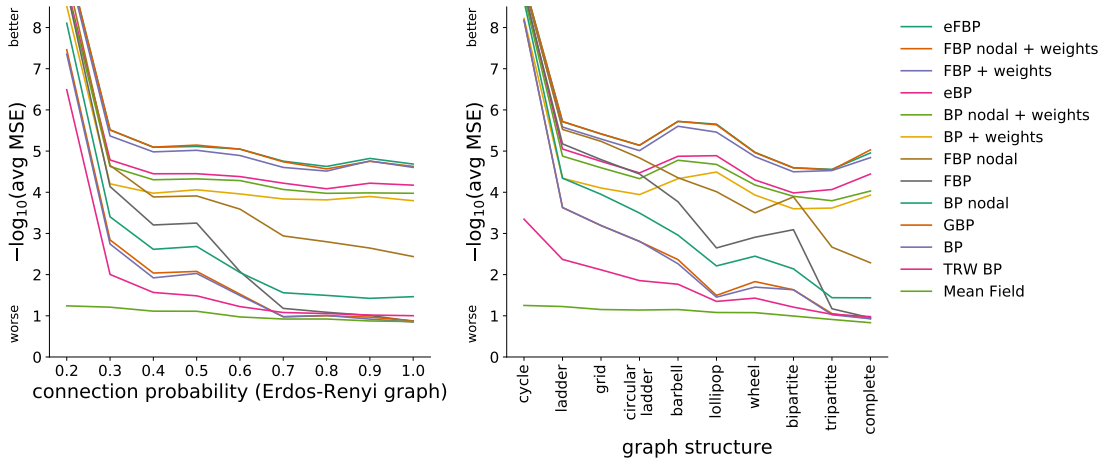

FIGURE 3.10: **From BP to eFBP: adding parameters helps generalize better**. Similarly to Figure 3.9, adding parameters to the algorithms from BP (no parameter) to extended Fractional BP (parameters $\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}$) helps generalize better to the test set.

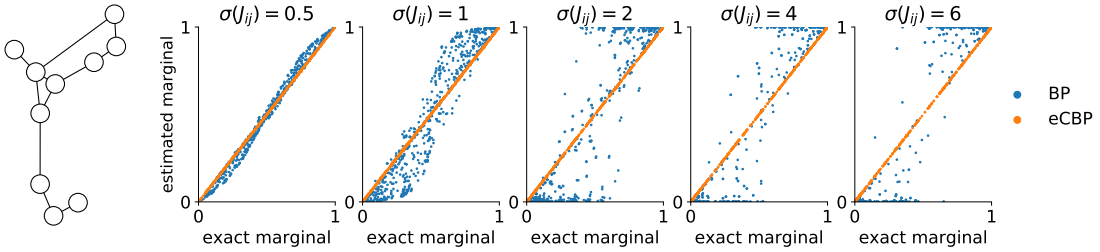

Figure 3.11: **Approximate marginals versus true marginals on the test set, for Belief Propagation and eCBP**. We consider a given graph randomly generated (shown on the left) with normally distributed weights $\boldsymbol{J} = \{J_{ij}\}$. By increasing the strength of the graph weights, BP gets worse, so as trained eCBP but at a lower rate (the decrease in performance of eCBP cannot be seen visually here but instead from the score measure defined in Equation (3.38)). In all cases, eCBP outperforms BP. BP shows here overconfidence over the graph, while eCBP is on average not overconfident nor underconfident.

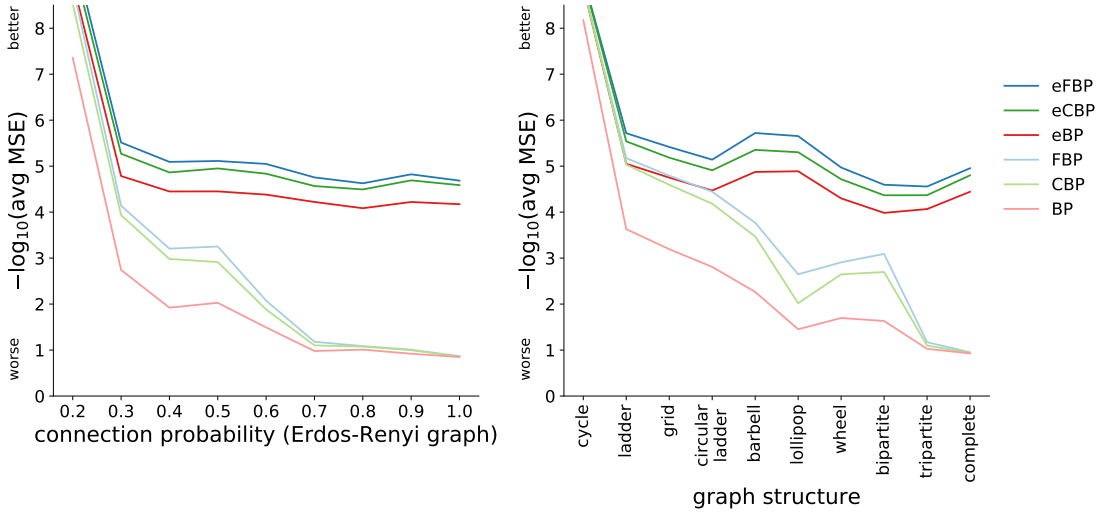

Figure 3.12: **(Extended) Circular BP is outperformed by (extended) Fractional BP but relatively close in performance**. Extended Circular and Fractional BP have the same number of parameters. Both (extended) Circular BP and (extended) Fractional BP outperform (extended) BP.

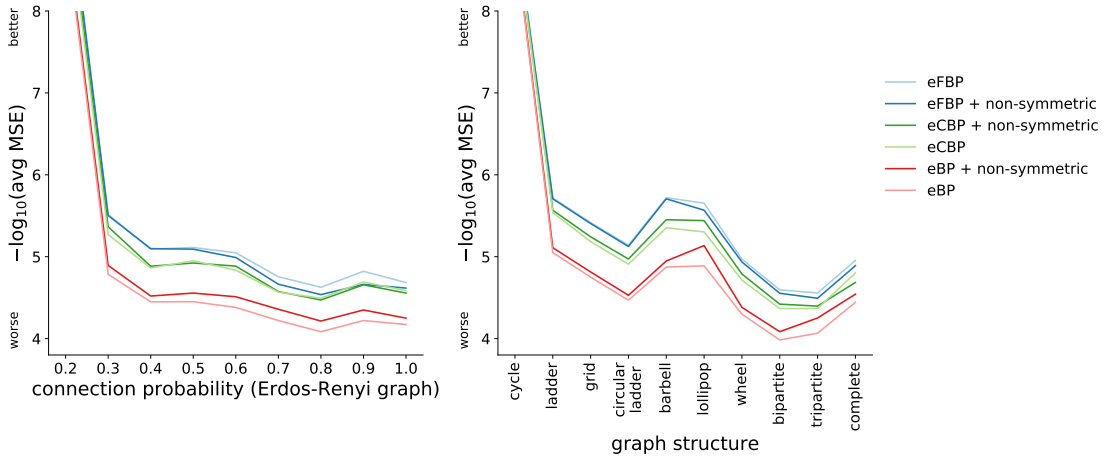

FIGURE 3.13: **Breaking the symmetry constraint in the parameter matrices α and β does not provide significant improvement to the model**, although it brings $2n_{\text{edges}}$ new parameters to fit (it even performs worse in the case of eFBP because of the high number of parameters to fit). However, it frees some constraints on the weights and synaptic scaling factors of the network, which provides stronger biological plausibility. For clarity, several points are not shown, as all models perform extremely well for trees and graphs close to trees.

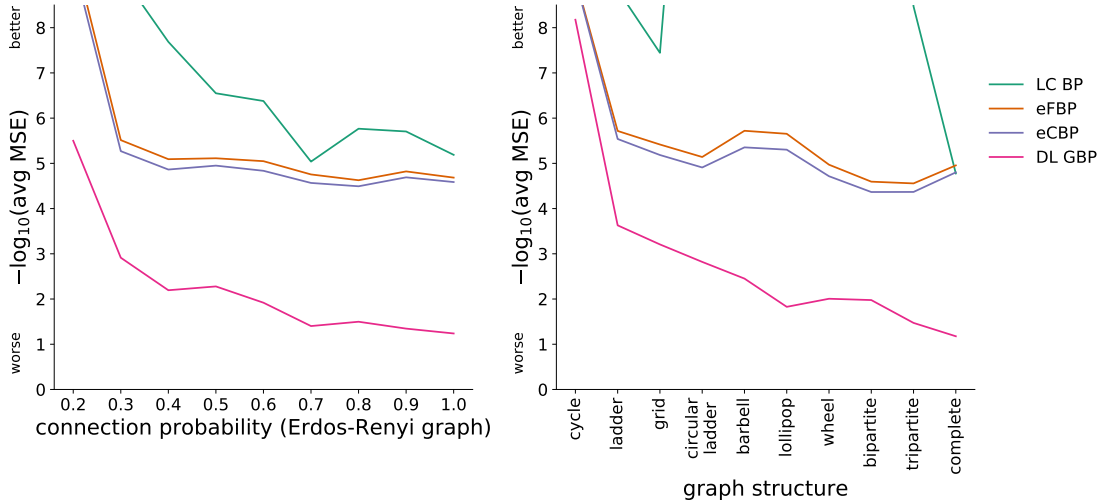

FIGURE 3.14: **Comparison between extended Circular and Fractional BP, and the more complex models Loop-corrected BP and Double-Loop Generalized BP**. Loop-corrected BP significantly outperforms eFBP and eCBP, especially for structured graphs and more generally for graphs with low density. For clarity, several points are not shown, in particular for the Loop-corrected BP algorithm.

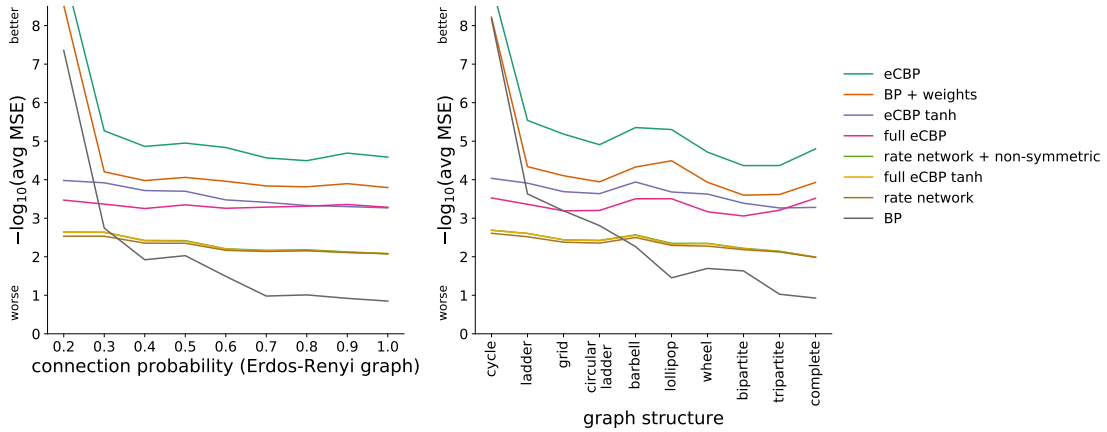

Figure 3.15: **Importance of the artanh non-linearity and of the subtraction $-\boldsymbol{\alpha_{ij}M_{j\to i}}$.** Extended Circular BP outperforms the classical rate network used in the theoretical neuroscience literature (which does not use the artanh non-linearity nor the subtraction as $\boldsymbol{\alpha} = \mathbf{0}$). More generally, none of the models ignoring the artanh non-linearity ("rate network" models and "eCBP tanh" models) perform well compared to their counterparts with artanh non-linearity - eCBP models - and even eCBP without subtraction, i.e., full eCBP. Besides, models with non-linearity artanh but non optimal level of information removal α_{ij} (eBP and full eCBP) get outperformed by eCBP which tweaks the level of subtraction α_{ij} to counter the influence of cycles. eBP (i.e., eCBP with $\boldsymbol{\alpha} = \mathbf{1}$) removes the information in the opposite direction without taking cycles into consideration, and full eCBP (i.e., eCBP with $\boldsymbol{\alpha} = \mathbf{0}$) ignores completely the removal as for mean-fied methods.

3.6 Circular BP with memory

Circular BP was not initially developed as a way of improving inference in cyclic graphs, but instead of impairing the quality of inference in acyclic graphs in which BP performs exactly. In this section, we propose an algorithm, "Circular BP with memory", similar to Circular BP, and which more closely relates to the initial idea of cancelling messages being reverberated through the network cycles.

Motivation When applying a message-passing algorithm to a graphical model containing cycles, messages will undoubtedly travel through loops, but this self-contribution should be subtracted at the moment when the message comes back. This idea was initially developed in Denève (2005) (and the related article Denève (2008)) which proposes a spiking network implementation of the Belief Propagation algorithm. This paper, in addition to considering HMMs (as discussed later in section 4.1), which is a quite restricted class of generative models, also considers a Bayesian causal network in time (coupled hidden Markov chain). In this network, BP performs only suboptimally because of the presence of cycles once unwrapping the graph through time, which causes information loops. The corresponding spiking network implementing BP therefore performs inference suboptimally as well, and the information loops sometimes lead to an explosion of the network activity (case when BP does not converge). This raises the need for appropriate amounts of *inhibitory control* to control the exchange of information and avoid positive feedback caused by cycles of the graphical model. More precisely, in the article, inhibitory control is hypothesized to take place through local inhibitory loops cancelling potentially reverberated spikes. This translates in the equation defining the evolution of the membrane potential into a self-inhibitory term, taken into account $2dt$ after a spike of the unit. This additional term is rather arbitrary, as it is not derived from normative principles, but is formulated based on the precisely defined intuition that a spike emitted at time t increases of spiking probability of receiver neurons at time $t+dt$ by a certain amount (defined by the synaptic weights), and therefore, the spiking probability of the initial neuron at time $t+2dt$. The article by Denève shows that the newly introduced inhibitory term stabilizes the spiking network and improves the quality of inference compared to BP (yet does not perform exact inference). This idea developed in Denève (2005) of having natural reverberation of information in a cyclic network was the basis for the Circular BP algorithm (Jardri and Denève, 2013a) which was designed to arbitrarily introduce such reverberations even in acyclic networks (see section 1.4.3). Circular BP artificially correlates messages going in opposite directions arbitrarily by defining one as function of the other: $M_{i\to j} = f_{ij}(B_i - \alpha_{ij} M_{j\to i}) = f_{ij}\big(\sum_{k\in\mathcal{N}(i)} M_{k\to i} + (1-\alpha_{ij}) M_{j\to i} + M_{\text{ext}\to i}\big)$ where $1-\alpha_{ij} \neq 0$ and represents the amount of circularity. Likewise, the precise modification of BP with the introduction of parameter $\boldsymbol{\alpha}$ does not come from normative principles but is strongly motivated at the intuition level by ideas of information being reverberated.

Going forward with the initial idea of Denève (2005), inference could be largely improved compared to BP if specific additional terms allowed to counter the influence of cycles of length 3, 4, 5, etc. at the appropriate time (in static graphs which are the only ones considered in this thesis, cycles of length 2 are already being exactly cancelled thanks to the subtraction in BP, see Figure 1.2). Similar ideas are developed in Raju and Pitkow (2016) and Yoon et al. (2018). Raju and Pitkow (2016) notes that "a longer memory could be used to discount past information sent at more distant times, thus avoiding the overcounting of evidence that arises from loops of length three and greater". Yoon et al. (2018) uses a multidimensional hidden state in their Graph Neural Network implementation of Bayesian inference (but not BP). This multidimensionality allows the information to be potentially retained over many timesteps (that is, not only the last timestep) and therefore acts against cycles of higher length than simply two.

Inspired by Pitkow's work on Bayesian inference Raju and Pitkow (2016); Yoon et al. (2018) and the initial idea of Denève from Denève (2005), we propose a model attempting to cancel the loops at the moment they come back. This cancellation obviously cannot be exact, but we hypothesize that the resulting model outperforms largely the initial (extended) Circular BP described in this thesis. We call this model "Circular BP with memory".

Formulation of Circular BP with memory The update equation on messages for the new algorithm is proposed as follows:

$$M_{i\to j}^{t+1} = f_{ij}\left(B_i^t - M_{j\to i}^t - \sum_{\ell} w_{i\to j,\ell} M_{j\to i}^{t-\ell}\right) \tag{3.40}$$

where f_{ij} is defined in Equation (1.20), and the sum is taken over all cycles passing through node i. Each cycle has some length ℓ and contributes to overcounting information. Note that there might be several cycles of given length ℓ, but their contributions show up as a single term in the sum. For a comparison with Circular BP, see next paragraph. The algorithm is defined here in the log-domain, but the corresponding formulation in the original domain (for any type of variable) is easy to infer , based upon the demonstration for Circular BP.

As a first approximation, we can determine the amount of subtraction needed given the graph weights:

$$w_{i\to j,\ell} = \sum_{\text{cycles}} \prod_{\text{edge}\in\text{cycle}} J_{\text{edge}} \tag{3.41}$$

where the sum is taken over minimal cycles of length ℓ passing through node i but not through the edge (i,j): see Figure 3.16 for an example. The reason for the subtracting factors in Equation (3.41) is that the message function can be linearized into $f_{ij}(x) \approx J_{ij} \times x$.

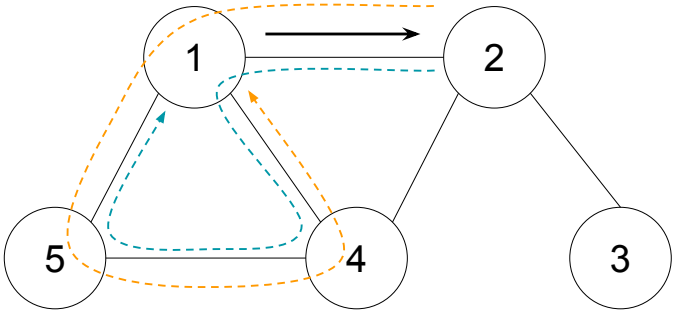

Figure 3.16: **Circular Belief Propagation with memory**. Removal of the opposite message in Circular BP $M_{j\to i}$ has to be carried out at the moment the message comes back to node i after travelling in a cycle. In this graph, there is only cycle passing through node $i = 1$ and not passing through the edge $(i,j) = (1,2)$: it is composed of nodes $1-5-4$ and has length 3. Because of this cycle, message $M_{2\to 1}$ corrupts messages $M_{5\to 1}$ and $M_{4\to 1}$ 3 timesteps later (see the propagation of $M_{2\to 1}$ in colored dotted lines). Therefore, to update $M_{1\to 2}$ (full black line), terms in $M_{2\to 1}$ need to be removed from $M_{5\to 1}$ and $M_{4\to 1}$ with a delay of 3 timesteps. We propose a linear removal: $M_{1\to 2}^{t+1} = f_{12}(B_1^t - M_{2\to 1}^t - w_{1\to 2,3} M_{2\to 1}^{t-3})$ where by definition $B_1^t - M_{2\to 1}^t = M_{5\to 1}^t + M_{4\to 1}^t$. We believe that removing the redundant information at the appropriate time (contrary to Circular BP) improves further the quality of inference.

Relation with Circular BP As a reminder, the Circular BP algorithm is defined by the following equations:

$$\begin{cases} M_{i\to j}^{t+1} = f_{ij}(B_i^t - \alpha_{ij} M_{j\to i}^t) & \text{(3.42a)} \\ B_i^t = \sum_{j\in\mathcal{N}(i)} M_{j\to i}^t + M_{\text{ext}\to i}^t & \text{(3.42b)} \end{cases}$$

The update equation on messages for Circular BP is exactly the same as for Circular BP with memory, provided we neglect the propagation time inside cycles of the graph. The update equation for CBP with memory (3.40) indeed becomes in this case:

$$M_{i\to j}^{t+1} = f_{ij}\left(B_i^t - M_{j\to i}^t - \sum_\ell w_{i\to j,\ell} M_{j\to i}^t\right)$$

which also provides a first approximation to loop correction factors $\{\alpha_{ij}\}$ in Circular BP, as $\alpha_{ij} = 1 + \sum_\ell w_{i\to j,\ell}$:

$$\alpha_{ij} = 1 + \sum_\ell \sum_{\text{cycles}} \prod_{\text{edge}\in\text{cycle}} J_{\text{edge}} \tag{3.43}$$

where cycles have length ℓ and should contain node i but not edge (i,j). This can be used as a starting point for a more subtle learning method (unsupervised learning or supervised learning, described previously in this thesis for Circular BP). See also section 3.1 for similar ideas.

Overall, we reconciliated the vision of Denève (2005) with the Circular BP algorithm as formulated in Jardri and Denève (2013a). It turns out that Circular BP, where the loop-correction parameter α intervenes at time $t-1$, makes sense provided we neglect the propagation time inside cycles. However, the newly formulated algorithm (Circular BP with memory) corresponds more to the idea of cancelling reverberating terms arising from loops of the network, even though Circular BP has the same flavor.

Parameters $\{w_{i\to j,l}\}$ of the new algorithm might be hard to fit because there are many more than in Circular BP (which only has one parameter α_{ij} per edge, contrary to the potentially linear number of terms in the number of nodes for Circular BP with memory), and because these parameters are very much interrelated: for a given oriented edge $i \to j$, the $\{w_{i\to j,l}\}$ have similar effects on the overcounting or undercounting by the graph. However, a good starting point would be to define these parameters using Equation (3.41), which most surely performs inference relatively well.

The new algorithm does not perform exact inference, because it does not cancel loops exactly but only in a linear manner. For a graph correponding to a single cycle composed of all its nodes, Circular BP with memory does not differ from BP and is thus suboptimal (although BP for single-cycle graphs is known to give the MAP solution and performs relatively well (Weiss, 2000)). Moreover, the new algorithm does not relate simply to Fractional BP like Circular BP did. It would be interesting to try and relate Circular BP with memory to Fractional BP and/or the Tree-Reweighted BP algorithm.

There are potentially other BP-inspired models which could come from the same idea of memory, and the update formula provided in Equation (3.40) is simply one of them. Future work will need to test these different propositions and propose plausible neural implementations, as chapter 3 did for Circular BP. More precisely, "testing" means fitting the model and compare its performance to Circular BP (which is different, that is, is not a generalization of Circular BP with memory nor a special case of it, but has less parameters) or extended Circular BP (which is different and has more parameters), among others.

A final note is that the idea of adding memory does not translate to Fractional BP, as Fractional BP resembles Equation (3.42a) for Circular BP, but the update function g_{ij} (defined in Equation 3.6) depends on α_{ij}, contrary to f_{ij} .

3.7 Conclusion: Circular BP can improve the quality of inference

The Belief Propagation algorithm, which approximates marginals of a given probability distribution, often struggles with probability distributions containing cyclic conditional dependencies, that is, whose associated probabilistic graphical model has cycles. In order to tackle this problem, several variants or generalization of BP have been proposed, ever since the interest in BP started to grow (Minka, 2001b; Sudderth et al., 2003; Ihler and McAllester, 2009).

In this chapter, we proposed an approximate inference method, the *extended Circular Belief Propagation* algorithm, defined for general factor graphs. This method is a very light modification of the Fractional Belief Propagation algorithm, which is a particular Generalized Belief Propagation algorithm (Yedidia et al., 2001). This provides a theoretical foundation to (extended) Circular BP, which is not a normative model per se but relates very closely to (extended) Fractional BP, which is normatively motivated.

We show in simulations, by considering particular probability distributions (Ising models) that this extended CBP algorithm outperforms previously proposed approaches improving BP such as Fractional BP, Power EP, α-BP as well as BP itself. The extended Circular BP algorithm is slightly outperformed by the extended Fractional BP algorithm, but the latter cannot be as easily implemented with rate units (see chapter 4) and as a consequence, lacks biological plausibility.

While the level of complexity of extended CBP stays comparable to the one of BP and its variants, we gain two useful features. First, the guarantee of stability (or rather, of being able to find parameters for which the system is stable), which has not been shown in these other models and according to simulations might not be true. Second, the higher accuracy in inference when all algorithms converge, that is, produced approximate marginals. Many more approximate inference methods exist that have a much higher complexity than our proposed approach and probably are better performing, in particular deep-learning based methods such as Graph Neural Networks (Yoon et al., 2018) or Belief Propagation Neural Networks (Kuck et al., 2020). It remains to be seen how extended Circular BP compares to these methods in terms of performance.

This work paves the way for future research on approximate inference. Equation (3.24) (or Equation (3.26a) in the log-domain) gives a general recipe to use rescaled correction in many algorithms. This "corrective multiplicative factor trick" can be used on top of algorithms which already improve BP while conserving the same form, like Kuck et al. (2020) or the message-GNN from Yoon et al. (2018), potentially leading to further improvement of these methods.

One limitation of the supervised learning procedure is the need to generate exact marginals which prevents from learning the model on large graphs. An unsupervised learning method to learn the corrective multiplicative factors, presented in section 4.5, brings a solution to this issue.

Another limitation of the approach is that, contrary to some deep-learning based methods as Graph Neural Networks (Yoon et al., 2018), extended CBP is trained to carry out approximate inference on any external inputs for each graph independently. Learning a mapping to be able to predict the corrective factors from the graph structure and weights could be an interesting future piece of work.

Overall, we showed, using numerical experiments, that it is possible to improve substantially the quality of inference carried out by BP by using (anti-)Circular Inference. This resolves the initial following conundrum: Circular BP models psychosis through disturbances of inference in humans, but its baseline (standard BP), which is supposed to model normal behavior, is not

carrying out inference with a high quality. This chapter shows that Circular BP is not only a good model of psychosis, but of normal behavior as well. Interestingly, Circular BP was initially thought to introduce suboptimality (to BP) in graphs without cycles. We use it instead to correct for (BP) suboptimalities naturally existing in cyclic graphs. Finally, we introduce the *Circular BP with memory* algorithm in the spirit of Denève (2005) that reverberated messages should be considered (here, removed from the signal they corrupt) only once they have travelled through cycles of the graph.

Chapter 4

Circular Belief Propagation and its neural implementation

Summary of Chapter 4

We showed in previous chapters that (extended) Circular Belief Propagation, an algorithm performing approximate inference on probabilistic graphs, is potentially a good model of suboptimal and optimal behavior. However, all models proposed remained at the algorithmic level, without providing a proposition of how the brain could implement this algorithm. In this chapter, we address this issue by presenting a possible neural implementation of the algorithm in the case of probability distributions over binary variables.

More precisely, we propose a rate network implementation as well as a more biologically-plausible spiking network implementation of the Circular BP algorithm. In both proposed implementations, "message" populations and "marginal" populations encode for the parameters of the probability distribution (in the binary case presented here, the log-odds, which is a transform of the mean), while "prediction error" populations encode for the information that a message population does not provide about the corresponding marginal population (not to be confused with prediction error units from Predictive Coding theories of Karl Friston and colleagues). Relationships between variables of the probability distribution are encoded in the synaptic weights. Lastly, parameters of extended Circular BP have a very clear meaning in the proposed implementation. α and κ represent the synaptic gains of the network units (respectively of the message population and of the marginal population), β scales the synaptic weights of the network, and γ scales its input weights. In the rate network implementation, parameters of the probability distribution (log-odds) are directly encoded in firing rates. On the contrary, in the spiking network implementation, the same parameters are encoded only implicitly, as they are obtained by integrating the output spike trains of neurons from the corresponding population. The instantaneous firing rate of a spiking neuron is proportional to the current *new* piece of sensory evidence (see also Denève (2008)).

Finally, we present a biologically plausible unsupervised learning rule to learn parameters of extended Circular BP, therefore providing an alternative to the supervised learning rule of chapter 3 which could take place in the brain.

4.1 Introduction

Chapter 2 examined how Circular BP could be used to model suboptimal behavior such as the bistable perception phenomenon, and how Circular BP looks compatible with particular disturbances of the functional connectivity in schizophrenia. Chapter 4 of this thesis explored the possible neural implementations of the Circular Belief Propagation algorithm, considering various possibilities such as rate networks or spiking networks in the binary case, and rate networks in the Gaussian case.

In this chapter, we focus on possible neural implementations of the extended Circular Belief Propagation algorithm, an approximate Bayesian inference algorithm. This is a question of the highest importance, because it could ultimately lead to testable predictions relative to neurobiological data. The confrontation between such predictions and experimental data could help understand whether Circular BP is the way the brain performs probabilistic inference, or instead whether the model should be modified or abandoned.

We start by reviewing existing work on this issue. While section 1.3 discussed different general theories of how Bayesian inference could be carried out in the brain, we focus here exclusively on implementations of the Belief Propagation algorithm (or sum-product algorithm) as well as existing work on implementations of the Circular Belief Propagation algorithm itself.

The reason why the implementations of BP are of interest for the potential ones of Circular BP is obvious: Circular BP equations highly resemble and extend the ones of BP, as seen in the Introduction chapter. A specificity of the BP algorithm is that it does not have a particularly simple form in its initial formulation. Indeed, its update equation (Equation (1.3)) involves summation and most importantly, multiplication and division. This latest operation cannot reasonably be implemented as such in the brain. A usual bypass to this problem, which is the one considered in this thesis, is to consider BP in the log domain (or log-odds domain), in which the multiplication operation becomes a summation operation and the division becomes a subtraction, as shown in section 1.4.4. Interestingly, (extended) Circular Belief Propagation also takes a very simple form in the log-odds domain: the additional parameters $\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, which are initially involved in multiplications as exponents, logically end up as multiplicative factors in the log-domain. This thus does not bring more complexity to the initial BP algorithm, which allows for propositions of neural implementations of Circular BP based on the existing ones for BP in the log-domain.

BP on Hidden Markov Models The Belief Propagation algorithm can be applied to graphical models evolving over time. This was historically the problem considered in the first propositions of neural implementations of BP (Rao, 2004; Denève, 2005, 2008; Beck and Pouget, 2007). More precisely, these papers consider a particular model: the hidden Markov model (HMM), in which a latent variable or hidden state evolves randomly at each time step with a given transition probability, and a noisy observation of the hidden state gets generated at each point in time. HMMs are a model for cue integration, decision-making, and motor control, for instance. Rao (2004) proposes a rate model in which the firing rate of a neuron is proportional to the log-probability of the event that it encodes. However, this model requires approximations, for instance of a log-of-sums with a sum-of-logs or alternatively use non-linear dendrites, and therefore performs BP only approximately. The rate model can be replaced with a stochastic spiking network, in which the membrane potential of the neuron is the log-probability and a neuron spikes stochastically with a probability (or instantaneous firing rate) scaling with the encoded probability itself. However, this spiking network has the same inconvenient: it carries out approximate BP. Very similarly, Denève (2005, 2008) propose a spiking model implementing BP exactly - up to the decoding weights of the neurons' spikes - on a HMM (similarly, Boerlin and Denève (2011)

models the integration of a dynamic stimulus over time in a drift-diffusion process instead of a HMM). However, contrary to Rao (2004), the network implements the evolution of the log-odds rather than log-probability. Furthermore, the instantaneous firing rates of neurons are proportional to the strength of the *new* piece of evidence (to be integrated in time) and only encode probabilities implicitly. Finally, Beck and Pouget (2007) proposes a quadratic non-linear rate model generalizing Denève (2005) (in particular, to cases where the encoded variable is discrete rather than binary), as well as Rao (2004). Beck and Pouget (2007) also defends the idea that the evolution of the likelihood itself can be mapped directly onto neural circuits rather than the log-probability or the log-odds. It is to be noted that online inference with BP on HMMs admits as special case the forward algorithm or Kalman Filter, for which neural implementations have been proposed as well: Denève et al. (2007) uses recurrent basis function networks and relies on population codes, while Wilson and Finkel (2009) uses a modification of a line attractor and is only approximate.

BP on other acyclic graphical models and the need for message exclusion Applying BP to graphical models like HMMs is very restrictive. In fact, as pointed out by Raju and Pitkow (2016), in such models, the propagation of BP messages is unidirectional. For instance, in HMMs, a hidden variable receives messages about its past state and from the current noisy observation of itself. This therefore ignores the main difficulty of implementing BP in any other graphical model, in which messages are exchanged bidirectionally. In such a case, BP avoids messages from being reverberated $i \to j \to i$, a positive feedback phenomenon, which is responsible for overcounting. To do so, BP, when computing a message sent by node i to node j, considers all the evidence available at node i except the message coming from node j. This message exclusion shows in BP message update equation with $\mathcal{N}(i) \setminus \boldsymbol{j}$, which can equivalently be seen in the subtraction in the log-odd domain $M_{i\to j} = f_{ij}(B_i - \boldsymbol{M_{j\to i}})$. This message exclusion is simply ignored in some models proposing a neural implementation of BP (Ott and Stoop, 2006; Litvak and Ullman, 2009; Yu et al., 2017; Zheng et al., 2020) making the inference too poor to reasonably be implemented by our brain circuits. Indeed, this approximate inference algorithm, which exactly corresponds to full Circular BP (that is, Circular BP with $\boldsymbol{\alpha} = \mathbf{0}$), often performs worse than more naive algorithms like mean-field inference (see Raju and Pitkow (2016)). Finally, Raju and Pitkow (2016) elegantly avoids the message exclusion problem by using the tree-based reparametrization of the BP algorithm (Wainwright. et al., 2002), formulated for the exponential family of distributions, and proposes a neural implementation of the resulting equation using a rate network with probabilistic population coding.

A mapping between the cortical architecture and graphical models? Approximately at the same time when neural implementations of BP were proposed on HMM, researchers considered another generative model: the chain of variables, dependent on each other (static model). The reason for that is that the hierarchical organization of the cortex which might reflect a generative model mirroring this hierarchy (vertical chain). An example is the visual system, in which edges and lines are encoded by neurons of the primary visual cortex or V1, shapes in V2, objects in V4, up to faces in the inferior temporal cortex or IT Felleman and Van Essen (1991). In the example of the hierarchical probabilistic graph with 3 nodes encoding for the presence or absence of "leaf", "tree", and "forest", it means that the hierarchical graphical model directly translates into the cortical hierarchy and in particular, that the neural unit encoding for the concept of leaf is lower in the cortical hierarchy than the neural unit encoding for the concept of tree. Neurons or populations of neurons are nodes of the generative model, and neural connections are edges of the generative model (conditional dependencies in a Bayesian network, factor in a factor graph). Shon and Rao (2005) proposes that relationships between variables (encoded in

edges of the graphical model) are encoded through small groups of synaptic connections in a log-probability rate model where messages in BP symbolize a time-average of the log current passed between neurons. Lee and Mumford (2003) suggests BP might model the interactive feed-forward and feedback cortical computations.

BP in more complex probabilistic graphs In all the examples considered above, the probability distribution can be described by at most pairwise relationships between variables. However, George and Hawkins (2009); Steimer et al. (2009); Steimer and Douglas (2013) all propose spike-based implementations of BP applied in more complicated graphical models than most other approaches including the present thesis, which consider only pairwise interactions between variables (see section 1.2.2). Steimer et al. (2009); Steimer and Douglas (2013) even consider general Forney factor graphs (Forney, 2001), a variant of factor graphs. In this case of complex graphical models, an extra neuronal pool is required to compute the messages and the neural structure does not exactly mirror anymore the graphical model.

Previous propositions of Circular BP implementations Initial papers from Jardri and Denève (Jardri and Denève, 2013a,b; Jardri et al., 2016) describing the Circular Belief Propagation algorithm, and most importantly its potential to explain hallucinations and the formation of delusional beliefs, made no precise proposition on how the algorithm was implemented in the brain. These articles simply suggested that variables x_i of the probability distribution were encoded in different units and that the brain structure mirrored the (hierarchical) generative model of the world. It is only later, in Leptourgos et al. (2017), that more detailed neural-like implementation ideas were provided, yet without a true correspondence between the algorithm and the circuit hypothesized to implement it. The first proposition of neural circuit uses local connections between pyramidal cells in a given cortical area and their corresponding pool of inhibitory interneurons. The alternative implementation uses long-range inhibitory connections, possibly involving thalamocortical or corticostriatal connections, in the same spirit as the local inhibitory feedback of Denève (2005). Finally, Leptourgos et al. (2021) proposes an even more detailed way of implementing Circular BP in a microcircuit where inhibitory control is implemented by layer-specific inhibition, allowing the information to propagate in the cortical hierarchy. However, the big downside of this work is that a modification of the Circular BP algorithm is used, not Circular BP itself, in a form of a proxy which requires fitting the non-linear synaptic functions (see also Rao (2005b,a)).

Goal of the chapter In this chapter, we provide more biological plausibility to the extended Circular BP algorithm. We propose an implementation of eCBP in continuous time (and approximations of it in the last section of the chapter). We tackle specifically the problem of Bayesian inference on a *static* graphical model, that is, a given probability distribution $p(x)$ with constant evidence (which can be seen as prior or sensory evidence) on variables x_i.[1] We assume here log-probability coding (which defines the representation of probabilities) together with the eCBP algorithm (which defines how probabilities are updated) in our proposed implementation of probabilistic inference (see section 1.3). This hypotheses lead straightforwardly to a rate model implementing the algorithm. Additionally, we propose an implementation with spiking networks, whose membrane potential encodes for the difference of the estimate probability and a prediction based on its own spikes. Overall, we provide more detail about how and where the inhibitory control (subtraction in eCBP) could be implemented and propose two solutions: using a particular "prediction error" unit or alternatively carrying out the operation at the dendrite of a unit,

[1] Although we will see that the model can also be used for time-varying inputs

a debate which reminds the ones on the neural implementations of predictive coding (Spratling, 2016). Finally, we propose in section 4.5 a biologically-plausible unsupervised learning rule for parameters of extended Circular BP so that the network learns without supervision (contrary to section 3.5) how to perform approximate probabilistic inference rather accurately.

4.2 Implementation of Circular BP with rate units

Here we show how the (extended) Circular BP algorithm on binary graphical models can be directly implemented by a rate network in continuous time, given two hypotheses.

The first hypothesis is that marginal probabilities $p_i(x_i)$ are encoded by different units in the network. In other words, distinct concepts are encoded by distinct neurons or populations of neurons.

The second hypothesis is that units encode the log-odds of the distribution (see also Gold and Shadlen (2001); Shon and Rao (2005); Denève (2005, 2008)), an hypothesis which is backed by neurophysiological findings suggesting that neural spike rates perform computations in the log domain (Carpenter and Williams, 1995; Shadlen and Newsome, 2001). Moreover, this hypothesis is in line with previous proposals assuming that the neural activity represents the parameters of the probability distribution (see Ma et al. (2006); Fiser et al. (2010)) and more specifically that firing rates represent log-probabilities (Pouget et al., 2013), as in probabilistic population codes (PPC) (Zemel et al., 1998; Denève et al., 2001; Ma et al., 2006). It is also a convenient hypothesis as Circular BP takes a simple form in the log-domain as shown in section 1.4, with operations (summation and subtraction) which can take place directly in neural circuits, contrary to the multiplication operation appearing in the initial formulation of Circular BP.

4.2.1 Circular BP in continuous time

Notion of damping Let a discrete system:

$$x_{t+1} = f(x_t)$$

A common technique to improve convergence of the system is to take partial (or damped) update steps:

$$x_{t+1} = (1-\epsilon)f(x_t) + \epsilon x_t$$

with $\epsilon \in [0;1[$. This procedure, called *damping*, does not modify the fixed points of the system: a fixed point of the original system is a fixed point of the damped system, and reciprocally.

A parallel can be drawn between the damped discrete system and the following continuous system:

$$\tau \dot{x}(t) = -x(t) + f(x(t)) \tag{4.1}$$

Indeed, this continuous system can be discretized (Euler approximation) into:

$$x_{t+\delta t} = (1-\epsilon)f(x_t) + \epsilon x_t \tag{4.2}$$

with $\epsilon = 1 - \delta t/\tau$ (i.e., $\tau = \delta t/(1-\epsilon)$). This corresponds to the discrete damped system.

Damped Circular BP We start by rewriting the message update equation for extended Circular BP (Equation (3.24)) without damping:

$$m^{\text{new}}_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i,x_j)^{\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i}\right)^{\kappa_i} \tag{4.3}$$

and its equivalent in the log-domain for probability distributions with binary variables (Equation (3.26a)):

$$M_{i\to j}^{\text{new}} = f_{ij}(B_i - \alpha_{ij} M_{j\to i}) \tag{4.4}$$

where $M_{i\to j} \equiv \frac{1}{2}\log\left(\frac{m_{i\to j}(x_j=+1)}{m_{i\to j}(x_j=-1)}\right)$, $M_{\text{ext}\to i} \equiv \frac{1}{2}\log\left(\frac{\psi_i(x_i=+1)}{\psi_i(x_i=-1)}\right)$, $B_i \equiv \frac{1}{2}\log\left(\frac{b_i(x_i=+1)}{b_i(x_i=-1)}\right) = \kappa_i\Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big)$, and function f_{ij} is given in Equation (3.27).

Damping is a technique commonly used while running Belief Propagation (special case $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$), and consists of taking partial message update steps in the log-space: $M_{i\to j}^{\text{new}} = f_{ij}(B_i - M_{j\to i})$ becomes $M_{i\to j}^{\text{new}} = (1-\epsilon) f_{ij}(B_i - M_{j\to i}) + \epsilon M_{i\to j}$. Damping improves the convergence properties of BP. We talk about *damped Belief Propagation* in this case (Murphy et al., 1999).

Similarly to damped BP, a damped extended Circular BP algorithm can be defined, by taking partial message update steps in the log-space. The message update equation for the damped algorithm is:

$$m_{i\to j}^{\text{new}}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i}\right)^{\kappa_i}\right)^{1-\epsilon_{i\to j}} \times m_{i\to j}(x_j)^{\epsilon_{i\to j}} \tag{4.5}$$

where $\epsilon_{i\to j}$ is the damping factor associated to the oriented edge $i \to j$ ($\boldsymbol{\epsilon} = \mathbf{0}$ means no damping, i.e., standard eCBP) and is often taken uniformly over the edges. It is equivalent in the log-domain to the following equation:

$$M_{i\to j}^{\text{new}} = (1-\epsilon_{i\to j}) f_{ij}(B_i - \alpha_{ij} M_{j\to i}) + \epsilon_{i\to j} M_{i\to j} \tag{4.6}$$

where

$$B_i = \kappa_i\Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big) \tag{4.7}$$

As in the previous section, we draw a parallel with the continuous-time equation:

$$\tau_{i\to j}\dot{M}_{i\to j}(t) = -M_{i\to j}(t) + f_{ij}\big(B_i(t) - \alpha_{ij} M_{j\to i}(t)\big) \tag{4.8}$$

which has identical fixed points to the discrete system. The first contribution on the r.h.s. is a leak term, which ensures that the log-message $M_{i\to j}$ decays back to the zero baseline (for which beliefs are uniform: $b_i(x_i=+1) = b_i(x_i=-1) = 0.5$) in absence of external inputs $M_{\text{ext}\to i}$. Equations (4.7) and (4.8) together define a continuous-time Circular BP, more specifically, continuous-time Circular BP in the binary case and for at most pairwise factors.

4.2.2 A rate network implementing Circular BP

Here we hypothesize that $\tau_{i\to j} = \tau_j$, i.e., that $\tau_{i\to j}$ does not depend on i. Continuous-time Circular BP from Equations (4.7) and (4.8) can be rewritten into the following system:

$$\begin{cases} \tau_j \dot{M}_{i\to j} = -M_{i\to j} + f_{ij}(B_i - \alpha_{ij} M_{j\to i}) & \text{(4.9a)} \\ \tau_i \dot{B}_i = -B_i + \kappa_i \sum_{j\in\mathcal{N}(i)} f_{ji}(B_j - \alpha_{ij} M_{i\to j}) + \kappa_i \gamma_i I_{\text{ext}\to i} & \text{(4.9b)} \end{cases}$$

where $I_{\text{ext}\to i} \equiv \tau_i \dot{M}_{\text{ext}\to i} + M_{\text{ext}\to i}$: external messages are allowed to vary with time (even though the generative model is static).

Interestingly, by defining

$$B_j^i = \alpha_{ij} M_{i \to j} \tag{4.10}$$

the prediction of variable x_j by variable x_i (with correction factor α_{ij}), we obtain:

$$\begin{cases} \tau_j \dot{B}_j^i = -B_j^i + \alpha_{ij} f_{ij}(B_i - B_i^j) & \text{(4.11a)} \\ \tau_i \dot{B}_i = -B_i + \kappa_i \sum_{j \in \mathcal{N}(i)} f_{ji}(B_j - B_j^i) + \kappa_i \gamma_i I_{\text{ext} \to i} & \text{(4.11b)} \end{cases}$$

We map directly the above system of equations, which defines extended Circular BP in continuous time, into a rate network with two types of nodes shown in Figure 4.1B, which implements extended Circular BP exactly. The first unit type is the *projection units*, encoding for $B_i = \frac{1}{2} \log \left(\frac{b_i(x_i=+1)}{b_i(x_i=-1)} \right)$ (transformation of the approximate marginal probability). The second unit type is the *control units*, encoding for B_j^i, the *rescaled* prediction of B_j (i.e., of variable x_j) by the local information at variable x_i. Units are connected through non-linear saturating synapses (transfer function f_{ij}) due to receptor saturation or finite reversal potentials (Dayan and Abbott, 2001). Common transfer functions used in modeling work are often the sigmoid function or the closely related hyperbolic tangent function (tanh); f_{ij} is also sigmoidal, as reminded below.

The external input $I_{\text{ext} \to i} = \tau_i \dot{M}_{\text{ext} \to i} + M_{\text{ext} \to i}$ is only received by the projection units, and can vary with time. $M_{\text{ext} \to i} \equiv \frac{1}{2} \log \left(\frac{\psi_i(x_i=+1)}{\psi_i(x_i=-1)} \right)$ is constant in a static environment, as when the goal is to compute marginals of a given probability distribution (which defines constant factors $\{\psi_{ij}\}$ and $\{\psi_i\}$). But $M_{\text{ext} \to i}$ is not constant in a dynamic environment as it is the case in the brain receiving time-varying external (sensory) inputs.

Parameters $\boldsymbol{\alpha}, \boldsymbol{\kappa}$, and $\boldsymbol{\gamma}$ represent very concrete quantities in the network. Parameter $\boldsymbol{\kappa}$ represents the synaptic scaling factor of the projection units: κ_i is associated to the projection unit encoding for B_i. Similarly, $\boldsymbol{\alpha}$ represent the synaptic scaling factors of the control unit: α_{ij} is associated to the control unit encoding for B_i^j (or B_i^j, as $\alpha_{i \to j} = \alpha_{j \to i} \equiv \alpha_{ij}$). Finally, γ_i is the weight of the external input (before the scaling by κ_i). The symmetry constraint on $\boldsymbol{\alpha}$ is equivalent to having different units (encoding for B_i^j and B_j^i) with exactly identical synaptic scaling factors. Similarly, the symmetry constraint on $\boldsymbol{\beta}$ is equivalent to having different anatomical connections with identical weights (as $\beta_{i \to j} J_{ij}$ represents the weight of the non-linear synapse through f_{ij}). Undeniably, these symmetry constraints hinder the hypothesis that extended Circular BP is implemented as such in the brain, which cannot force biological components to be exactly identical. In particular, it is implausible that a change in the synaptic scaling factor of a neuron (for instance, of the unit encoding for B_i^j) reflects directly the same exact change on the one of another neuron (unit encoding for B_j^i). Therefore, we relax the symmetry constraint, motivated by biological arguments, in addition to the fact that symmetry breaking did not hinder the supervised learning method as described in cite section 3.5.4.

Equation (4.11b) means that projection unit i receives from projection unit j everything that j knows (B_j) minus what i already knows about j (B_i^j). The subtraction can be seen as (inhibitory) control in neural terms. Note that the exact same subtraction appears in Equation (4.11b), and initially comes from ignoring the message in the opposite direction in the Belief Propagation algorithm: for instance, in the message update equation (Equation (4.3)), the product appearing in the expression of the updated message is incomplete: "$k \in \mathcal{N}(i) \backslash j$" means that the message from node i to node j is a function of all messages coming to i (from neighbors of i in the graph), except from the one coming from j.

Importantly, we do not talk about excitatory units and inhibitory units, but instead about projection units and control units. Projection units represent the true signal that we want to encode (beliefs), sending signal from other areas, whereas control units regulate the exchange of

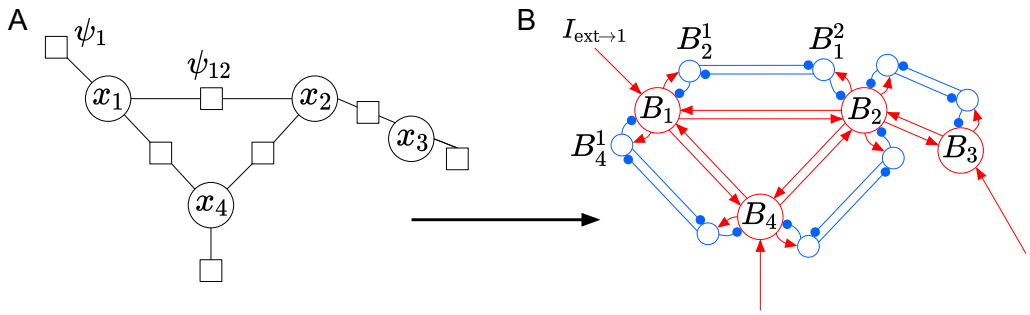

Figure 4.1: **From a probability distribution to a rate network implementing Circular BP. (A)** The probability distribution $p(\mathbf{x})$ is represented by a factor graph, with pairwise potentials ψ_{ij} and unitary potentials ψ_i. **(B)** Rate network with two types of units, implementing Circular Belief Propagation (see System (4.11)). Connection weights depend on ψ_{ij}, while external input depend on ψ_i. Projection units in red encode the approximate marginal probability $b_i(x_i) \approx p_i(x_i)$ (or rather, the half log-likelihood ratio or half log-odds $B = \log(p_i(x_i = +1)/p_i(x_i = -1))$) and receive external inputs, while control units in blue remove the information being reverberated between projection units. Colors of connections (or equivalently, shape of arrows) determines the influence of the sending unit onto the receiving unit: red indicates a positive effect and blue indicates a negative effect; see Equations (4.11a) and (4.11b).

information between projection units. Indeed, f_{ij} is not necessarily an increasing function of its argument $B_i - B_i^j$. For instance, in the case of the Ising model where $f_{ij}(x) = \phi^{-1}(\phi(\beta_{ij}J_{ij})\phi(x))$, $f_{ij}(x)$ has the same sign and evolves as $J_{ij} \times x$ where weights J_{ij} can be positive or negative. However, the minus sign in Equations (4.11a) and (4.11b) means that the effect of projection units goes against the effect of control units. For very restricted probability distributions for which all $J_{ij} > 0$ then we could say that projection units are excitatory and control units are inhibitory; however, it is not the case in general.

The detailed implementation of extended Circular BP (and its special case BP for $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$) is similar to neural implementations of predictive coding (Rao and Ballard, 1999; Friston and Kiebel, 2009; Spratling, 2017), and contains three types of units, which are estimate units whose rate is proportional to the log-odds $\{B_i\}$, to the predictions $\{B_i^j\}$, and to their difference $\{B_i - B_i^j\}$,; see Figure 4.2. We hypothesize here, as in Spratling (2016), that neural population represents separately the "prediction errors" $B_i - B_i^j$. This quantity indeed symbolizes the difference between the true belief of node i and its prediction by node j, and can be seen as a prediction error from predictive coding theories if node j is above node i in the equivalent hierarchical Bayesian network (i.e., concept x_j causes concept x_i). We insist here on the fact that the error units, encoding for $\{B_i - B_i^j\}$, are not the same as prediction error units from predictive coding (encoding for the difference between the prediction and the sensory signal). Here, the error $B_i - B_i^j$ is the difference between the estimated *probability* and partial information (estimate of the probability of a variable by another variable). The difference can be made clear on an example: when the network converged, the error $B_i - B_i^j \neq 0$, while the prediction error from predictive coding must be $= 0$ as inputs to neurons are all being perfectly predicted. Instead of using a third type of neural population to represent separately the "prediction errors", we could alternatively have proposed, as in Spratling and Johnson (2003), that the correction or error-detection (of the belief by its rescaled prediction) takes place directly at the non-linear

dendrites of the network units; see also Poirazi et al. (2003); Thalmeier et al. (2016).

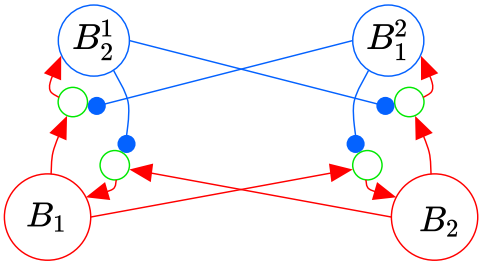

FIGURE 4.2: **Detailed implementation of Circular BP with a rate network, focusing on the connection between B_1 and B_2 from Figure 4.1.** Error neurons in green explicitly encode the "error" $B_i - B_i^j$ (difference between the estimate of x_i and the estimate of x_i from x_j), and $f_{ij}(B_i - B_i^j)$ is sent to estimate neurons in red or blue.

General case As a reminder, in the general case (for probability distributions on binary variable and with at most pairwise factors), function f_{ij} is given by:

$$f_{ij}(x) = \frac{1}{2}\log\left(\frac{\left(\psi_{ij}^{1,1}\right)^{\beta_{ij}}e^{2x} + \left(\psi_{ij}^{1,0}\right)^{\beta_{ij}}}{\left(\psi_{ij}^{0,1}\right)^{\beta_{ij}}e^{2x} + \left(\psi_{ij}^{0,0}\right)^{\beta_{ij}}}\right) \tag{4.12}$$

which means that synapses would depend on 4 parameters as the pairwise potential $\boldsymbol{\psi_{ij}}$ is a 2x2 matrix.[2] Note that the introduction of parameter $\boldsymbol{\beta}$ simply amounts to replacing the pairwise potentials ψ_{ij} from the probability distribution with $\psi_{ij}^{\beta_{ij}}$.

Ising model case The complexity of function f_{ij} in the general case contrasts with the much simpler function $f_{ij}(x) = \phi^{-1}(\phi(J_{ij})\phi(x))$ for an Ising model. It comes:

$$\begin{cases} \tau_j \dot{B}_j^i = -B_j^i + \alpha_{ij}\phi^{-1}\left[\phi\left(\beta_{ij}J_{ij}\right)\phi\left(B_i - B_i^j\right)\right] & \text{(4.13a)} \\ \tau_i \dot{B}_i = -B_i + \kappa_i \sum_{j\in\mathcal{N}(i)} \phi^{-1}\left[\phi\left(\beta_{ij}J_{ji}\right)\phi\left(B_j - B_j^i\right)\right] + \kappa_i\gamma_i I_{\text{ext}\to i} & \text{(4.13b)} \end{cases}$$

which means that synapses only depend on one value $\phi(\beta_{ij}J_{ij})$, the weight of the connection between projection unit i and projection unit j.

Differences with the classical rate model Note that there are several differences between the equations above and the classical rate model equation (see for instance Mastrogiuseppe (2017)) given by:

$$\tau_i \dot{x}_i(t) = -x_i(t) + \sum_{j\in\mathcal{N}(i)} W_{ij}^{\text{rec}}\phi(x_j(t)) + W_i^{\text{inp}} I_{\text{ext}\to i}(t) \tag{4.14}$$

[2] In fact f_{ij} depends only on 3 parameters as $K\boldsymbol{\psi_{ij}}$ and $\boldsymbol{\psi_{ij}}$ lead to the same function.

The first difference is the type of non-linearity used. In the classical rate model, the variable x_i (interpreted as the net input current entering the cell) is transformed into an output firing rate through $\phi = \tanh$ and weighted by the recurrent weight W_{ij}^{rec} (strength of the synapse). On the contrary, in the rate network implementing extended Circular BP (see system (4.13)), the output firing rate is multiplied by $\phi(\beta_{ij}J_{ij})$ (bounded in absolute value by 1), then passed through a non-linearity $\phi^{-1} = \text{artanh}$, and eventually weighted by a scaling factor associated to the postsynaptic neuron (α_{ij} or κ_i).

The second difference, and most important one, is the presence or absence of subtraction inside the non-linearity $\phi = \tanh$ in the classical rate model. An absence of subtraction means that $B_i^j \equiv \alpha_{ij}M_{j\to i} = 0$, i.e., $\boldsymbol{\alpha} = \mathbf{0}$ (not to be mistaken with $\boldsymbol{\alpha} = \mathbf{1}$ which corresponds to BP). $\boldsymbol{\alpha} = \mathbf{0}$ meaning that the message $m_{i\to j}$ sent from node i to node j depends on all the information collected by i. In other words, node i communicates to j information without taking into account that j already knows some of the communicated information. At present time, it is not clear whether the subtraction in system (4.13) could be implemented at the postsynaptic dendrite (see Figure 4.2A), or instead whether an additional unit type encoding for the errors would be needed (see Figure 4.2B).

4.3 Implementation of Circular BP with spiking neurons

Most neuroscience models are based on firing rates, that is, the average number of spikes per time unit of a neuron. However, real neurons communicate through spikes, which are discontinuous quantities. Recently, theoretical neuroscientists have been considering how to perform complex computations accurately and represent continuous quantities using spiking networks (for a review, see Abbott et al. (2016)). In this section, we build upon Denève and Machens' theory of so-called *spike-coding networks* (Denève, 2005, 2008; Boerlin et al., 2013; Denève and Machens, 2016). According to this theory, membrane potentials of neurons implement predictive coding, and more specifically, spikes are generated when a mismatch occurs between the prediction of the variable encoded by the neuron (based upon its own previous spikes, and therefore not encoded explicitly) and the signal arriving to the neuron. In other words, a spike is emitted when the prediction error of the neuron, encoded directly in its membrane potential, exceeds a threshold. Denève (2005) and Denève (2008) discuss the problem of Bayesian inference but in a very specific case (Hidden Markov Models) and these papers do not directly relate to the work presented here (inference in a static factor graph).

The theory by Denève and Machens leads us to propose an implementation of extended Circular BP by a network of spiking neurons, and more specifically, of leaky integrate-and-fire neurons. More precisely, we extend the framework, which uses a single population to encode a (possibly multidimensional) variable following a given differential equation in time. We use instead several populations, connected with specific sparsity and weights, to carry out approximate inference and more specifically, extended Circular BP, with the constraint that each population should encode for a given variable x_i from the probability distribution $p(\mathbf{x})$ on which to perform inference.

4.3.1 Spiking model

In section 4.2, a network composed of rate neurons is proposed to implement (extended) Circular BP. However, the model lacks biological realism, mostly for two reasons. First, neurons cannot have negative rates in reality, while in the model it simply means that the half log-odds $B_i = \frac{1}{2}\log\left(\frac{b_i(x_i=+1)}{b_i(x_i=-1)}\right) < 0$, that is, to beliefs $b_i(x_i = +1) < 0.5$ which is as probable as the contrary $b_i(x_i = +1) > 0.5$. The second reason, and most important one, is that in reality, a neuron does

not transmit an analogous signal (rate) to other neurons but instead discontinuous spikes, which are closer to a digital signal (presence or absence of spike).

Therefore, to increase the biological plausibility of the model, we propose a neural implementation of extended Circular BP using spiking neurons, which have proven capable of encoding continuous variables (see Abbott et al. (2016)).

The proposition of implementation with spiking neurons of extended Circular BP goes one step further than the rate model: instead of encoding variables (log-odds B_i of variable x_i in the probability distribution, or B_i^j) through a single rate unit as in the rate network, the spiking network instead encodes them in *populations* of spiking neurons. These populations of neurons can naturally encode positive as well as negative log-odds B, thanks to having diverse neurons in each population. Note that in the simple case considered here, the variables encoded ($\{B_i\}$ and $\{B_i^j\}$) are one-dimensional, therefore the population consists of only two neurons: one encoding for positive values and one encoding for negative values. However, the multidimensional case can be dealt with as well, very similarly to what is explained below. In fact, the extension from binary to discrete (or equivalently, multidimensional binary) is possible: probabilities of each outcome of the discrete variable x can be encoded by several units using a log-odds code where the activity of neurons is proportional to $\log(x = \theta_i) - \log(x = \theta_j)$ (to be compared to the only neuron needed in the binary case to encode the log-odds $\log(x = 1) - \log(x = -1)$); see Beck and Pouget (2007).

We start by rewriting the equations defining extended Circular BP in continuous time:

$$\begin{cases} \tau_i \dot{B}_i^j = -B_i^j + \alpha_{ij} f_{ij}(B_j - B_j^i) & \text{(4.15a)} \\ \tau_i \dot{B}_i = -B_i + \kappa_i \sum_{j \in \mathcal{N}(i)} f_{ji}(B_j - B_j^i) + \kappa_i \gamma_i I_{\text{ext} \to i} & \text{(4.15b)} \end{cases}$$

Predictions $\{B_i^j\}$ and beliefs $\{B_i\}$ are assumed to be encoded by populations $\{P_i^j\}$ and $\{P_i\}$ of spiking neurons. A spike is a digital quantity (taking discrete values: 0 or 1), contrary to rates which are analog quantities taking any value. Therefore, a spiking network cannot encode analog quantities such as predictions $\{B_i^j\}$ and beliefs $\{B_i\}$ in an exact manner but can encode them approximately through *estimates* $\{\hat{B}_i^j\}$ and $\{\hat{B}_i\}$ (the hat notation indicates estimated quantities by the spiking network). The natural hypothesis is to assume that the continuous value is encoded through the sum of many discrete contributions. Here we assume, as in Boerlin et al. (2013), that estimates $\{\hat{B}_i\}$ are obtained by a weighted, leaky integration of the spike trains $O_k^i(t) = \sum_s \delta(t - t_s^{ki})$ where t_s^{ki} is the time of the s^{th} spike of neuron k of population P_i, and similarly for $\{\hat{B}_i^j\}$ (see Figure 4.3):

$$\begin{cases} \dot{\hat{B}}_i^j = -\lambda_i^j \hat{B}_i^j + \sum_{k \in P_i^j} D_k^{ij} O_k^{ij}(t) & \text{(4.16a)} \\ \dot{\hat{B}}_i = -\lambda_i \hat{B}_i + \sum_{l \in P_i} D_l^i O_l^i(t) & \text{(4.16b)} \end{cases}$$

where $\boldsymbol{D}^i$ (respectively $\boldsymbol{D}^{ij}$) represents the decoding weights of population P_i (respectively P_i^j), and parameters $\{\lambda\}$ are the decay rates for estimated variable $\{\hat{B}\}$ or read-out variable. Equation (4.16a) means that the quantity B_i^j (prediction of variable x_i by variable x_j) or rather its estimated value $\hat{B}_i^j$, is encoded implicitly by population P_i^j through spikes of the neurons belonging to this population (k is the neuron index inside population P_i^j). A spike of neuron k occurs for $O_k^{ij} = 1$. Similarly, Equation (4.16b) means that the quantity B_i (half log-odds of variable

x_i) or rather its estimated value $\hat{B}_i$, is encoded implicitly by population P_i through spikes of the neurons belonging to this population (l is the neuron index inside population P_i).

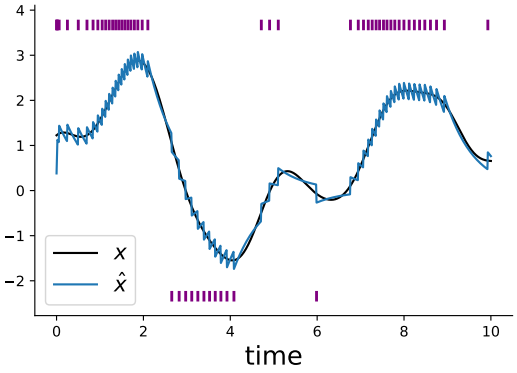

Figure 4.3: **A population of spiking neurons can encode approximately an analog signal thanks to synapses with exponential decrease**. See Boerlin et al. (2013). Here the population has two neurons: the ON neuron encoding for positive values and the OFF neuron encoding for negative values. Spikes of the ON neuron (resp. OFF neuron) are represented at the top (resp. bottom) of the figure in purple. $\hat{x}(t)$ is a reconstruction of the true signal $x(t)$ formed based on spikes of the population. More precisely, $\hat{x}(t)$ is the leaky integration of the spike trains by the population, with weights associated to each neuron. A spike is emitted when the difference $x - \hat{x}$ (difference between the true signal and the prediction of it based on spikes emitted so far) reaches a certain threshold.

We further define, as in Boerlin et al. (2013), the membrane potential of a neuron as a difference between the true variable and its estimate:

$$\begin{cases} V_k^{ij} = D_k^{ij}(B_i^j - \hat{B}_i^j) & \text{(4.17a)} \\ V_l^i = D_l^i(B_i - \hat{B}_i) & \text{(4.17b)} \end{cases}$$

where the weights $\boldsymbol{D}^{ij}$ (resp. $\boldsymbol{D}^i$) are the decoding weights of population P_i^j (resp. P_i) defined previously.

It comes, by writing (4.15a) - $\tau_i\times$(4.16a) and (4.15b) - $\tau_i\times$(4.16b) and assuming that $\lambda_i^j = 1/\tau_i$ and $\lambda_i = 1/\tau_i$:

$$\begin{cases} \tau_j \dot{V}_k^{ij} = -V_k^{ij} + \alpha_{ij} D_k^{ij} f_{ij}(B_j - B_j^i) - \tau_j D_k^{ij} \sum_{m \in P_i^j} D_m^{ij} O_m^{ij} & \text{(4.18a)} \\ \tau_i \dot{V}_l^i = -V_l^i + \kappa_i D_l^i \sum_{j \in \mathcal{N}(i)} f_{ji}(B_j - B_j^i) - \tau_i D_l^i \sum_{n \in P_i} D_n^i O_n^i + D_l^i \kappa_i \gamma_i I_{\text{ext}\to i} & \text{(4.18b)} \end{cases}$$

We need to go a step further in the approximation as B_i and B_i^j are not encoded directly in the network, contrary to $\hat{B}_i$ and $\hat{B}_i^j$. We hypothesize that the spiking network manages to approximate $B_j^i \approx \hat{B}_j^i$ and $B_i \approx \hat{B}_i$ (which is the goal of the network, as stated above). We obtain the following equation, which defines the spiking model implementation of extended Circular BP:

$$\begin{cases} \tau_j \dot{V}_k^{ij} = -V_k^{ij} + \alpha_{ij} D_k^{ij} f_{ij}(\hat{B}_j - \hat{B}_j^i) - \tau_j D_k^{ij} \sum_{m \in P_i^j} D_m^{ij} O_m^{ij} & \text{(4.19a)} \\ \tau_i \dot{V}_l^i = -V_l^i + \kappa_i D_l^i \sum_{j \in \mathcal{N}(i)} f_{ji}(\hat{B}_j - \hat{B}_j^i) - \tau_i D_l^i \sum_{n \in P_i} D_n^i O_n^i + D_l^i \kappa_i \gamma_i I_{\text{ext} \to i} & \text{(4.19b)} \end{cases}$$

It is simply the equation of a leaky integrate-and-fire neuron model (Lapicque, 1907). Overall, the evolution of membrane potentials is lead by a mix between functions which are the integrated spikes, and the spikes themselves (see Equation (4.19b)). Indeed, O are spike trains and is therefore not a continuous variables (delta functions). On the contrary, $\hat{B}$ represents a current as it is a convolution of spike trains (similarly, Rao (2004) considers synaptic currents corresponding to the instantaneous firing rate).

We explicit here the different terms in the differential equation (4.19b) defining the evolution of the membrane potential of neuron l of population i: a leak term, a slow current, a spike term, and an input term (see also Boerlin and Denève (2011); Boerlin et al. (2013); Thalmeier et al. (2016), among others). The **leak term** "$-V$" ensures that the membrane potential decays back to the neutral value of zero in the absence of inputs to the neuron. The **spike term** "$-\tau_i D_l^i \sum_{n \in P_i} D_n^i O_n^i$" represents lateral connections, i.e., within population i. It includes the reset process after a spike (instantaneous self-inhibition of a neuron at the time of a spike - $n = l$ - which resets its membrane potential) and the instantaneous influence of this spike on other neurons (with weight $-D_l^i \times D_n^i$). The **input term** "$+D_l^i \kappa_i \gamma_i I_{\text{ext} \to i}$" represents the influence of the external input on neurons of population P_i (but not on neurons of populations P_i^j, see Equation (4.19a)). Last, the **slow current term** "$+\kappa_i D_l^i \sum_{j \in \mathcal{N}(i)} f_{ji}(\hat{B}_j - \hat{B}_j^i)$" is a different type of contribution called *slow current* in opposition to the infinitely fast connections (term in O which represent spikes propagating instantaneously); the non-linearity allows the network to perform non-linear computations (Poirazi et al., 2003; Abbott et al., 2016) as needed with the (Circular) Belief Propagation algorithm and more generally probabilistic inference. Importantly, all terms are eventually weighted by the decoding weight D_l^i of the neuron. Last, note that if λ and τ are not assumed to be related, a last term "$+D(\tau\lambda - 1)\hat{B}$" appears, which is an additional slow current (postsynaptic potential).

As stated above, in the simple example (which is the case considered here) where the encoded variable is a scalar, a population consists of two neurons: one encoding for positive values and one encoding for negative values. In other words, the decoding weight is positive for ON neurons and negative for OFF neurons: $\boldsymbol{D}^i = (D_+^i; D_-^i)$ and $\boldsymbol{D}^{ij} = (D_+^{ij}; D_-^{ij})$ where D_+^i and $D_+^{ij} > 0$, D_-^i and $D_-^{ij} < 0$. The influence of an ON neuron l of population i to the OFF neuron of the same population is positive as $-D_+^i \times D_-^i > 0$; see the different types of arrows indicating the connections within a population in Figure 4.4. More generally, for encoded variables with a higher dimension, neurons with similar kernels within the same population ($(D_l^i)^T D_n^i > 0$), that is, encoding for similar features, inhibit each other, while neurons with opposite kernels ($(D_l^i)^T D_n^i < 0$) excite each other. This interpretation, although contrary to the main belief in the area of theoretical neuroscience is supported by experimental evidence. For instance, Chettih and Harvey (2019) shows competition between neurons encoding for similar features, thus supporting the whole theory of spike-coding networks (Denève and Machens, 2016) whose principles have been used to design the spiking network.

4.3.2 Spiking condition

Now the most important question remains: when should neurons spike? The predictive coding hypothesis of Boerlin et al. (2013) states that the network should minimize the distance between

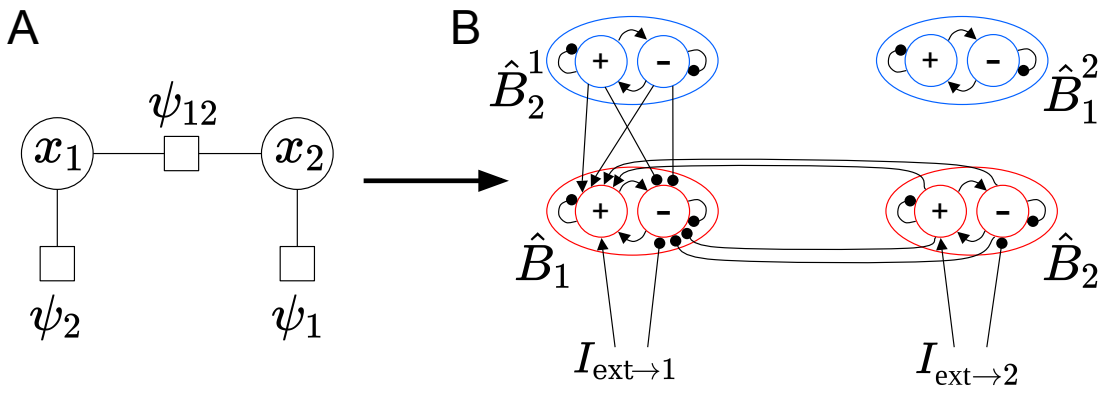

Figure 4.4: **Neural implementation of the (extended) Circular BP algorithm with a spiking network**. **(A)** Example of probabilistic graph, representing the probability distribution $p(x_1, x_2) = \psi_{12}(x_1, x_2)\psi_1(x_1)\psi_2(x_2)$. **(B)** The corresponding spiking network is composed of several populations ($\{P_i\}$ in red and $\{P_i^j\}$ in blue) encoding respectively for the estimates $\{\hat{B}_i\}$ and $\{\hat{B}_i^j\}$ of the half log-odds $\{B_i\}$ and $\{B_i^j\}$. Each population is composed of an ON neuron ("+", encoding for positive values of $\hat{B}$) and an OFF neuron ("-", encoding for negative values). Neurons spike when a correction of the estimation $\hat{B}$ needs to be made. For clarity, the only connections shown are the ones within each population and the ones going to population P_1 encoding for $\hat{B}_1$. Arrows indicate positive connections, while full circles indicate negative connections (see system (4.19)). The true signal B is not known by the network, but it is implicitly encoded in the membrane potentials as $V \propto B - \hat{B}$.

the true variables and their estimates. In other words, a neuron should spike when the estimate $\hat{x}$ (here $\hat{B}_i^j$ or $\hat{B}_i$) of x (here B_i^j or B_i) has to be corrected. The basic idea is that if the estimate $\hat{x}$ is too far (for instance too low) compared to the true value x, then a spike should be emitted for the estimate to get closer to the truth (here increase); see Equations (4.16a) and (4.16b). The more practical rule is that a spike is emitted if the approximation $\hat{x} \approx x$, as measured by some distance $E(t)$, becomes better thanks to this additional new spike.

The minimization of the distance between the true variables and their estimates leads to the following spiking condition or firing rule (see below for a demonstration):

$$O_l^i(t) = 1 \text{ (neuron } l \text{ of population } i \text{ spikes) if } V_l^i(t) > T_l^i \quad \text{where} \quad T_l^i \equiv \frac{(D_l^i)^2}{2} \tag{4.20}$$

Justification of the firing rule Here we demonstrate why the minimization by the network of the distance between the true variables and their estimates naturally leads to the spiking condition stated in Equation (4.20). The demonstration follows closely the one of Boerlin et al. (2013).

We start by defining the distance which we want the network to minimize:

$$E(t) = \sum_{\text{pop}} \int_0^t (B_{\text{pop}}(u) - \hat{B}_{\text{pop}}(u))^2 du \tag{4.21}$$

$$= \sum_i \int_0^t (B_i(u) - \hat{B}_i(u))^2 du + \sum_{(i,j)} \int_0^t (B_i^j(u) - \hat{B}_i^j(u))^2 du \tag{4.22}$$

Note that for simplicity and contrary to Boerlin et al. (2013), we do not include here any additional term symbolizing the metabolic cost of spiking: linear regularization "$+\nu\|\boldsymbol{r}_i(u)\|_1$" or quadratic regularization "$+\mu\|\boldsymbol{r}_i(u)\|_2^2$", where $\boldsymbol{r}_i$ is the firing rate of neurons from population P_i ($r_i(u) = \{r_l^i(u)\}$ where $\dot{r}_l^i(t) = -\lambda r_l^i(t) + \lambda O_l^i(t)$).

The distance $E(t)$ is the sum of errors in each population. However, populations of neuron only have access to part of this information. That is why the only way of minimizing $E(t)$ is to minimize each of its parts locally. For instance, population P_i aims at minimizing

$$E_i(t) = \int_0^t (B_i(u) - \hat{B}_i(u))^2 du \tag{4.23}$$

At each time, neuron l of population P_i should "decide" whether to spike or not. Because the neuron cannot predict future spikes (which also depend on the neuron spiking or not at present time), it necessarily operates greedy minimization on the distance $E_i(t)$: neuron l spikes at time t if and only if spiking reduces $E_i(t)$ in the immediate future after t:

$$E_i(t+\epsilon|\text{neuron } l \text{ spikes}) < E_i(t+\epsilon|\text{neuron } l \text{ does not spike}) \tag{4.24}$$

$$\Leftrightarrow \int_0^{t+\epsilon} (B_i(u) - \hat{B}_i(u) - D_l^i h(u-t))^2 du < \int_0^{t+\epsilon} (B_i(u) - \hat{B}_i(u))^2 du \tag{4.25}$$

Indeed, a spike of neuron l from population i at time t leads to adding function $\delta(u-t)$ to the spike train $O_l^i(u)$ of neuron l, and therefore $h(u-t)$ (weighted by D_l^i) to the read-out variable $\hat{B}_i(u)$, where $h(u-t) = \exp(-\lambda(u-t))$ for $u \geq t$ and 0 otherwise.

It comes, by cancelling terms and using the fact that $h(u-t) = 0$ for $u < t$:

$$\int_t^{t+\epsilon} 2D_l^i h(u-t)(B_i(u) - \hat{B}_i(u) du > \int_t^{t+\epsilon} (D_l^i)^2 h(u-t)^2 du \tag{4.26}$$

As seen previously, the network cannot predict the future, therefore it can only consider the immediate future (greedy optimization): $\epsilon \approx 0$ or stated otherwise, $\epsilon << \lambda$. Therefore, all terms under the integrals are approximately constant (and $h(u-t) \approx 1$, its value for $u = t$). We eventually obtain:

$$2D_l^i(B_i(t) - \hat{B}_i(t)) > (D_l^i)^2 \tag{4.27}$$

By defining the membrane potential of neuron i as $V_l^i \equiv D_l^i(B_i - \hat{B}_i)$ and the threshold $T_l^i \equiv (D_l^i)^2/2$, we eventually obtain the spiking condition provided in Equation (4.20). □

4.3.3 Quality of the approximation

Choosing the decoding weights The quality of the approximation by the spiking network highly depends on the values of the decoding weights D. As a matter of fact, when a spike occurs, the estimate $\hat{B}$ of B instantaneously changes by D. Because the network is designed to minimize the distance between the eCBP value B and its estimate $\hat{B}$ by the spiking network, then a spike occurs when $B - \hat{B} = \pm D/2$, after which still $B - \hat{B} = \mp D/2$: the distance $E(t)$ is still roughly the same after the spike but $\hat{B}$ jumped to the other side of B (see spiking condition from Equation (4.27)). The precision of the approximation therefore highly depends on the value of decoding weights D: the lower the decoding weights, the better the approximation.

Results of the numerical simulations We simulate the spiking network defined by the system of Equations (4.19) (taken in discrete time) and the spiking condition of Equation (4.20).

For technical reasons, we do not take $D_+ = -D_-$ but instead $D_+ = 0.95D_-$ in order to avoid the so-called *ping-pong effect*. This effect corresponds to a spike from the neuron encoding for positive values (resp. negative) automatically triggering a spike at the neuron encoding for negative values (resp. positive) at the next time step and then back to the initial neuron, etc.; see also the Supplementary Material of Boerlin et al. (2013).

We take $\tau = 1/\lambda = 20$ (τ is the time constant of the leak of $\hat{B}$, while λ is the damping (leak) parameter in the damped eCBP algorithm).

Parameters of eCBP are taken randomly: $\kappa_{ij} \sim \Gamma(20, 0.05)$, $\gamma_{ij} \sim \Gamma(20, 0.05)$, $\alpha_{ij} \sim \mathcal{N}(1, 0.5)$, and $\beta_{ij} \sim \mathcal{U}(0.8, 1.2)$.

The probability distribution was picked randomly: the graph is generated using the Erdos-Renyi model with 10 nodes and connection probability 0.4. Existing connections are weighted with $J_{ij} \sim \mathcal{N}(0, 1)$.

Figure 4.5 shows that the spiking network approximates well the eCBP algorithm. Intriguingly, as Figure 4.5A shows, not only both system converge to the same value of the belief, but the dynamics are also the same.

4.3.4 Comparison with the rate network and biological plausibility

Comparison with the rate network We saw previously that the spiking network implements extended Circular BP only approximately, as it uses spikes and therefore cannot encode beliefs with infinite precision. This is contrary to the rate network proposed in section 4.2, which implements eCBP exactly. Indeed, the rate network can encode directly for the continuous variables $\{B_i\}$ and $\{B_{ij}\}$ in the rate of its units.

However, there are some similarities between this proposed implementation of eCBP with a spiking network and the rate network. In fact, in the case considered here where the encoded variable is one-dimensional (B_i is half the log-odds the binary variable x_i, therefore its dimension is equal to one), the spiking network behaves similarly to a rate network at convergence. As shown in see Figure 4.5A, neurons spike regularly once the system converges with a rate proportional to the encoded variable ($\hat{B} \approx B$).

However, this regularity in the spiking only comes from the hypothesis that the estimated variable $\hat{B}$ is the result of a *leaky* integration of spikes. For a leak term $\lambda = 0$ in Equations (4.16a) and (4.16b), spiking only takes place when the encoded quantities change.

As shown in section 2.3.8.3, the rate of the spiking neurons is proportional to $\left|B + \lambda \dot{B}\right|$: both the encoded variable (through B) and the temporal variation of this variable (through $\dot{B}$) play a role in the neural activity (see also Figure 4.3). Therefore, it is only after convergence that the

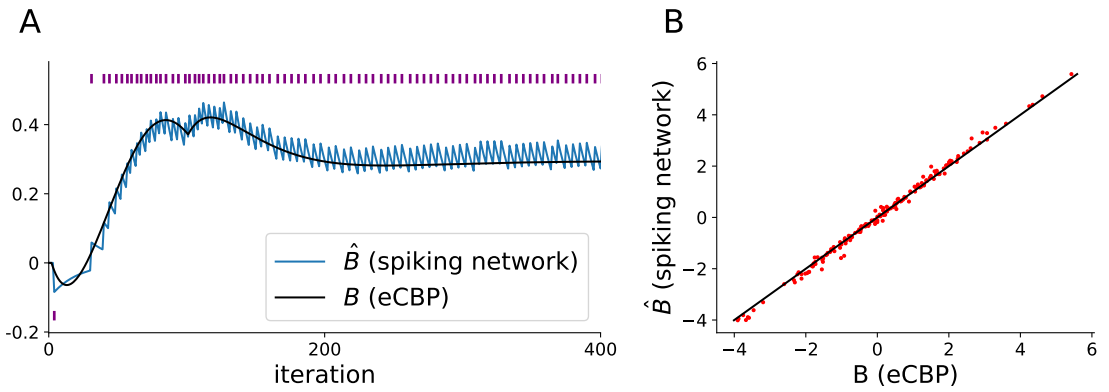

FIGURE 4.5: **The spiking network of Figure 4.4 implements extended Circular BP, approximately because it uses spikes, but very closely**, using the spiking rule of Equation (4.20). **(A)** Reconstructed signal (estimated beliefs $\hat{B}$) versus true signal (beliefs B, obtained by running the eCBP algorithm), for one example population P_i. Spikes of the ON neuron (resp. OFF neuron) are represented in purple of the figure at the top (resp. bottom). **(B)** Reconstructed signal versus true signal on many random graphs and for many random external input examples, after 800 iterations. Overall, the spiking network approximates well the eCBP algorithm, given that it only uses spikes (zeros and ones) to encode for continuous variables and therefore cannot possibly implement eCBP exactly.

spiking network strongly resembles a rate network, as $\dot{B} = 0$ and therefore rate $\propto |B|$. It is even possible to have zero spiking at a point in time when B is strong if $\dot{B} = -1/\lambda B$.

Interestingly, the spiking network could implement approximate inference for more complicated problems, e.g., if variable x_i is not binary but discrete with $k > 2$ states, *in an efficient manner*, i.e., with the population of neurons collectively encoding for all encoded variables (Denève and Machens, 2016). The spiking network and the rate network become less and less similar as the dimension of the encoded variable increases.

Biological plausibility The spiking network model is more biologically realistic than the rate network previously proposed. Dynamics of the leaky integrate-and-fire neurons' membrane potentials are biophysically plausible and result in Poisson-like spiking as observed in the cortex (Denève and Machens, 2016). However, it is still not a completely realistic model of neuronal activity. Recent work has been carried out towards this goal. Schwemmer et al. (2015) extends the spike-based approach of Boerlin and Denève (2011); Boerlin et al. (2013) known as "spike coding network" and which serves as a starting point here, into a biophysically plausible networks with conductance-based neurons and slower synapses than the instantaneous lateral connections ("spike term" from above). Maraš (2019) investigates the effects on the quality of the approximation by the network introduction and its robustness to noise due to synaptic delays, which goes against the instantaneous "fast connections" from the model. The work by Maras also considers experimentally observed sparse connectivity within a population, which could potentially be extended in the present network (which contains several populations) to sparse connectivity *between* populations. Additionally, real neural networks abide by Dale's law, according to which connections leaving a given neuron should all have an identical sign. The current spiking model violates Dale's law, as the connectivity within a population i is $(D_l^i)^T D_n^i$. Dale's law can be

respected by separating cost functions for excitatory and inhibitory neurons, as shown in the Supplementary Material of Boerlin et al. (2013). Finally, many features of the model do not correspond to biology, like the simplicity of the leafy integrate-and-fire model, the absence of adaptation meaning that the same sequence of stimuli will trigger exactly the same response from the network, etc.

4.4 Approximations of Circular BP and their neural implementation

In this section, we discuss the limitations of the rate model presented implementing extended Circular BP in the binary case. This rate model is defined by Equations (4.11a) and (4.11b)) and is represented in Figure 4.1 and 4.2. There are two limitations to this rate model.

The first limitation is that the same computation takes place twice: the operation $B_i - B_i^j$ is required for the computation of B_j ($B_i - B_i^j$ is encoded at the error interneuron or equivalently at the dendrite of node j) but also for the computation of B_j^i ($B_i - B_i^j$ is encoded by the interneuron).

The second one is that one neuron (or population of neurons) is required to encode $M_{i\to j}$, assigned to the oriented edge $i \to j$. The necessary number of coding units is thus proportional to n^2 for graphs like random graphs, where n is the number of variables in the probability distribution. This is far from being optimal, given that the goal is to find the n marginals of the distribution. In this section, we propose approximations of Circular BP, and the corresponding neural implementation, which use n coding units.

Here we consider several approximations of extended Circular BP in the binary case, and their associated neural implementation.

Transfer function approximation

In the field of neuroscience, a traditional transfer function is $\phi = \tanh$ is used. This contrasts with function f_{ij}, the update function associated to BP or Circular BP, defined by $f_{ij}(x) = \phi^{-1}(\phi(J_{ij})\phi(x))$.

Using the approximation $\phi^{-1}(\phi(x)\phi(y)) \approx \phi(x)\phi(y)$, we get the following *approximate* system:

$$\begin{cases} \tau_j \dot{B}_j^i = -B_j^i + W_{ij}^c \phi\Big(B_i - B_i^j\Big) & \text{(4.28a)} \\ \tau_i \dot{B}_i = -B_i + \sum\limits_{j\in\mathcal{N}(i)} W_{ij}^p \phi\Big(B_j - B_j^i\Big) + W_{ii}^{in} I_{\text{ext}\to i} & \text{(4.28b)} \end{cases}$$

where the effective recurrent weights are $W_{ij}^c = \alpha_{ij}\phi(J_{ij})$ to a control unit, and $W_{ij}^p = \kappa_i \phi(J_{ij})$ to a projection unit. The input matrix is $W^{in} = \text{diag}\big(\{\kappa_i \gamma_i\}\big)$.

Circular BP in continuous time (more general case) For more general probability distributions, the connectivity W does not need to be symmetric and the transfer function of the rate neurons becomes $\phi(\cdot + d_{ij}) + c_{ij}$ instead of ϕ.

While still taking into account probability distribution over binary variables, we consider the case where the pairwise factor is any 2x2 matrix.

In this more general case, $f_{ij}(x) \approx W_{ij}\phi(x + c_{ij}) + d_{ij}$ where parameters W_{ij}, c_{ij}, d_{ij} depend on the coefficients of the 2x2 matrix ψ_{ij}. The transfer function of the rate neuron is then not $\phi(x)$ as in the Ising model case but instead a shifted version: $\phi(x + c_{ij}) + d_{ij}$ (also used in

neuroscience).

$$\begin{cases} \tau_j \dot{B}_j^i = -B_j^i + W_{ij}^c \phi\Big(B_i - B_i^j + c_{ij}\Big) + d_{ij} & \text{(4.29a)} \\ \tau_i \dot{B}_i = -B_i + \sum\limits_{j \in \mathcal{N}(i)} W_{ij}^p \phi\Big(B_j - B_j^i + c_{ij}\Big) + W_{ii}^{in} I_{\text{ext}\to i} + f_i & \text{(4.29b)} \end{cases}$$

with $f_i \equiv \sum\limits_{j \in \mathcal{N}(i)} d_{ij}$.

This allows the connectivity W to be non-symmetric (although constrained: it does not seem possible to propose a 2x2 factor corresponding to any couple (W_{ij}, W_{ji})).

However, the parallel with rate networks can be not only drawn for Ising models, for which $J_{ij} = J_{ji}$, but for any probability distribution over binary variables with pairwise interactions. In this more general case, the connectivity can be non-symmetric, and the transfer function of the rate neurons is not the hyperbolic tangent ϕ as above but a shifted version of it: $\phi(\cdot + d_{ij}) + c_{ij}$

Circular BP without subtraction

We propose in this section a simple implementation of a mean-field Circular BP algorithm. As stated above, the Belief Propagation algorithm has a particular feature which makes it hard to implement neural circuits (Raju and Pitkow, 2016): the subtraction "$-M_{j\to i}$" in the message update equation $M_{i\to j}^{\text{new}} = f_{ij}(\sum_{k\in\mathcal{N}(i)\setminus j} M_{k\to i} + M_{\text{ext}\to i}) = f_{ij}(B_i - M_{j\to i})$ in the log-domain, or equivalently, the partial product over $k \in \mathcal{N}(i)\setminus j$ in the original formulation. Some papers ignore this subtraction completely while proposing a neural implementation of probabilistic inference (Ott and Stoop, 2006; Litvak and Ullman, 2009) . The Belief Propagation algorithm without subtraction actually corresponds to the Circular BP algorithm with $\boldsymbol{\alpha} = \mathbf{0}$. Note that this particular algorithm is different from mean-field inference (see section 3.2.1.2); in fact, it often performs worse than mean-field inference. Having $\boldsymbol{\alpha} = \mathbf{0}$ in Circular BP represents a situation where inferences are fully circular in the sense of Jardri and Denève (2013a). We name this special case of Circular BP algorithm the "full" Circular BP algorithm.

We consider here the full extended Circular BP algorithm, i.e., extended Circular BP in the particular case $\boldsymbol{\alpha} = \mathbf{0}$. This leads to the following update equations:

$$\begin{cases} M_{i\to j}^{\text{new}} = f_{ij}(B_i) & \text{(4.30a)} \\ B_i = \sum\limits_{j \in \mathcal{N}(i)} M_{j\to i} + M_{\text{ext}\to i} & \text{(4.30b)} \end{cases}$$

where

$$f_{ij}(x) = \phi^{-1}\big(\phi(J_{ij})\phi(x)\big) \tag{4.31}$$

This particular case of extended Circular BP, which can be seen as an approximation to it (approximation at order 0 in α), can be implemented by a rate network with only one unit type. Indeed, the system above can be rewritten as a single equation with the log-odds $\mathbf{B}$ (but without the messages $\mathbf{M}$):

$$B_i = \sum_{j \in \mathcal{N}(i)} f_{ji}(B_j) + M_{\text{ext}\to i} \tag{4.32}$$

which is similar to the classical rate network described in section 4.4, although with a more complicated transfer function.

Note that in this network, connection weights are symmetrical as $J_{ij} = J_{ji}$, and $\big|W_{ij}\big| = \big|\phi(J_{ij})\big| < 1$.

In this case, messages are encoded at the neurons' dendrites, and all messages are summed as the belief, encoded in the soma.

Approximating into a classical rate network

The network implementing (extended) Circular BP is similar to the classical rate network equation:

$$\tau \dot{x}_i = -x_i + \sum_{j \in \mathcal{N}(i)} W_{ij}^{\text{rec}} \phi(x_j) + W_i^{\text{in}} \boldsymbol{I}_{\text{ext}} \tag{4.33}$$

The major difference between equations (4.11a) and (4.33) lies in the presence or absence of correction inside the non-linearity ϕ. Without this correction (i.e., for $\boldsymbol{\alpha} = \mathbf{0}$), the quality of inference becomes catastrophic: the model without correction obtained by fitting the recurrent weights $\boldsymbol{J}$ is even outperformed by BP (see section 3.5.5).

Going a step further, one can approximate f_{ij} with:

$$f_{ij}(x) = \phi^{-1}\big(\phi(J_{ij})\phi(x)\big) \tag{4.34}$$

$$\approx \phi(J_{ij})\phi(x) \tag{4.35}$$

This leads to the following update equation:

$$B_i = \sum_{j \in \mathcal{N}(i)} W_{ij} \phi(B_j) + M_{\text{ext} \to i} \tag{4.36}$$

which exactly corresponds to the classical rate network described in section 4.4.

In this network, the connectivity $W_{ij} = \phi(J_{ij})$ is symmetrical ($W_{ij} = W_{ji}$) and bounded ($\big|W_{ij}\big| < 1$). Furthermore, the input weight matrix is diagonal. Finally, the output weight matrix is equal to the identity matrix: each unit encodes a particular variable x_i of the probability distribution $p(x)$. The alternative would have been to decode the marginals based on the activity in the units composing the network.

4.5 Unsupervised learning of Circular BP parameters: how the brain might balance the probabilistic reasoning network

We start by recalling the extended Circular BP equations (3.26a) and (3.26b):

$$\begin{cases} M_{i \to j} = f_{ij}(B_i - \alpha_{ij} M_{j \to i}) & \text{(4.37a)} \\ B_i = \kappa_i \Big(\sum_{j \in \mathcal{N}(i)} M_{j \to i} + \gamma_i M_{\text{ext} \to i} \Big) & \text{(4.37b)} \end{cases}$$

which can be implemented by following rate network in continuous time (see Equations (4.11a) and (4.11b)):

$$\begin{cases} \tau_j \dot{B}_j^i = -B_j^i + \alpha_{ij} f_{ij}(B_i - B_i^j) & \text{(4.38a)} \\ \tau_i \dot{B}_i = -B_i + \kappa_i \sum_{j \in \mathcal{N}(i)} f_{ji}(B_j - B_j^i) + \gamma_i \kappa_i I_{\text{ext} \to i} & \text{(4.38b)} \end{cases}$$

where $B_j^i = \alpha_{ij} M_{i \to j}$. κ_i is simply a scaling factor for the log-odds $\{B_i\}$, and α_{ij} is a scaling factor for the corrected predictions $\{B_i^j\}$.

4.5.1 Motivation for an unsupervised learning method

In section 3.5.5, I considered supervised learning, and more specifically, learnt the parameters minimizing the mean square error between the marginals predicted by (extended) Circular BP and the exact marginals. Alternatively, another supervised learning method is proposed in Wiegerinck and Heskes (2002), which consists of minimizing the KL divergence between the predicted pairwise marginals and the true pairwise marginals. Another example of supervised learning is Yoon et al. (2018) which minimizes the KL divergence between the unitary marginals to learn the parameters of the model.

The inconvenient of all these methods is the need to generate training examples (true marginals, unitary or pairwise), which has an exponential complexity in the number of graph nodes. As a consequence, these fitting procedures are not scalable to bigger complex graphs (e.g., highly connected Erdos-Renyi graphs with a high number of nodes).

A second inconvenient of supervised learning is its lack of biologically plausibility as a potential learning algorithm used by the brain. Therefore, it would be interesting to propose an unsupervised learning method to learn the corrective multiplicative factors. Inspired by the literature of balanced networks (Renart et al., 2010; Tetzlaff et al., 2012), we describe in this section a preliminary unsupervised learning method derived from decorrelating messages going in opposite directions.

4.5.2 Formulation of the unsupervised learning rule

We propose here a way of learning online the parameters of extended Circular BP for $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1})$ fixed. For discussion on the way that one might learn ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$), see section 4.5.5.

The proposed learning rule on parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ is:

$$\begin{cases} \Delta\alpha_{ij} = \eta_1 M_{j\to i}\Big(B_i - \alpha_{ij} M_{j\to i}\Big) + \eta_1 M_{i\to j}\Big(B_j - \alpha_{ij} M_{i\to j}\Big) & \text{(4.39a)} \\ \Delta\kappa_i = \eta_2 M_{\text{ext}\to i}(B_i - M_{\text{ext}\to i}) & \text{(4.39b)} \end{cases}$$

where messages and beliefs are taken after $T = 100$ iterations of Circular BP, and η_1 and η_2 are learning rates. The second term in the right-hand side of Equation (4.39a) ensures that matrix $\boldsymbol{\alpha}$ is symmetric, i.e., that α_{ij} is associated to the *unoriented* edge (i, j) in the approximation of the Gibbs Free Energy (see Equation (3.14)). If there is not symmetry constraint on $\boldsymbol{\alpha}$ (in this case, Circular BP is written $M_{i\to j} = f_{ij}(B_i - \alpha_{i\to j} M_{j\to i})$ in an Ising model; see also Equation (1.12) for the general case), the learning rule on $\boldsymbol{\alpha}$ becomes instead:

$$\Delta\alpha_{i\to j} = \eta_1 M_{j\to i}\Big(B_i - \alpha_{i\to j} M_{j\to i}\Big) \tag{4.40}$$

where $\alpha_{i\to j}$ is used in the computation of $M_{i\to j}$ as a weight to the message going in the opposite direction $M_{j\to i}$; in this case, $\alpha_{i\to j}$ is the scaling factor associated B_j^i as the corrected prediction of x_j by x_i is $B_j^i = \alpha_{i\to j} M_{i\to j}$.

4.5.3 Results of the unsupervised learning

The unsupervised learning rule manages to learn parameters of Circular BP and achieve with good performance in all tested cases ; see Figure 4.6. We used 5000 training examples, and learning rates starting from $\eta_1 = 0.03$ and $\eta_2 = 0.0003$ (which were both decreased by half after one third of the optimization, and after two thirds of the optimization). Note that we add some amount of damping to the algorithm, $\boldsymbol{\epsilon} = \mathbf{0.7}$ (see section 4.2.1 about damping), contrary to the supervised learning case.

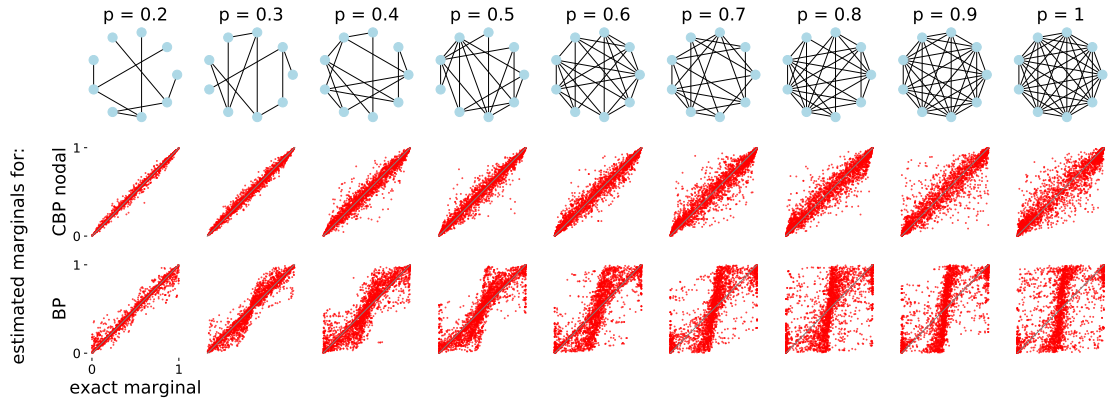

Figure 4.6: **Results of the unsupervised learning rule on Erdos-Renyi graphs**. Nodal Circular BP (i.e., eCBP with $(\boldsymbol{\beta}, \boldsymbol{\gamma} = (\mathbf{1}, \mathbf{1}))$ outperforms BP for all connection probabilities. The performance of the unsupervised learning rule is comparable in highly connected graphs to the one of the supervised learning procedure (see "CBP nodal" line in Figure 3.6). Both algorithms are taken with damping, which helps for convergence (especially for BP).

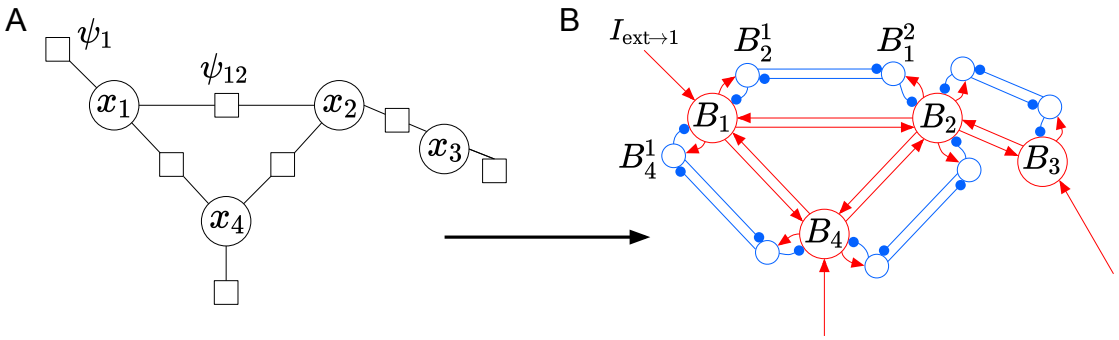

Figure 4.7: **Balancing the network to carry out near-optimal probabilistic inference**. **(A)** The probability distribution $p(\mathbf{x})$ is represented by a factor graph, with pairwise potentials ψ_{ij} and unitary potentials ψ_i. **(B)** Rate network with two types of units, implementing (extended) Circular Belief Propagation; see Equation (4.38). Connection weights depend on ψ_{ij}, while external input depend on ψ_i. Projection units in red encode the approximate marginal probability $b_i(x_i) \approx p_i(x_i)$, while control units in blue remove information being reverberated between projection units. For example, B_2^1 removes from B_1 the message which went through $B_1 \to B_2 \to B_1$ (reciprocal connection), but also for instance $B_1 \to B_4 \to B_2 \to B_1$ (cycle). In contrast, BP only removes redundant information caused by reciprocal connections.

4.5.4 Understanding the learning rule

Here we motivate the learning rule proposed above, and more specifically, why the learning rule on $\boldsymbol{\kappa}$ is a homeostasic rule, and the learning rule on $\boldsymbol{\alpha}$ is anti-Hebbian (or inhibitory Hebbian).

Learning rule on $\boldsymbol{\alpha}$ The unsupervised learning rule for $\boldsymbol{\alpha}$ minimizes the total amount of information sent in the network $\sum_{(i,j)} M_{i\to j}^2$, or equivalently, minimizes the quantity of so-called prediction errors $\sum_{(i,j)} (B_i - B_i^j)^2$:

$$L = \sum_{(i,j)} M_{i\to j}^2 \tag{4.41}$$

$$\approx \sum_{(i,j)} \left(\phi^{-1}\Big[\phi(J_{ij})\phi(B_i - \alpha_{ij}M_{j\to i})\Big]\right)^2 \tag{4.42}$$

We compute the derivative of the loss function L w.r.t. $\boldsymbol{\alpha}$ by considering that $\frac{\partial(B_i - \alpha_{ij}M_{j\to i})}{\partial \alpha_{kl}} \approx -M_{j\to i}\delta_{jk}\delta_{il}$, i.e., by neglecting higher-order dependencies. Consequently,

$$\frac{\partial L}{\partial \alpha_{ij}} \approx \frac{\partial M_{i\to j}^2}{\partial \alpha_{ij}} + \frac{\partial M_{j\to i}^2}{\partial \alpha_{ij}} \tag{4.43}$$

where

$$\frac{\partial M_{i\to j}^2}{\partial \alpha_{ij}} \approx -2M_{i\to j}(\phi^{-1})'\Big(\phi(J_{ij})\phi(B_i - \alpha_{ij}M_{j\to i})\Big)\phi(J_{ij})\phi'(B_i - \alpha_{ij}M_{j\to i})M_{j\to i} \tag{4.44}$$

which has the same sign as $-2M_{i\to j}\phi(J_{ij})M_{j\to i}$ (symmetrical in (i,j)). We thus propose the following learning rule:

$$\Delta\alpha_{ij} \propto -\frac{\partial L}{\partial \alpha_{ij}} \propto M_{i\to j}\phi(J_{ij})M_{j\to i} \tag{4.45}$$

This learning rule tends to make the correlations between opposite messages disappear, i.e., $< M_{i\to j}M_{j\to i} >_{\text{examples}} \approx 0$. In the simulations, we used an alternative learning rule (see Equation (4.39a)):

$$\Delta\alpha_{ij} \propto (B_i - \alpha_{ij}M_{j\to i})M_{j\to i} + (B_j - \alpha_{ij}M_{i\to j})M_{i\to j} \tag{4.46}$$

(the right-hand terms of Equations (4.45) and (4.46) have identical signs). This can be seen as a inhibitory Hebbian learning rule (or anti-Hebbian learning rule) "$\Delta w \propto r_I(r_E - wr_I)$", which tends to balance the network.

Learning rule on $\boldsymbol{\kappa}$ The learning rule on $\boldsymbol{\kappa}$ aims at avoiding reverberation of external information by cycles. κ_i acts as a scaling factor on the log-odds (see Equation (3.26b)). The proposed learning rule does synaptic scaling (or homeostatic scaling). It aims at decorrelating the information truly received by node i from the outside world ($M_{\text{ext}\to i}$), and all the information received by node i except this true external information ($B_i - M_{\text{ext}\to i}$):

$$\Delta\kappa_i \propto M_{\text{ext}\to i}(B_i - M_{\text{ext}\to i}) \tag{4.47}$$

Note that a learning rule consisting in from minimizing, as for $\boldsymbol{\alpha}$, the quantity of messages sent throughout the graph, would not work here as $\boldsymbol{\kappa} \to \mathbf{0}$ would be enough to have all messages go to zero. Finally, the learning rule $\Delta\kappa_i \propto M_{\text{ext}\to i}(B_i - \kappa_i M_{\text{ext}\to i})$ would not work either as the sign of the correlation between $M_{\text{ext}\to i}$ and $B_i - \kappa_i M_{\text{ext}\to i} \propto M_{j\to i}$ does not depend on κ_i but instead on the graph topology and weights $\boldsymbol{J}$.

Neural interpretation and balanced network In the special case of probability distribution over binary variables, we can draw a parallel with a rate network composed of two types of units. While projection units encode the marginal probability of the associated variable, control units attempt to remove redundancies by predicting the information received by the projection units (see also Li and Pehlevan (2020)), thus balancing the system and decorrelating information sent between regions, allowing for efficient probabilistic inference.

The right amount of control prevents the network from overamplifying the messages. More specifically, it allows the network to control the flow of information by avoiding to double count messages and therefore by only spreading meaningful information. We call this state *Balanced* Circular Belief Propagation: indeed, this mechanism is analogous to balancing recurrent excitation by local inhibition in a neural network (Vogels et al., 2011; Brendel et al., 2020). Here we talk about *tight* balance, as each input sent by a projection neuron is individually balanced by a prediction from a control unit and encoded at the error neuron (or equivalently without error neuron, at every dendrite of the projection neuron receiving the signal).

The proposed approach not only improves the quality of probabilistic inferences but also brings better higher stability to the network, which is a known feature of balanced excitatory-inhibitory networks.

4.5.5 Learning the remaining parameters of eCBP

In all section 4.5 until now, we considered the case where $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1})$ and proposed a way of learning these parameters in an unsupervised fashion. Here we ask the question of the learning of these additional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

As a reminder, parameter $\boldsymbol{\beta}$ simply appears in the formulation of eCBP as a multiplicative factor to the true weights $\boldsymbol{J}$ (defined by $\psi_{ij}(x_i, x_j) = \exp(J_{ij} x_i x_j)$). Similarly, parameter $\boldsymbol{\gamma}$ simply appears as a multiplicative factor to the true external inputs $\boldsymbol{M}_{\text{ext}}$ (defined by $\psi_i(x_i) = \exp(M_{\text{ext}\to i} x_i)$). Therefore, fitting $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is equivalent to fitting $\boldsymbol{J}$ and $\boldsymbol{M}_{\text{ext}}$, respectively the network weights and the input to the network (where $\boldsymbol{J}$ and $\boldsymbol{M}_{\text{ext}}$ are not related anymore to the factors of the probability distribution).

Parameter $\boldsymbol{\beta}$ can be learnt similarly to Mongillo and Deneve (2008); Jardri and Denève (2013a) which proposes a expectation maximization (EM) algorithm (Dempster et al., 1977) and alternatively a Hebbian-like learning rule using stochastic gradient descent, to learn the factors $\{\psi_{ij}\}$. This approach is based on the fact that messages from BP and its variants allow not only to compute unitary marginal probabilities $\{p_i(x_i)\}$ but also pairwise marginal probabilities $\{p_{ij}(x_i, x_j)\}$. For instance, for eCBP, one can assume that:

$$b_{ij}(x_i, x_j) \propto \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \psi_i(x_i)^{\kappa_i \gamma_i} \psi_j(x_j)^{\kappa_j \gamma_j} \prod_{k \in \mathcal{N}(i)\setminus j} m_{k\to i}(x_i)^{\kappa_i} \times \prod_{k \in \mathcal{N}(j)\setminus i} m_{k\to j}(x_j)^{\kappa_j} m_{j\to i}(x_i)^{\kappa_i - \alpha_{ij}} m_{i\to j}(x_j)^{\kappa_i - \alpha_{ji}} \tag{4.48}$$

$$\implies b_{ij}(x_i, x_j) \propto \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \times \frac{b_i(x_i)}{m_{j\to i}(x_i)^{\alpha_{ij}}} \times \frac{b_j(x_j)}{m_{i\to j}(x_j)^{\alpha_{ij}}} \tag{4.49}$$

which has the same form as the expression of the message update equation for eCBP: $m_{i\to j}(x_j)$ depends on $b_i(x_i)/m_{j\to i}(x_i)^{\alpha_{ij}}$ in Equation (3.24). In particular, for an Ising model, Equation (4.49) gives:

$$\log\left(\frac{b_{ij}(1,1)}{b_{ij}(1,0)}\right) = 2\alpha_{ij}\beta_{ij}J_{ij} + (B_j - \alpha_{ij}M_{i\to j}) \tag{4.50}$$

which shows the importance of the quantity $B_j - \alpha_{ij} M_{i \to j}$, together with the message update equation of eCBP $M_{j \to i} = f_{ij}(B_j - \alpha_{ij} M_{i \to j})$. Note, however, that Equations (4.48), (4.49) and (4.50) are only approximations for eCBP. Indeed, Equation (4.48) is the expression of the pairwise beliefs for extended Fractional BP and not extended Circular BP (see Equation (A.7)). Such an equation does not exist for extended Circular BP, as eCBP does not come from any approximation of the Gibbs free energy (which is the starting point to yield such an expression). However, as seen previously, eCBP can be defined as an approximation to eFBP with a slightly simpler message update equation. By making the hypothesis that even though eCBP has a different update equation for the messages, it eventually computes the (approximate) solution to eFBP, then Equation (4.48) is (approximately) valid.

Jardri and Denève (2013a) proposes, in the particular situation where $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$ (i.e., for the original Circular BP), an online update of the factor parameters (here, simply J_{ij} or equivalently, β_{ij}) based on the pairwise beliefs $\{b_{ij}(x_i, x_j)\}$, with a Hebbian-like learning rule. Learning the external message $M_{\text{ext} \to i}$ (or equivalently, its weight γ_i) could be done similarly by defining an abstract "external neuron" connected to x_i through weight γ_i.

Overall, it might be possible (this still needs to be demonstrated numerically) to learn all parameters of eCBP in an unsupervised way: $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. The learning on $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ is anti-Hebbian (or inhibitory Hebbian) and the learning on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is Hebbian.

4.6 Conclusion

We showed that for Ising models, the proposed algorithm can be implemented by a rate network. In this network, removal of redundant probabilistic information created by cycles is carried out by control units while projection units try to perform inference over variables. This offers an analogy with excitatory-inhibitory balanced networks, hence the term *Balanced* Circular Belief Propagation. The corresponding spiking neural network would need to be micro-balanced, with each recurrent excitatory input controlled by an inhibitory input of similar strength (Denève and Machens, 2016; Li and Pehlevan, 2020; Ahmadian and Miller, 2021) , which results in very efficient and fast inference systems (Hennequin et al., 2014; Aitchison and Lengyel, 2016; Echeveste et al., 2020).

However, such tight balance is unlikely to be perfectly achieved in brain networks. Excitation-inhibition imbalance in the brain has also been associated to hallucinatory experiences and more generally is a biological marker of autism and schizophrenia (Sohal and Rubenstein, 2019; Jardri et al., 2016). Small deviations from tight balance are frequent and could drive some normal perceptual features like bistability (Leptourgos et al., 2020a). A lack of precise control would introduce deviations from exact inference in human observers. This could lead to general overconfidence or even aberrant beliefs and eventually modified behavior (Jardri and Denève, 2013a; Bouttier et al., 2021); see also section 2.3.

One limitation of the supervised learning procedure (presented in section 3.5), despite the need to generate exact marginals which is infeasible in large graphs, is its lack of plausibility. The unsupervised learning method to learn the corrective multiplicative factors bring a solution to this issue. Inspired by the literature of balanced networks (Renart et al., 2010; Tetzlaff et al., 2012), we describe in section 4.5 a preliminary unsupervised learning method based on decorrelating messages going in opposite directions.

Crucially, we showed in section 4.5 that parameters of Circular BP can be learned through anti-Hebbian learning and homeostatic plasticity, leading to a decorrelation of the network.

Chapter 5

Circular Belief Propagation in more general cases

Summary of Chapter 5

In this chapter, we expand the range of distributions which can be dealt with by the Circular Belief Propagation algorithm. We present the extended Circular Belief Propagation in the general case, that is, in a general factor graph. Additionally, we particularly investigate the case where variables are Gaussian, which leads to a very simple formulation of (Gaussian) Circular BP, very similar to the one in the binary case. This allows us to expand the neural model presented in chapter 4, which applied exclusively to pairwise Markov Random Fields with binary variables, by addressing the Gaussian case. Similarly to the binary case, populations of the neural model encode for the parameters of the probability distribution (mean and inverse variance), which requires twice as many units as in the binary case. Three types of populations are involves: message populations, marginal populations, and prediction error population. In each population, the firing rate of units is proportional to the value they encode: one unit encodes for the inverse variance, and the other one encodes for the product between the mean and the inverse variance. As in the binary case, the rate model implements exactly eCBP, and could be translated into a spiking network as in the binary case.
Next, we describe the effects of circularity in the Gaussian case by simulating Gaussian Circular BP. Similarly to the binary case where circularity lead to overconfidence, estimates of probabilities are overconfident (as measured by the inverse variance or precision). Additionally, the MAP estimation (mean of the distribution) also gets strongly modified by circularity, contrary to the binary case; simulations show that the means are in general overestimated.
Last, we include preliminary discussions on the comparison between Gaussian Circular BP and impaired predictive coding, which both model brain processes underlying the emergence of sub-optimal behavior and psychosis, and are currently the two main types of Bayesian models in computational psychiatry.

5.1 Introduction

In this chapter, we expand the range of distributions which can be dealt with by the Circular Belief Propagation algorithm. We present the extended Circular Belief Propagation in the general case, that is, in a general factor graph. Additionally, we particularly investigate the case where variables are Gaussian, which leads to a very simple formulation of (Gaussian) Circular BP, very similar to the one in the binary case. This allows us to expand the neural model presented in chapter 4, which applied exclusively to pairwise Markov Random Fields with binary variables by addressing the Gaussian case, therefore expanding the restrictive binary case. In this neural model, a population encodes for the parameters of the distribution (mean and inverse variance), which requires twice as many units as in the binary case. The firing rate of units is proportional to the value they encode; one unit encodes for the inverse variance, and the other one encodes for the product between the mean and the inverse variance. As in the binary case, the rate model implements exactly eCBP, and could be translated into a spiking network as in the binary case.

Next, we describe the effects of circularity by simulating Gaussian Circular BP. Similarly to the binary case where circularity lead to overconfidence, estimates of probabilities are overconfident (as measured by the inverse variance or precision). However, and contrary to the binary case, the MAP estimation (as measured by the mean of the distribution) also gets strongly modified by circularity (it is also the case in the binary case, but only rarely). Simulations show that the means are in general overestimated.

Last, we include preliminary discussions on the comparison between Gaussian Circular BP and impaired predictive coding, which both model brain processes underlying the emergence of suboptimal behavior and psychosis, and are currently the two main types of Bayesian models in computational psychiatry.

5.2 Circular BP in general factor graphs

In previous chapters, we only considered pairwise factors graphs, that is, probability distributions $p(x)$ which could be written as the product of unitary and pairwise potentials: $p(\mathbf{x}) \propto \prod_{(i,j)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i)$. In this section, we present the general case.

Similarly to the pairwise factors case, Circular BP is defined as an approximation of Fractional BP, and extended Circular BP is defined based on extended Fractional BP (a generalized BP algorithm using the Kikuchi approximation; see Yedidia et al. (2001)).

Inference in graphical models A general Markov random field $p(\mathbf{x})$ can be written:

$$p(\mathbf{x}) \propto \prod_c \psi_c(\mathbf{x}_c) \tag{5.1}$$

Potentials ψ_c are called "factors" and are associated to a clique $\mathbf{x}_c$ ($\mathbf{x}_c$ is a group of variables $\{x_i\}$ or simply one variable x_i). As Figure 5.1 shows, the probability distribution can be represented graphically as a factor graph composed of variable nodes x_i and factor nodes ψ_c, with links between ψ_c and all the variable nodes in $\mathbf{x}_c$.

Belief Propagation Belief Propagation can be defined on a factor graph (Kschischang et al., 2001). At every iteration, the algorithm updates messages from factor nodes to variable nodes and from variable nodes to factor nodes:

$$m^{\text{new}}_{\psi_c \to x_i}(x_i) \propto \sum_{\mathbf{x}_c \setminus x_i} \psi_c(\mathbf{x}_c) \prod_{x_j \in \mathcal{N}(\psi_c) \setminus x_i} m_{x_j \to \psi_c}(x_j) \tag{5.2}$$

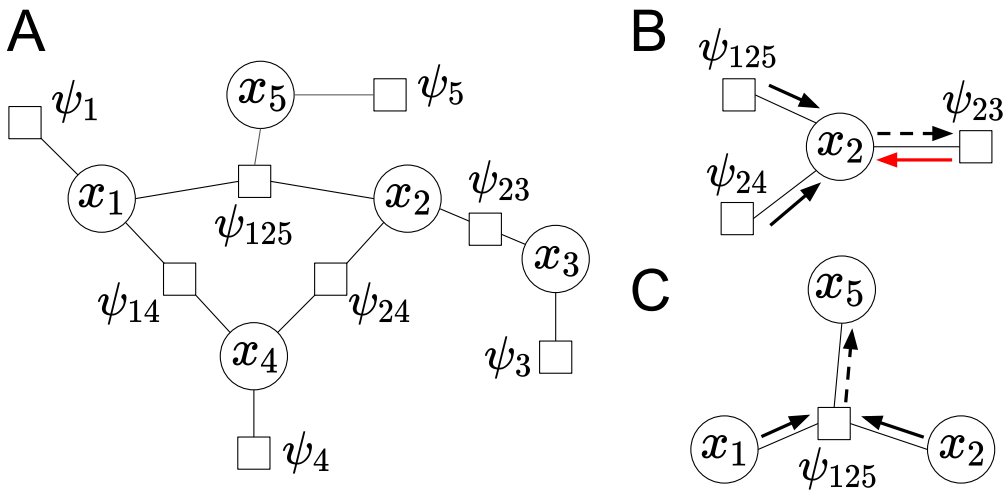

FIGURE 5.1: Belief Propagation and Circular Belief Propagation on a general factor graph. **(A)** Example of a factor graph with higher-order interactions (i.e., not only unitary and pairwise factors), here representing the probability distribution $p(\mathbf{x}) = \psi_{125}(x_1, x_2, x_5)\psi_{23}(x_2, x_3)\psi_{14}(x_1, x_4)\psi_{24}(x_2, x_4)\psi_1(x_1)\psi_3(x_3)\psi_4(x_4)\psi_5(x_5)$. **(B)** Belief Propagation updates a given variable-to-factor message (from x to ψ, dotted black line) according to the messages received by node x from other factor nodes than ψ (full black lines). For Circular BP, the opposite message (coming from ψ to x, full red line) is also taken into account. **(C)** Both BP and Circular BP update factor-to-variable messages (from ψ to x, dotted black line) according to the messages collected by ψ from other variable nodes than x (full black lines), and the interaction function ψ.

$$m^{\text{new}}_{x_j \to \psi_d}(x_j) \propto \psi_j(x_j) \prod_{\psi_c \in \mathcal{N}(x_j) \setminus \psi_d} m_{\psi_c \to x_j}(x_j) \tag{5.3}$$

where $\mathcal{N}(x)$ are the neighbors of node x in the factor graph. Neighbors of variable nodes are factor nodes, and reciprocally.

Once messages have converged (or at some given maximum iteration), approximate marginal probabilities or *beliefs* are computed as:

$$b_i(x_i) \propto \psi_i(x_i) \prod_{\psi_c \in \mathcal{N}(x_i)} m_{\psi_c \to x_i}(x_i) \tag{5.4}$$

A message $m_{\psi_c \to x_i}(x_i)$ from factor node to variable node correspond to probabilistic information about variable x_i collected by the factor node ψ_c. The message is based on the information available elsewhere in the network (observed variables, prior distribution over variables) received by ψ_c, and takes into account the probabilistic interactions between x_i and its neighbors (i.e., the interaction factor ψ_c).

A message $m_{x_j \to \psi_d}(x_j)$ from variable node to factor node is simply the sum (in the log-domain) of the local information at x_j (e.g., noisy observation or prior) with the messages received by x_j from all factors neighboring x_j except ψ_d.

Note that if factors are all pairwise (case considered in the main text: $\psi_c = \psi_{ij}$) then BP equations (5.2) and (5.3) can be written with messages going from factor to variable node only. We recover Equations (1.3) and (1.4) from the main text by defining $m_{i \to j}(x_j) \equiv m_{\psi_c \to x_j}(x_j)$.

Extended Fractional BP We consider here a modification of Belief Propagation based on a parametric approximation of the entropy of the approximating distribution $b(\mathbf{x})$, with the parameters $(\boldsymbol{\kappa}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. κ_i and γ_i are assigned to the variable node x_i, while α_c and β_c are assigned to the factor ψ_c. We obtain the following modified update equations (see following paragraph for the demonstration):

$$m^{\text{new}}_{\psi_c \to x_i}(x_i) \propto \Big(\sum_{\mathbf{x}_c \setminus x_i} \psi_c(\mathbf{x}_c)^{\beta_c \alpha_c} \prod_{x_j \in \mathcal{N}(\psi_c) \setminus x_i} m_{x_j \to \psi_c}(x_j) \Big)^{\kappa_i / \alpha_c} \tag{5.5}$$

$$m^{\text{new}}_{x_j \to \psi_d}(x_j) \propto \psi_j(x_j)^{\gamma_j \kappa_j} \left(\prod_{\psi_c \in \mathcal{N}(x_j) \setminus \psi_d} m_{\psi_c \to x_j}(x_j) \right) m_{\psi_d \to x_j}(x_j)^{1 - \alpha_c / \kappa_i} \tag{5.6}$$

and beliefs (approximate marginal probabilities) are computed using:

$$b_i(x_i) \propto \psi_i(x_i)^{\kappa_i} \prod_{\psi_c \in \mathcal{N}(x_i)} m_{\psi_c \to x_i}(x_i) \tag{5.7}$$

The special case where $(\boldsymbol{\kappa}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$ corresponds to BP.

Similarly to above, if factors are all pairwise (case considered in the main text) then extended Fractional BP equations (5.5) and (5.6) can be written with messages going from factor to variable node only. We recover Equations (A.10) and (3.19) from the main text by defining $m_{i \to j}(x_j) \equiv m_{\psi_c \to x_j}(x_j)$.

Theoretical background for eFBP Here we provide the theoretical foundations underlying the modification of BP given in Equations (5.5), (5.6), and (5.7).

Note that in the demonstration, we cover all the special cases of eFBP, including BP (for which $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$), as well as Fractional BP, Power EP and α-BP (which all correspond to $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$).

As stated in section A of the Appendix, the approach taken by BP to compute marginals $p(\mathbf{x})$, whose formula is known but whose marginals are hard to compute, is to approximate $p(\mathbf{x})$ with distribution $b(\mathbf{x})$ (called the *variational* distribution) whose marginals are easier to compute. The Gibbs free energy (that we would like to minimize) is given by:

$$G = U_b - S_b \tag{5.8}$$

where the variational average energy U_b can be computed easily:

$$\begin{aligned} U_b &= \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x}) \\ &= -\sum_{\mathbf{x}} b(\mathbf{x}) \sum_{\text{cliques } c} \psi_c(\mathbf{x}_c) - \sum_x b(\mathbf{x}) \sum_i \psi_i(x_i) \\ &= -\sum_{\text{cliques } c} \sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) \psi_c(\mathbf{x}_c) - \sum_i \sum_{x_i} b_i(x_i) \psi_i(x_i) \end{aligned} \tag{5.9}$$

, contrary to the variational entropy S_b.

As it is not possible to easily compute S_b, Belief Propagation estimates it as if the factor graph representing $b(x)$ was a tree (i.e., was acyclic). This means that:

$$b(\mathbf{x}) \approx \prod_{\text{cliques } c} \left(\frac{b_c(\mathbf{x}_c)}{\prod_{i \in \mathcal{N}(c)} b_i(x_i)} \right) \prod_{\text{nodes } i} b_i(x_i) \tag{5.10}$$

where $b_c(\mathbf{x}_c) \equiv \sum_{x \setminus \mathbf{x}_c} b(\mathbf{x})$ (where for instance $\mathbf{x}_c = (x_1, x_2)$) and $b_i(x_i) \equiv \sum_{x \setminus x_i} b(\mathbf{x})$.

The equation above can also be written:

$$b(\mathbf{x}) \approx \prod_{\text{cliques } c} b_c(\mathbf{x}_c) \prod_{\text{nodes } i} b_i(x_i)^{1-|\mathcal{N}(i)|}$$

where $|\mathcal{N}(i)|$ is the number of neighbors of node i in the graph representation of the distribution.

The approximation of $b(x)$ given in Equation (5.10) is equivalent to approximating the entropy S_b of $b(\mathbf{x})$ as follows:

$$\begin{aligned} -S_b &= \sum_{\mathbf{x}} b(\mathbf{x}) \log(b(\mathbf{x})) \\ &\approx \sum_{\mathbf{x}} b(\mathbf{x}) \sum_{\text{cliques } c} \log(b_c(\mathbf{x}_c)) + \sum_{\mathbf{x}} b(\mathbf{x}) \sum_{\text{nodes } i} (1 - |\mathcal{N}(i)|) \log(b_i(x_i)) \\ &\approx \sum_{c} \sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) \log(b_c(\mathbf{x}_c)) + \sum_{i} (1 - |\mathcal{N}(i)|) \sum_{x_i} b_i(x_i) \log(b_i(x_i)) \end{aligned} \tag{5.11}$$

In contrast, the extended Fractional BP algorithm consists of approximating the variational distribution $b(\mathbf{x})$ as:

$$b(\mathbf{x}) \approx \prod_{\text{cliques } c} \left(\frac{b_c(\mathbf{x}_c)}{\prod_{i \in \mathcal{N}(c)} b_i(x_i)} \right)^{1/\alpha_c} \prod_{\text{nodes } i} b_i(x_i)^{1/\kappa_i} \tag{5.12}$$

which can also be written:

$$b(\mathbf{x}) \approx \prod_{\text{cliques } c} b_c(\mathbf{x}_c)^{1/\alpha_c} \prod_{\text{nodes } i} b_i(x_i)^{1/\kappa_i - |\mathcal{N}(i)|/\alpha_i} \quad \text{where} \quad \frac{1}{\alpha_i} \equiv \frac{1}{|\mathcal{N}(i)|} \sum_{i \in \mathcal{N}(c)} \frac{1}{\alpha_c} \tag{5.13}$$

This leads to the following parametric approximation of the variational entropy:

$$\begin{aligned} -S_b &= \sum_{\mathbf{x}} b(\mathbf{x}) \log(b(\mathbf{x})) \\ &\approx \sum_{\mathbf{x}} b(\mathbf{x}) \sum_{\text{cliques } c} \frac{1}{\alpha_c} \log(b_c(\mathbf{x}_c)) + \sum_{\mathbf{x}} b(\mathbf{x}) \sum_{\text{nodes } i} \Big(\frac{1}{\kappa_i} - \frac{|\mathcal{N}(i)|}{\alpha_i} \Big) \log(b_i(x_i)) \\ &\approx \sum_{\text{cliques } c} \frac{1}{\alpha_c} \sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) \log(b_c(\mathbf{x}_c)) + \sum_{i} \Big(\frac{1}{\kappa_i} - \frac{|\mathcal{N}(i)|}{\alpha_i} \Big) \sum_{x_i} b_i(x_i) \log(b_i(x_i)) \end{aligned} \tag{5.14}$$

Hence the following approximation of the Gibbs free energy $G \approx G_{\text{approx}}$ (we recover the Bethe free energy of BP with $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$):

$$\begin{aligned} G_{\text{approx}} = \sum_{\text{cliques } c} \frac{1}{\alpha_c} \sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) \log \left(\frac{b_c(\mathbf{x}_c)}{\prod_{i \in \mathcal{N}(c)} b_i(x_i)} \right) - \sum_{\text{cliques } c} \beta_c \sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) \log \Big(\psi_c(\mathbf{x}_c) \Big) \\ + \sum_{i} \frac{1}{\kappa_i} \sum_{x_i} b_i(x_i) \log \Big(b_i(x_i) \Big) - \sum_{i} \gamma_i \sum_{x_i} b_i(x_i) \log \Big(\psi_i(x_i) \Big) \end{aligned} \tag{5.15}$$

From Gibbs free energy to messages In what follows, we derive the message-passing update equations of extended Fractional BP (5.5), (5.6), and (5.7) (and its special cases BP, Fractional BP, Power EP and α-BP, among others) based on the approximation of the Gibbs Free Energy G_{approx} given in Equation (5.15). The message update equations are simply fixed-point equations of G_{approx}, meaning that the beliefs computed by Balanced BP are stationary points of the approximate Gibbs free energy G_{approx}. This demonstration is similar to the one for BP (Yedidia et al., 2001, 2003) and Fractional BP (Wiegerinck and Heskes, 2002), with additional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$.

We form the Lagrangian by adding Lagrange multipliers to G_{approx}. Lagrange multiplier μ_i (resp. μ_c) corresponds to the normalization constraint $\sum_{x_i} b_i(x_i) = 1$ (resp. $\sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) = 1$), while $\lambda_c(x_j)$ corresponds to the marginalization constraint $\sum_{\mathbf{x}_c \setminus x_j} b_c(\mathbf{x}_c) = b_j(x_j)$. We obtain:

$$\begin{aligned}\mathcal{L} = G_{\text{approx}} + \sum_i \mu_i \Big(\sum_{x_i} b_i(x_i) - 1 \Big) + \sum_{\text{cliques} c} \mu_c \Big(\sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) - 1 \Big) \\ + \sum_{\text{cliques } c} \sum_{x_j} \lambda_c(x_j) \Big(\sum_{\mathbf{x}_c \setminus x_j} b_c(\mathbf{x}_c) - b_j(x_j) \Big)\end{aligned} \tag{5.16}$$

The partial derivatives of the Lagrangian are:

$$\frac{\partial \mathcal{L}}{\partial b_i(x_i)} = - \sum_{c \in \mathcal{N}(i)} \frac{1}{\alpha_c} + \frac{1}{\kappa_i} + \frac{1}{\kappa_i} \log(b_i(x_i)) - \log(\psi_i(x_i)) + \mu_i - \sum_{c \in \mathcal{N}(i)} \lambda_c(x_i) \tag{5.17}$$

$$\frac{\partial \mathcal{L}}{\partial b_c(\mathbf{x}_c)} = \frac{1}{\alpha_c} + \frac{1}{\alpha_c} \log \left(\frac{b_c(\mathbf{x}_c)}{\prod_{i \in \mathcal{N}(c)} b_i(x_i)} \right) - \log(\psi_c(\mathbf{x}_c)) + \sum_{i \in \mathcal{N}(c)} \lambda_c(x_i) + \mu_c \tag{5.18}$$

It comes, by cancelling the partial derivatives of the Lagrangian:

$$b_i(x_i) \propto \psi_i(x_i)^{\kappa_i} \prod_{c \in \mathcal{N}(i)} \exp \Big(\kappa_i \lambda_c(x_i) \Big) \tag{5.19}$$

and

$$b_c(\mathbf{x}_c) \propto \Big(\prod_{i \in \mathcal{N}(c)} b_i(x_i) \Big) \psi_c(\mathbf{x}_c)^{\alpha_c} \prod_{i \in \mathcal{N}(c)} \exp \Big(- \alpha_c \lambda_c(x_i) \Big) \tag{5.20}$$

$$\begin{aligned}\implies b_c(\mathbf{x}_c) \propto \psi_c(\mathbf{x}_c)^{\alpha_c} \Big(\prod_{i \in \mathcal{N}(c)} \psi_i(x_i)^{\kappa_i} \Big) \Big(\prod_{i \in \mathcal{N}} \prod_{\tilde{c} \in \mathcal{N}(i) \setminus c} \exp \Big(\kappa_i \lambda_{\tilde{c}}(x_i) \Big) \Big) \\ \times \Big(\prod_{i \in \mathcal{N}(c)} \exp \Big(\lambda_c(x_i) \Big(\kappa_i - \alpha_c \Big) \Big) \Big)\end{aligned} \tag{5.21}$$

We obtain, for $m_{c \to i}(x_i) \equiv \exp \Big(\kappa_i \lambda_c(x_i) \Big)$, the following expression of the (approximate) marginal and pairwise beliefs:

$$b_i(x_i) \propto \psi_i(x_i)^{\kappa_i} \prod_{c \in \mathcal{N}(i)} m_{c \to i}(x_i) \tag{5.22}$$

$$b_c(\mathbf{x}_c) \propto \psi_c(\mathbf{x}_c)^{\alpha_c} \left(\prod_{i \in \mathcal{N}(c)} \psi_i(x_i)^{\kappa_i} \right) \left(\prod_{i \in \mathcal{N}} \prod_{\tilde{c} \in \mathcal{N}(i) \setminus c} m_{\tilde{c} \to i}(x_i) \right) \left(\prod_{i \in \mathcal{N}(c)} m_{c \to i}(x_i)^{1-\alpha_c/\kappa_i} \right) \tag{5.23}$$

Eventually, thanks to the constraint $\sum_{\mathbf{x}_c \setminus x_j} b_c(\mathbf{x}_c) = b_j(x_j)$, we obtain the fixed point equations for the messages:

$$m_{c \to j}(x_j) \propto \left(\sum_{\mathbf{x}_c \setminus x_j} \psi_c(\mathbf{x}_c)^{\alpha_c} \prod_{i \in \mathcal{N}(c) \setminus j} \psi_i(x_i)^{\kappa_i} \prod_{i \in \mathcal{N}(c) \setminus j} \prod_{\tilde{c} \in \mathcal{N}(i) \setminus c} m_{\tilde{c} \to i}(x_i) \right.$$
$$\left. \prod_{i \in \mathcal{N}(c) \setminus j} m_{c \to i}(x_i)^{1-\alpha_c/\kappa_i} \right) m_{c \to j}(x_j)^{1-\alpha_c/\kappa_j} \tag{5.24}$$

$$\Leftrightarrow m_{c \to j}(x_j) \propto \left(\sum_{\mathbf{x}_c \setminus x_j} \psi_c(\mathbf{x}_c)^{\alpha_c} \prod_{i \in \mathcal{N}(c) \setminus j} \psi_i(x_i)^{\kappa_i} \prod_{i \in \mathcal{N}(c) \setminus j} \prod_{\tilde{c} \in \mathcal{N}(i) \setminus c} m_{\tilde{c} \to i}(x_i) \right.$$
$$\left. \prod_{i \in \mathcal{N}(c) \setminus j} m_{c \to i}(x_i)^{1-\alpha_c/\kappa_i} \right)^{\kappa_j/\alpha_c} \tag{5.25}$$

The extended Fractional BP algorithm consists of running iteratively the fixed-point equation (5.25). This single equation, which involves only messages from factor node to variable node, can be rewritten into the two following equations by introducing messages from variable node to factor node:

$$m^{\text{new}}_{\psi_c \to x_i}(x_i) \propto \Big(\sum_{\mathbf{x}_c \setminus x_i} \psi_c(\mathbf{x}_c)^{\alpha_c} \prod_{x_j \in \mathcal{N}(\psi_c) \setminus x_i} m_{x_j \to \psi_c}(x_j) \Big)^{\kappa_i/\alpha_c} \tag{5.26}$$

$$m^{\text{new}}_{x_j \to \psi_d}(x_j) \propto \psi_j(x_j)^{\kappa_j} \left(\prod_{\psi_c \in \mathcal{N}(x_j) \setminus \psi_d} m_{\psi_c \to x_j}(x_j) \right) m_{\psi_d \to x_j}(x_j)^{1-\alpha_c/\kappa_i} \tag{5.27}$$

The expression of the unitary beliefs is given by Equation (5.22).

Note that one could also use directly Equation (5.24) instead of (5.25) to define the extended Fractional BP algorithm. In fact, Equations (5.24) and (5.25) correspond to the damped versus undamped update equation; see section 4.2.1 about damping. There is no absolute better choice: fixed points obtained are identical in both cases, and damping provides better convergence properties but slows down the system (see section 4.2.1).

Comparison with related models: BP, Fractional BP, Power EP, alpha BP, and Circular BP The special case of Belief Propagation is recovered for $(\boldsymbol{\alpha}, \boldsymbol{\kappa}) = (\mathbf{1}, \mathbf{1})$.

Fractional BP (Wiegerinck and Heskes, 2002), Power EP (Minka, 2004) and α-BP (Liu et al., 2019) use the damped message update equation (A.8) (with $\boldsymbol{\kappa} = \mathbf{1}$) rather than its undamped version (A.9) (see section 4.2.1). .

Circular BP (Jardri and Denève, 2013a) also considers $\boldsymbol{\kappa} = \mathbf{1}$, and modifies the message update equation from variable to factor (Equation (5.3) for BP) into:

$$m_{x_j \to \psi_d}(x_j) \propto \psi_j(x_j) \left(\prod_{\psi_c \in \mathcal{N}(x_j) \setminus \psi_d} m_{\psi_c \to x_j}(x_j) \right) m_{\psi_d \to x_j}(x_j)^{1-\alpha_{x_j \to \psi_d}} \tag{5.28}$$

Note that Circular BP was only defined on pairwise factor graphs, meaning that all cliques c have at most 2 elements: the product of Equation (5.27) does not appear in this case.

Notably, extended Fractional BP, BP, Fractional BP and Circular BP (respectively Equations (3.20a), (1.17a), (3.5a) and (1.19a)) are similar but differ in the number of degrees of liberty. BP has no degree of liberty. Fractional BP (as well as Power EP and α-BP) has n_{factors} degrees of liberty. Balanced BP has $n_{\text{factors}} + n_{\text{variables}}$ degrees of liberty. Circular BP has as as many degrees of liberty as the number of edges in the factor graph.

Extended Circular BP on general factor graphs Similarly to the pairwise factor case, we can define an extended Circular BP on general factor graphs, based on the modification of Equation (5.25) into:

$$m_{c\to j}(x_j) \propto \left(\sum_{\mathbf{x}_c \setminus x_j} \psi_c(\mathbf{x}_c) \prod_{i\in\mathcal{N}(c)\setminus j} \psi_i(x_i)^{\kappa_i} \prod_{i\in\mathcal{N}(c)\setminus j} \prod_{\tilde{c}\in\mathcal{N}(i)\setminus c} m_{\tilde{c}\to i}(x_i) \prod_{i\in\mathcal{N}(c)\setminus j} m_{c\to i}(x_i)^{1-\alpha_c/\kappa_i} \right)^{\kappa_j} \tag{5.29}$$

In this case of general factor graphs, contrary to the pairwise case, there is no simple way of writing the message update in the log domain, making it unclear how this algorithm could be implemented in the brain.

5.3 Gaussian Circular BP

In all this thesis with the exception of the present section, variables x_i of the probability distribution $p(x_1, \dots, x_n)$ are assumed to be binary variables: $x_i \in \{-1; +1\}$. Equations (3.26a) and (3.26b) relate (the log-transforms of) the means of the belief distribution to the ones of the message distribution. These relations were used to propose a neural implementation of extended Circular BP in the log-domain (see sections 4.2 and 4.3).

However, in our day-to-day lives, humans often deal with continuous variables. For instance, before taking the bus, one might want to estimate the time to reach a certain destination, in order to make sure to arrive there on time. This estimate is based on external information. In this example, the external information is for instance the level of congestion on the road this day, the probability of having to wait for several buses in case they are full, and the average time between two bus arrivals.

In this section, we consider instead the Gaussian case, which is an example of continuous distribution commonly considered. We show that the extended Circular BP algorithm translates into relations between the parameters (mean and variance) of the beliefs and the parameters (mean and variance) of the messages, very similarly to the binary case. This allows us to propose a neural implementation of extended Circular BP in the Gaussian case, where each population encodes for the parameters of the distribution (here, the mean of the gaussian distribution, and the product of the precision and the mean, by using one neuron for each) as in the binary case. This goes along the hypothesis according to which neural responses might represent the parameters of the posterior probability distribution as in probabilistic population codes (PPCs) and their predecessors, kernel density estimator codes and distributional population codes; see Fiser et al. (2010). We describe the corresponding rate network, which highly resembles the one proposed in the binary case. A spiking neuron could be constructed as well, following the same technique for the binary case described in section 4.3.

5.3.1 Circular BP in Gaussian Markov Random Fields

5.3.1.1 Gaussian Markov Random Fields

We consider in this section a Gaussian Markov Random Field, which is by definition a Markov Random Field (MRF) in which the joint distribution is Gaussian:

$$p(x) = \sqrt{\frac{|P|}{(2\pi)^n}} \exp\Big(-\frac{1}{2}(x-\mu)^T P (x-\mu)\Big) \tag{5.30}$$

and is commonly used in the fields of computer vision and sensor networks, for instance. The distribution is parametrized by its mean $\mu = \mathbb{E}[x]$ and its *precision matrix* P. P is the inverse covariance matrix: $P = \Sigma^{-1}$ where the covariance matrix is $\Sigma = \mathbb{E}[(x-\mu)(x-\mu)^T]$. Matrix P is also known as the *information matrix*, as its sparsity exactly matches the corresponding graphical model: $P_{ij} = 0$ indicates an absence of edge between nodes i and j in the graph.

We use a slightly different parametrization of the Gaussian distribution with its natural parameters (P, v) where $v \equiv P\mu$ is often called the *potential vector* (Raju and Pitkow, 2016), instead of using parameters P and μ:

$$p(x) \propto \exp\Big(-\frac{1}{2} x^T P x + v^T x\Big) \tag{5.31}$$

We assume, as for the binary case, that the probability distribution can be written as a factor of pairwise interactions and unitary interactions:

$$p(x) = \frac{1}{Z} \prod_{(i,j)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i) \tag{5.32}$$

where potentials have the following Gaussian expression (see section 5.3.1.4):

$$\begin{cases} \psi_{ij}(x_i, x_j) = \exp\Big(-\frac{1}{2} \begin{pmatrix} x_i \\ x_j \end{pmatrix}^T P_{x_i,x_j} \begin{pmatrix} x_i \\ x_j \end{pmatrix} \Big) & (5.33) \\ \psi_i(x_i) = \exp\Big(-\frac{1}{2} P_{\text{ext}\to i} x_i + v_i x_i \Big) & (5.34) \end{cases}$$

where $P_{x_i,x_j} \equiv \begin{pmatrix} P^{x_i,x_i}_{x_i,x_j} & P^{x_i,x_j}_{x_i,x_j} \\ P^{x_j,x_i}_{x_i,x_j} & P^{x_j,x_j}_{x_i,x_j} \end{pmatrix}$ is a 2×2 matrix.

Note that Equation 5.33 is equivalent to writing $x_i = Cx_j + N$ where N is Gaussian noise with zero mean (with $C = -P^{x_i,x_j}_{x_i,x_j} / P^{x_i,x_i}_{x_i,x_j}$ and precision $P^{x_i,x_i}_{x_i,x_j}$ for N).

5.3.1.2 BP in Gaussian graphical models

The Belief Propagation algorithm has been studied empirically and theoretically in the Gaussian case specifically; see for instance Malioutov (2008). It is also the case for extensions of BP like Expectation Propagation (Minka, 2001b,a) or Fractional Belief Propagation (Cseke and Heskes, 2011; Liu et al., 2020). When applied on probabilistic graphs with cycles, Belief Propagation often performs reasonably well. As shown in Weiss and Freeman (1999), when the algorithm converges it produces the correct means, but generally not the correct variances. A nice example of application of Belief Propagation in Gaussian Markov Random Fields is provided in Weiss (1997), where BP is used to estimate the direction of motion automatically given a video of a hand moving.

5.3.1.3 Formulation of Circular BP in Gaussian MRF

In the case described in the above section 5.3.1.1, marginals $\{p_i(x_i)\}$ of the distribution can be approximated using BP and its extension, the extended Circular BP algorithm. Extended Circular BP consists of very simple update equations on the parameters of the distribution, by sending (as in the binary case) messages between nodes of the graph to spread information to all nodes. The relations between the mean μ_i (or rather the potential vector $v_i \equiv P_i\mu_i$) and precision P_i of the beliefs and the mean $\mu_{i\to j}$ and precision $P_{i\to j}$ of the messages are the following (see the following section for a demonstration):[1]

$$\begin{cases} P_{i\to j}^{\text{new}} = g_{ij}(P_i - \alpha_{ij}P_{j\to i}) & (5.35) \\ v_{i\to j}^{\text{new}} = h_{ij}\Big(v_i - \alpha_{ij}v_{j\to i},\, P_i - \alpha_{ij}P_{j\to i}\Big) & (5.36) \end{cases}$$

where

$$\begin{cases} P_i = \kappa_i\Big(\sum\limits_{j\in N(i)} P_{j\to i} + \gamma_i P_{\text{ext}\to i}\Big) & (5.37) \\ v_i = \kappa_i\Big(\sum\limits_{j\in N(i)} v_{j\to i} + \gamma_i v_{\text{ext}\to i}\Big) & (5.38) \end{cases}$$

and where functions g_{ij} and h_{ij} are given by

$$g_{ij}(y) = \beta_{ij}P^{x_j,x_j}_{x_i,x_j} - \frac{(\beta_{ij}P^{x_i,x_j}_{x_i,x_j})^2}{\beta_{ij}P^{x_i,x_i}_{x_i,x_j} + y} \tag{5.39}$$

and

$$h_{ij}(x,y) = -\frac{\beta_{ij}P^{x_i,x_j}_{x_i,x_j}x}{\beta_{ij}P^{x_i,x_i}_{x_i,x_j} + y} \tag{5.40}$$

These equations logically extend the ones obtained for BP (see for instance Weiss and Freeman (1999); Plarre and Kumar (2004)) for which parameters $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ take value $\mathbf{1}$.

Equations (5.35) and (5.36), which relate the parameters $(P_{i\to j},\, v_{i\to j})$ of the message distribution to the parameters $(P_i,\, v_i)$ of the marginal distribution and the parameters $(P_{j\to i},\, v_{j\to i})$ of the opposite message distribution, have exactly the same form as in the binary case, with the subtraction of the "belief variable" by the "opposite message variable" with weight α_{ij}. Indeed, in the binary case, the evolution of the only parameter of the message distribution (its expected value $m_{i\to j}(x_j = +1)$, or rather, the transform $M_{i\to j} \equiv \frac{1}{2}\log\left(\frac{m_{i\to j}(x_j=+1)}{m_{i\to j}(x_j=-1)}\right)$) is expressed as a function of a quantity associated to the marginal distribution $B_i \equiv \frac{1}{2}\log\left(\frac{b_i(x_i=+1)}{b_i(x_i=-1)}\right)$ and the parameter of the opposite message distribution $M_{j\to i}$ weighted by α_{ij} as $M_{i\to j} = f_{ij}(B_i - \alpha_{ij}M_{j\to i})$. In both the Gaussian and the binary case, the quantity associated to the marginal distribution gets subtracted by α_{ij} times the quantity associated to the message distribution (for the message going in the opposite direction). As a reminder, the sign of function f_{ij} is determined by the interaction between x_i and x_j ($J_{ij} > 0$ indicates a positive interaction) and the sign of its argument. Similarly, the sign of h_{ij} depends on the sign of its first argument, as well as the sign of the interaction between variables x_i and x_j (a positive interaction means that to $P^{x_i,x_j}_{x_i,x_j} < 0$ as the precision is the inverse covariance). Function g_{ij} is positive as precision matrices are positive definite and symmetrical matrices by definition. g_{ij} is an increasing function of its argument $P_i - \alpha_{ij}P_{j\to i}$. The absolute value of function h_{ij}, $\big|h_{ij}\big|$, increases as function of its first argument $v_i - \alpha_{ij}v_{j\to i}$ and decreases as function its second argument $P_i - \alpha_{ij}P_{j\to i}$.

[1] Note that all the parameters in Equations (5.35)-(5.40) are scalars.

Likewise, Equations (5.37) and (5.38) giving the expression of the (transform of the) marginal distribution parameter are similar to the expression in the binary case $B_i = \kappa_i\Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big)$. The variable of interest associated to node i (function of the parameter(s) of the $b_i(x_i)$) is the sum of all the contributions from its neighbors and from the external world.

Running these equations iteratively (see Algorithm 5) allows one, as in the binary case, to marginalize the distribution. In the Gaussian case, this means obtaining estimates of the mean and variance of the variables $\{x_i\}$, given the external information $\{\psi_i(x_i)\}$ (through $\{P_{\text{ext}\to i}\}$ and $\{\mu_{\text{ext}\to i}\}$) and the conditional dependencies between variables $\{\psi_{ij}(x_i, x_j)\}$ (through $\{P_{x_i,x_j}\}$).

We also write an equivalent formulation of the algorithm, which first consists of running the updates on precisions and only then on potential vectors (or equivalently, on means). Indeed, in the equations above, precisions $\{P_i\}$ and $\{P_{i\to j}\}$ do not depend on the potentials $\{v_i\}$ and $\{v_{i\to j}\}$. Thefore, precisions can be computed at first and fixed once for all, and the "real" algorithm would consist in running the updates on the means with potentially time-varying inputs $\{\mu_{\text{ext}\to i}\}$. By writing $\mu = (P\mu)/P = v/P$, we obtain after some mathematical manipulations the following update equations on the means:

$$\begin{cases} \mu_i = \displaystyle\sum_{j\in N(i)} k_{j\to i}\mu_{j\to i} + k_{\text{ext}\to i}\mu_{\text{ext}\to i} & (5.41) \\ \mu_{i\to j}^{\text{new}} = h_i\mu_i - h_{j\to i}\mu_{j\to i} & (5.42) \end{cases}$$

with coefficients $k_{j\to i} \equiv \dfrac{P_{j\to i}}{\displaystyle\sum_{j\in N(i)} P_{j\to i} + \gamma_i P_{\text{ext}\to i}}$, $k_{\text{ext}\to i} \equiv \dfrac{\gamma_i P_{\text{ext}\to i}}{\displaystyle\sum_{j\in N(i)} P_{j\to i} + \gamma_i P_{\text{ext}\to i}}$, $h_i \equiv \dfrac{-P_{x_i,x_j}^{x_i,x_j} P_i}{\beta_{ij}|P_{x_i,x_j}| + P_{x_i,x_j}^{x_j,x_j}(P_i - \alpha_{ij}P_{j\to i})}$, and $h_{j\to i} \equiv \dfrac{-P_{x_i,x_j}^{x_i,x_j}\alpha_{ij}P_{j\to i}}{\beta_{ij}|P_{x_i,x_j}| + P_{x_i,x_j}^{x_j,x_j}(P_i - \alpha_{ij}P_{j\to i})}$. Coefficients $\{k_{j\to i}\}$ and $k_{\text{ext}\to i}$ are positive and sum to 1. Coefficients h_i and $h_{j\to i}$ can be positive or negative, and their difference or sum is not equal to 1 in general (however, they have identical signs for $\alpha_{ij} > 0$ which is a reasonable hypothesis, and they are positive if x_i and x_j positively interact, that is, $P_{x_i,x_j}^{x_i,x_j} < 0$, and for a "reasonable" choice of parameters). Equations (5.41) and (5.42) would be easy to implement in a network in which the firing rate of neurons is proportional to the mean of the encoded variable, if precisions were implicitly encoded in the recurrent connections and input weights. However, the way connection strengths ($k_{j\to i}, k_{\text{ext}\to i}$, h_i and $h_{j\to i}$) are set up is not clear given the complexity of formulas providing the expression of coefficients $k_{j\to i}, k_{\text{ext}\to i}$, h_i and $h_{j\to i}$. Additionally, precisions ($\{P_{j\to i}\}$ and most importantly P_i) would not be encoded anywhere in the network, which seems necessary for the brain to have access to the estimation of marginal probabilities (defined by the mean and the precision) and take decisions based on that. Therefore, a direct implementation of Equations (5.41) and (5.42) is not considered here.

5.3.1.4 Demonstration

Mean and precision of messages We start by writing the message update equation for extended Circular Belief Propagation (the integral replaces the sum as x takes continuous values):

$$m_{i\to j}(x_j) \propto \int_{x_i} \psi_{i,j}(x_i, x_j)^{\beta_{ij}} \left[\prod_{k\in N(i)\setminus j} m_{k\to i}(x_i) m_{y_i\to x_i}(x_i)^{\gamma_i}\Big(m_{j\to i}(x_i)\Big)^{1-\alpha_{ij}/\kappa_i}\right]^{\kappa_i} dx_i \quad (5.43)$$

Here we hypothesize that messages have a normal distribution. We use, by slight abuse of notation, $\mathcal{N}$ to denote the density of the normal distribution (and not the law): $\mathcal{N}_x(\mu, P) \equiv$

Algorithm 5 Gaussian extended Circular BP algorithm in a pairwise factor graph

for all directed edges $i \to j$ **do**
  $P_{i\to j} \leftarrow$ random positive value {Initialize the messages' precisions}
  $v_{i\to j} \leftarrow$ random value {Initialize the messages' potentials}
end for
repeat
  for all nodes x_i **do**
    $P_i = \kappa_i\Big(\sum_{j\in N(i)} P_{j\to i} + \gamma_i P_{\text{ext}\to i}\Big)$ {Compute the beliefs' precisions}
    $v_i = \kappa_i\Big(\sum_{j\in N(i)} v_{j\to i} + \gamma_i v_{\text{ext}\to i}\Big)$ {Compute the beliefs' potentials}
  end for
  for all directed edges $x_i \to x_j$ **do**
    $P^{\text{new}}_{i\to j} = g_{ij}(P_i - \alpha_{ij}P_{j\to i})$ {Update the messages' precisions}
    $v^{\text{new}}_{i\to j} = h_{ij}\Big(v_i - \alpha_{ij}v_{j\to i},\, P_i - \alpha_{ij}P_{j\to i}\Big)$ {Update the messages' potentials}
  end for
  $P \leftarrow P^{\text{new}}$
  $\mu \leftarrow \mu^{\text{new}}$
until convergence
for all nodes x_i **do**
  $b_i(x_i) \leftarrow \sqrt{\frac{P_i}{2\pi}}\exp\Big(-\frac{1}{2}P_i x_i^2 + v_i x_i\Big)$ {Compute the beliefs}
end for

$\sqrt{|P|/(2\pi)} \times \exp\Big(-\frac{1}{2}(x-\mu)^T P(x-\mu)\Big)$ where μ is the mean of the distribution and $P = \Sigma^{-1}$ is the precision matrix or inverse covariance matrix (note that here we consider for simplicity that all variables x_i have dimension 1, thus $\mu_{i\to j}, \mu_i$, $P_{i\to j}$ and P_i are scalars). The message distribution is thus written:

$$m_{i\to j}(x_j) = \mathcal{N}_{x_j}(\mu_{i\to j},\, P_{i\to j}) \tag{5.44}$$

We also hypothesize that the potentials have a normal distribution. As in Weiss and Freeman (1999), we assume, without loss of generality, that the joint means are zero:

$$\begin{cases} \psi_{ij}(x_i, x_j) = \mathcal{N}_{(x_i,x_j)}(0,\, P_{x_i,x_j}) & (5.45) \\ \psi_i(x_i) = m_{y_i\to x_i}(x_i) = \mathcal{N}_{(x_i,y_i)}(0, P_{x_i,y_i}) \propto \mathcal{N}_{x_i}(\mu_{x_i|y_i},\, P_{x_i|y_i}) & (5.46) \end{cases}$$

where y_i is a noisy observation related to x_i only, P_{x_i,x_j} and P_{x_i,y_i} are 2x2 matrices, and the parameters of the Gaussian distribution in Equation (5.46) are given by $P^{x_i,x_i}_{x_i,y_i} \times \mu_{x_i|y_i} = -P^{x_i,y_i}_{x_i,y_i} \times y_i$ and $P_{x_i|y_i} = P^{x_i,x_i}_{x_i,y_i}$ (conditional distribution). In the following, $P_{x_i|y_i}$ is written $P_{\text{ext}\to i}$ and $\mu_{x_i|y_i}$ is written $\mu_{\text{ext}\to i}$.

$$\begin{aligned} \implies m_{i\to j}(x_j) \propto \int_{x_i} \mathcal{N}_{(x_i,x_j)}(0, P_{x_i,x_j})^{\beta_{ij}} \Bigg[\prod_{k\in N(i)\setminus j} \mathcal{N}_{x_i}(\mu_{k\to i}, P_{k\to i}) \mathcal{N}_{x_i}(\mu_{\text{ext}\to i}, P_{\text{ext}\to i})^{\gamma_i} \\ \times \mathcal{N}_{x_i}(\mu_{j\to i}, P_{j\to i})^{1-\alpha_{ij}/\kappa_i} \Bigg]^{\kappa_i} dx_i \end{aligned} \tag{5.47}$$

Using the fact that $\big(\mathcal{N}_x(\mu,P)\big)^k \propto \mathcal{N}_x(\mu, P\times k)$, it comes:

$$m_{i\to j}(x_j) \propto \int_{x_i} \mathcal{N}_{(x_i,x_j)}(0,\tilde{P}_{x_i,x_j})\Bigg[\prod_{k\in N(i)\setminus j} \mathcal{N}_{x_i}(\mu_{k\to i},P_{k\to i})\mathcal{N}_{x_i}(\mu_{\text{ext}\to i},\tilde{P}_{\text{ext}\to i})$$
$$\tilde{P}_{x_i,x_j}\begin{pmatrix}x_i\\x_j\end{pmatrix}\Big)\times \mathcal{N}_{x_i}(\mu_{j\to i},(1-\alpha_{ij}/\kappa_i)P_{j\to i})\Bigg]^{\kappa_i} dx_i \tag{5.48}$$

where by definition,

$$\begin{cases}\tilde{P}_{\text{ext}\to i} = \gamma_i P_{\text{ext}\to i} & \text{(5.49a)}\\ \tilde{P}_{x_i,x_j} = \beta_{ij} P_{x_i,x_j} & \text{(5.49b)}\end{cases}$$

It comes:

$$m_{i\to j}(x_j) \propto \int_{x_i} \exp\Big(-\frac{1}{2}\begin{pmatrix}x_i\\x_j\end{pmatrix}^T \tilde{P}_{x_i,x_j}\begin{pmatrix}x_i\\x_j\end{pmatrix}\Big)\mathcal{N}_{x_i}(\tilde{\mu}_i,\tilde{P}_i)dx_i \tag{5.50}$$

where

$$\begin{cases}\tilde{P}_i\tilde{\mu}_i = \kappa_i\Big(\sum_{k\in N(i)\setminus j} P_{k\to i}\mu_{k\to i} + \tilde{P}_{\text{ext}\to i}\mu_{\text{ext}\to i} + \Big(1-\frac{\alpha_{ij}}{\kappa_i}\Big)P_{j\to i}\mu_{j\to i}\Big) & \text{(5.51a)}\\ \tilde{P}_i = \kappa_i\Big(\sum_{k\in N(i)\setminus j} P_{ki} + \tilde{P}_{\text{ext}\to i} + \Big(1-\frac{\alpha_{ij}}{\kappa_i}\Big)P_{j\to i}\Big) & \text{(5.51b)}\end{cases}$$

(as a product of Gaussian densities: precisions P add up, and precisions multiplied by means $P\mu$ add up as well).

We now define

$$\begin{cases}P_0^{i,j} = \tilde{P}^{x_i,x_i}_{x_i,y_i} + \sum_{k\in N(i)\setminus j} P_{k\to i} & \text{(5.52a)}\\ P_0^{i,j}\mu_0^{i,j} = \tilde{P}_{\text{ext}\to i}\mu_{\text{ext}\to i} + \sum_{k\in N(i)\setminus j} P_{k\to i}\mu_{k\to i} & \text{(5.52b)}\end{cases}$$

It then comes:

$$\begin{cases}\tilde{P}_i\tilde{\mu}_i = \kappa_i\Big(P_0^{i,j}\mu_0^{i,j} + \Big(1-\frac{\alpha_{ij}}{\kappa_i}\Big)P_{j\to i}\mu_{j\to i}\Big) & \text{(5.53a)}\\ \tilde{P}_i = \kappa_i\Big(P_0^{i,j} + \Big(1-\frac{\alpha_{ij}}{\kappa_i}\Big)P_{j\to i}\Big) & \text{(5.53b)}\end{cases}$$

Finally, we write the result of Equation (5.50): $m_{i\to j}(x_j)$ is the integral over x_i of the product between a bivariate Gaussian density $\mathcal{N}_{(x_i,x_j)}(0,\tilde{P}_{x_i,x_j})$ and a univariate Gaussian density $\mathcal{N}_{x_i}(\tilde{\mu}_i,\tilde{P}_i)$, thus the message $m_{i\to j}(x_j)$ has a normal distribution $\mathcal{N}_{x_j}(\mu_{i\to j},P_{i\to j})$ with the following mean and precision:

$$\begin{cases}P_{i\to j} = \dfrac{\tilde{\Sigma}_{x_i,x_j,11} + \frac{1}{\tilde{P}_i}}{|\tilde{\Sigma}_{x_i,x_i}| + \tilde{\Sigma}_{x_i,x_j,22}\frac{1}{\tilde{P}_i}} & \text{(5.54)}\\ \mu_{i\to j} = \tilde{\mu}_i\dfrac{\tilde{\Sigma}_{x_i,x_j,12}}{\frac{1}{\tilde{P}_i}+\Sigma_{x_i,x_j,11}} & \text{(5.55)}\end{cases}$$

It comes:

$$\begin{cases} P_{i\to j} = \dfrac{\frac{\tilde{P}_{x_i,x_j,22}}{|\tilde{P}_{x_i,x_j}|} + \frac{1}{\tilde{P}_i}}{\frac{1}{|\tilde{P}_{x_i,x_j}|} + \frac{\tilde{P}_{x_i,x_j,11}}{|\tilde{P}_{x_i,x_j}|}\frac{1}{\tilde{P}_i}} = \dfrac{\tilde{P}_{x_i,x_j,22}\tilde{P}_i + |\tilde{P}_{x_i,x_j}|}{\tilde{P}_i + \tilde{P}_{x_i,x_j,11}} = \tilde{P}_{x_i,x_j,22} - \dfrac{(\tilde{P}_{x_i,x_j,12})^2}{\tilde{P}_i + \tilde{P}_{x_i,x_j,11}} \\ \mu_{i\to j} = \tilde{P}_i\tilde{\mu}_i \dfrac{-\frac{\tilde{P}_{x_i,x_j,21}}{|\tilde{P}_{x_i,x_j}|}}{1 + \tilde{P}_i\frac{\tilde{P}_{x_i,x_j,22}}{|\tilde{P}_{x_i,x_j}|}} = -\tilde{P}_i\tilde{\mu}_i\dfrac{\tilde{P}_{x_i,x_j,21}}{|\tilde{P}_{x_i,x_j}| + \tilde{P}_i\tilde{P}_{x_i,x_j,22}} = -\tilde{P}_i\tilde{\mu}_i\dfrac{\tilde{P}_{x_i,x_j,21}}{P_{i\to j}(\tilde{P}_i + \tilde{P}_{x_i,x_j,11})} \end{cases}$$

We thus obtain the following formulas for $P_{i\to j}$ and $\mu_{i\to j}$:

$$\begin{cases} P_{i\to j} = \beta_{ij}P^{x_j,x_j}_{x_i,x_j} - \dfrac{(\beta_{ij}P^{x_i,x_j}_{x_i,x_j})^2}{\beta_{ij}P^{x_i,x_i}_{x_i,x_j} + \kappa_i\Big(P_0^{i,j} + (1-\alpha_{ij}/\kappa_i)P_{j\to i}\Big)} & (5.56) \\ \mu_{i\to j} = -\dfrac{\beta_{ij}P^{x_i,x_j}_{x_i,x_j}\kappa_i[P_0^{i,j}\mu_0^{i,j} + (1-\alpha_{ij}/\kappa_i)P_{j\to i}\mu_{j\to i}]}{P_{i\to j}[\beta_{ij}P^{x_i,x_i}_{x_i,x_j} + P_0^{i,j} + (1-\alpha_{ij}/\kappa_i)P_{j\to i}]} & (5.57) \end{cases}$$

Mean and precision of the beliefs We now write the expression of the beliefs:

$$b_i(x_i) \propto \left(\prod_{j\in N(i)} m_{j\to i}(x_i)\psi_i(x_i)^{\gamma_i}\right)^{\kappa_i} \tag{5.58}$$

where

$$\begin{cases} m_{j\to i}(x_i) \propto \mathcal{N}_{x_i}(\mu_{j\to i},\, P_{j\to i}) & (5.59) \\ \psi_i(x_i) \propto \mathcal{N}_{x_i}(\mu_{\text{ext}\to i},\, \Sigma_{\text{ext}\to i}) & (5.60) \end{cases}$$

It comes from Equation (5.58) that the beliefs follows a Gaussian distribution ($\log(b_i(x_i)) = C - \frac{1}{2}(x-\mu_i)^T P_i(x-\mu_i)$) with precision P_i and mean μ_i such that:

$$\begin{cases} P_i = \kappa_i\Big(\sum_{j\in N(i)} P_{j\to i} + \gamma_i P_{\text{ext}\to i}\Big) & (5.61) \\ P_i\mu_i = \kappa_i\Big(\sum_{j\in N(i)} P_{j\to i}\mu_{j\to i} + \gamma_i P_{\text{ext}\to i}\mu_{\text{ext}\to i}\Big) & (5.62) \end{cases}$$

Relations between means and precisions of the messages and the beliefs Here we combine Equations (5.56) and (5.57) giving the expression of the mean and precision of the messages, with Equations (5.61) and (5.62) giving the expression of the mean and precision of the beliefs.

$$\begin{cases} \kappa_i P_0^{i,j} = P_i - \kappa_i P_{j\to i} & (5.63a) \\ \kappa_i P_0^{i,j}\mu_0^{i,j} = \mu_i P_i - \kappa_i\mu_{j\to i}P_{j\to i} & (5.63b) \end{cases}$$

So $P_0^{i,j} + (1-\frac{\alpha_{ij}}{\kappa_i})P_{j\to i} = \frac{1}{\kappa_i}\Big(P_i - \alpha_{ij}P_{j\to i}\Big)$
and $P_0^{i,j}\mu_0^{i,j} + (1-\frac{\alpha_{ij}}{\kappa_i})P_{j\to i}\mu_{j\to i} = \frac{1}{\kappa_i}\Big(\mu_i P_i - \alpha_{ij}\mu_{j\to i}P_{j\to i}\Big)$

We thus obtain from Equation (5.56):

$$P_{i\to j} = \beta_{ij} P^{x_j,x_j}_{x_i,x_j} - \frac{(\beta_{ij} P^{x_i,x_j}_{x_i,x_j})^2}{\beta_{ij} P^{x_i,x_i}_{x_i,x_j} + P_i - \alpha_{ij} P_{j\to i}} \tag{5.64}$$

$$= g_{ij}(P_i - \alpha_{ij} P_{j\to i}) \tag{5.65}$$

where $g_{ij}(y) = \beta_{ij} P^{x_j,x_j}_{x_i,x_j} - \frac{(\beta_{ij} P^{x_i,x_j}_{x_i,x_j})^2}{\beta_{ij} P^{x_i,x_i}_{x_i,x_j} + y}$.

Similarly, Equation (5.57) gives:

$$P_{i\to j}\mu_{i\to j} = -\frac{\beta_{ij} P^{x_i,x_j}_{x_i,x_j} [P_i\mu_i - \alpha_{ij}\mu_{j\to i} P_{j\to i}]}{\beta_{ij} P^{x_i,x_i}_{x_i,x_j} + P_i - \alpha_{ij} P_{j\to i}} \tag{5.66}$$

$$= h_{ij}\Big(P_i\mu_i - \alpha_{ij} P_{j\to i}\mu_{j\to i},\ P_i - \alpha_{ij} P_{j\to i}\Big) \tag{5.67}$$

where $h_{ij}(x,y) = -\frac{\beta_{ij} P^{x_i,x_j}_{x_i,x_j} x}{\beta_{ij} P^{x_i,x_i}_{x_i,x_j} + y}$. □

5.3.2 Implementation of Gaussian Circular BP with rate units

As in the binary case, we map directly equations defining eCBP (Equations (5.35), (5.36), (5.37), and (5.38)) onto a rate model implementing exactly the extended Circular BP algorithm in the Gaussian case. In this rate model, shown in Figure 5.2, populations are associated to oriented edges $x_i \to x_j$ (messages) or to unitary variables x_i (beliefs) and each population is composed of two rate units. In the "belief populations", one unit encodes for the product between the precision (or inverse variance) and the mean $v_i = P_i \times \mu_i$ and another unit encodes for the precision P_i. Similarly, in the "message populations", one unit encodes for $v_{i\to j} = P_{i\to j} \times \mu_{i\to j}$ and another unit encodes for $P_{i\to j}$. This makes this implementation part of other parametric representational schemes like PPCs in which the activities of neurons determine the parameters of the distribution (Fiser et al., 2010).

The "message synapses" are, as in the binary case, non-linear (see functions h_{ij} and g_{ij}), collecting the (weighted) difference between the quantity associated to the marginal and the one associated to the opposite message. Function g_{ij}, acting on a difference of precisions, is a positive and increasing function, with a slope, minimum and maximum depending on the statistical dependencies P_{x_i,x_j} between variables x_i and x_j. Function h_{ij} has a very similar shape to g_{ij} with an additional multiplication by a difference of potentials (i.e., precision × mean): $h_{ij}(x,y) = A + B g_{ij}(y) x$. Therefore $g_{ij}(y)$ can be seen as a modulating factor to the linear dependency of function h_{ij} on x. Similarly, in the binary case, the interaction function $f_{ij}(x) = \phi^{-1}(\phi(\beta_{ij} J_{ij})\phi(x))$ depends almost linearly (except at its bounds) on x, with a scaling factor J_{ij} depending on the statistical dependency between x_i and x_j, and represents the strength of the connection in the neural implementation. Overall, as the algorithm does not differ much between the binary and the Gaussian case, it indeed reflects in the proposed neural implementations, which strongly look alike as well.

5.3.3 Effects of circularity

In this paragraph, we consider the Circular BP algorithm (i.e., extended Circular BP in the particular case $(\boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$) and look at the effects of circularity on the means and the precisions μ_i and P_i of the belief or marginal distribution. The term "circularity" is defined by

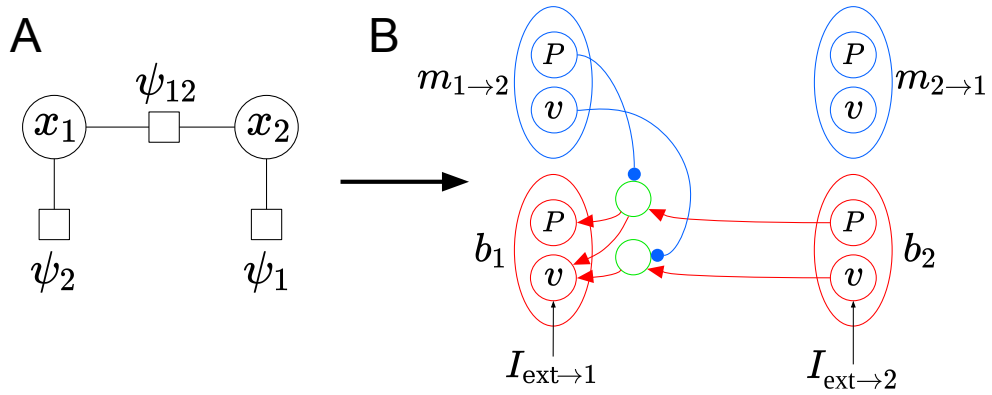

Figure 5.2: **Detailed neural implementation of Gaussian Circular BP with a rate network**, focusing on the connection between B_1 and B_2 from Figure 4.1B. **(A)** Example of probabilistic graph, representing the probability distribution $p(x_1, x_2) = \psi_{12}(x_1, x_2)\psi_1(x_1)\psi_2(x_2)$. **(B)** Each population is composed of a unit encoding for the precision P and a unit encoding for the potential v (product between the precision and the mean: $v = P \times \mu$). Error neurons in green explicitly encode the errors $P_i - \alpha_{ij}P_{j\to i}$ and most importantly $v_i - \alpha_{ij}v_{j\to i}$, which represents the information on node i unknown to node j, that is, not brought by j.

$1 - \alpha$, i.e., the distance to the value of 1 which corresponds to BP (and is optimal for acyclic graphs, in which BP carries out exact inference).

First, increased amounts of circularity (that is, decreased α_{ij}) leads to increased precisions (or equivalently, smaller variance). This indicates an increased confidence in the beliefs. A mathematical justification is that $P_{i\to j}$ is an increasing function of all precisions of the incoming messages and of the circularity level $(1 - \alpha)$ as g_{ij} is positive and increases as function of its argument $\sum_{k\in\mathcal{N}(i)\setminus j} P_{k\to i} + (1 - \alpha_{ij})P_{j\to i}$. The increase in overconfidence due to circularity in the Gaussian case reminds the binary case, which had the same exact conclusions (see for instance Figure 1.4). This overconfidence effect is even increased in graphs with cycles, in which information not only reverberates within an edge ($i \to j \to i \to \dots$ because of $\alpha_{ij} \neq 1$, but also naturally through the cycles. Because all precisions are positive, it increases the already existing overconfidence brought by the circularity $1 - \alpha$. For a demonstration that BP overestimates precisions (or equivalently underestimates variances) in graphs with loops, see Weiss and Freeman (1999).

As in the binary case, the effect of circularity seen in Gaussian Circular BP is not restricted to overconfidence but also to wrongness of the estimation (and potentially, of the subsequent decision). In the binary case, estimated beliefs (between 0 and 1) are not on the same side of the neutral value of 0.5, as shown in Figures 1.3 and 1.4. In the Gaussian case, the estimation of the means $\{\mu_i\}$ depends on the value of α and is only approximate. Interestingly, Weiss and Freeman (1999) shows that means obtained by BP ($\boldsymbol{\alpha} = \mathbf{1}$) are exact if BP converges, even in graphs with cycles. Because function h_{ij} determining $|v_{i\to j}| \equiv P_{i\to j}|\mu_{i\to j}|$ increases as function of its first argument and decreases as function of its second argument, and because μ can be negative, it is impossible to predict the evolution of $v_i = P_i \times \mu_i$ (nor the means μ_i themselves) with circularity. Nevertheless, simulations show that increased circularity (that is, decreased levels of α) generally leads to more extreme estimations (i.e., increased absolute means $|\mu_i|$) in addition to overestimated precisions, as shown in Figure 5.3.

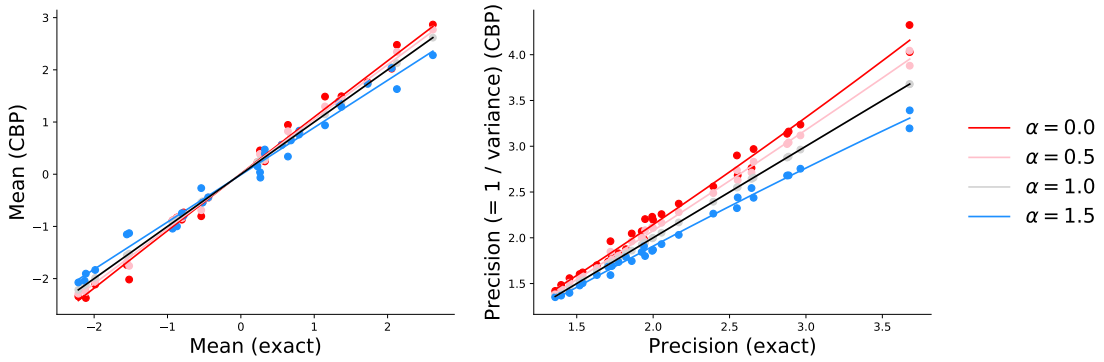

FIGURE 5.3: **Effects of Circular BP on the estimation of means and variances**. We consider an acyclic probabilistic graph and generate its weights (interactions between variables) and its inputs (unitary potentials) randomly. Full lines represent the best fitting polynomial function of degree 3. **(A)** Estimated means increase in absolute value due to circularity (that is, when α decreases from its value of 1 (which corresponds to BP). **(B)** Circularity increases confidence, as Circular BP overestimates precisions for $\alpha < 1$ and underestimates them for $\alpha > 1$.

5.3.4 Circular BP and predictive coding

Circular BP and impaired predictive coding are currently the two main types of Bayesian models in computational psychiatry. Both model brain processes underlying the emergence of suboptimal behavior and psychosis.

Predictive coding The predictive coding theory (Rao and Ballard, 1999; Friston and Kiebel, 2009; Spratling, 2016, 2017), briefly mentioned in section 1.3, is currently the most advanced theory of how the brain performs probabilistic inference. According to this theory, predictions are implemented across the cortical hierarchy and prediction errors represent the mismatch between the top-down prediction (feedback) and the signal (feedforward), with the overall goal of minimizing the prediction errors or surprise caused by a new stimulus, which is equivalent to minimizing the free energy (Friston, 2005). This explains for instance the dampening of neural responses when the sensory input can be predicted, as for the mismatch-negativity signal (MMN) (Wacongne et al., 2012; Lieder et al., 2013).

Predictive coding and BP It remains to be seen how BP (or extended Circular BP) relates to predictive coding, and more specifically, how Equations (5.41) and (5.42) relate to the predictive coding equations, which have a very similar form. However, if we consider BP on a hierarchical probabilistic graph (chain or tree) representing a hierarchy of concepts, node i located at the bottom of the hierarchy receives information from the top ("prior") and from the bottom ("likelihood") and $\mu_i = \mu_{j\to i} + W(\mu_{\text{ext}\to i} - \mu_{j\to i})$ where $W = \frac{P_{\text{ext}\to i}}{P_{j\to i}+P_{\text{ext}\to i}}$ (for eCBP, $P_{\text{ext}\to i}$ gets replaced with $\gamma_i P_{\text{ext}\to i}$). This corresponds to the idea of predictive coding that

$$\mu_{\text{posterior}} = \mu_{\text{prior}} + \frac{P_{\text{likelihood}}}{P_{\text{posterior}}}(\mu_{\text{likelihood}} - \mu_{prior})$$

where $(\mu_{\text{likelihood}} - \mu_{prior})$ represents the *prediction error* (see for example Figure 1 of Sterzer et al. (2018)). A more detailed parallel between predictive coding and (Circular) BP still needs

to be proposed, particularly in the more general case where node i is not at the bottom of the hierarchy.

We insist here on the fact that the notions of "prediction error" from Predictive Coding theories and from the current work are different. In Predictive Coding theories, prediction errors represent the mismatch between the prediction arising from the top of the hierarchy (prior) and the information from the bottom of the hierarchy (likelihood). In cases where likelihoods are predicted (for instance, if sensory evidence does not change over time), then prediction errors from Predictive Coding become zero. On the contrary, in the proposed neural implementation of Circular BP, prediction errors are the difference between a knowledge of the marginal (that is, of all the messages arriving to a node) and the knowledge brought by a single message. For this reason, prediction errors do not become zero, even when sensory inputs do not vary in time.

Impaired predictive coding and Circular BP Alterations of predictive coding have been widely considered as a model of psychosis and schizophrenia (Fletcher and Frith, 2009; Corlett et al., 2009, 2011; Adams et al., 2013; Wacongne, 2016; Sterzer et al., 2018), including the specific phenomenon of bistable perception (Schmack et al., 2013, 2015; Weilnhammer et al., 2017; Schmack et al., 2017) (see also section 2.2) and even non-clinical hallucinations or tendencies towards delusional ideation (Powers et al., 2017; Stuke et al., 2017, 2019) and the effects of the ketamine drug (Corlett et al., 2007, 2016). More specifically, schizophrenia and/or psychosis have been linked to aberrant weighting of prior beliefs and sensory information. Depending on the specific task, conclusions are different: some studies conclude on the weakening of prior beliefs in psychosis, while other studies present the exact opposite conclusions (see Sterzer et al. (2018) and (Corlett et al., 2019) for reviews). However, as explained by Sterzer and colleagues, these seemingly contradictory conclusions can be reconciled by assuming that priors are not impacted the same way in all sensory modalities and at all levels of the hierarchy. Interestingly, the predictive coding theory can be used to model behavior but also neural activity; for instance, Weilnhammer et al. (2017, 2018) used model-based fMRI to find neural correlates of quantities used in the behavioral model like (low-level or high-level) predictions or prediction errors.

It has been stated several times that because of their similarity, predictive coding theories could be reconciled with BP and its variants (Jardri and Denève, 2013a; Notredame et al., 2014; Denève and Machens, 2016; Jardri et al., 2017). The current work brings a contribution towards that goal, by formulating (extended) Circular BP in the Gaussian case, and most importantly, by showing through Equations (5.41) and (5.42), that the mean associated to the marginal distribution can be written as a linear combination of means associated to the messages and the mean of the external input. Although the idea remains to be backed by simulations, impaired predictive coding models could be related to Gaussian Circular BP by hypothesizing an asymmetry in the loop correction factor (matrix $\boldsymbol{\alpha}$). This relates to the idea of ascending and descending loops with different strengths, an idea which was only used sparingly in this thesis (see section 2.2 on the modeling of bistable perception) but might turn out crucial in this case. This could be a first step towards a deeper comprehension of the link between predictive coding theories of psychosis and the Circular Belief Propagation algorithm.

5.4 Conclusion

Learning to carry out approximate inference in the general case The experiments reported in chapter 3 demonstrated successes of supervised learning on rather small and binary graphical models with pairwise interactions. However, many situations where humans perform inference nearly optimally involve much more generative models. A follow-up of this work using the theory developed in the current chapter could focus on more general graphical models, with

higher-order interactions and/or with continuous variables, and investigate whether parameters can be learnt (in a supervised but also unsupervised manner) in these cases.

More specifically, one might wonder what Circular BP could bring in the Gaussian case. Indeed, Weiss and Freeman (1999) shows that in Gaussian factor graphs, when BP converges then it produces the right means. Therefore, any change in the loop correction factor from Circular BP would then alter the estimate of the mean. Nevertheless, what Circular BP can possibly do is to approximate the precisions a lot better than BP, while controlling the error made while estimating the means. Moreover, Circular BP could improve the convergence properties of BP and thus help, as in the binary case, carrying out its role as an approximate inference algorithm. Indeed, BP produces the correct means whenever it converges, but the beliefs produced in absence of convergence have in general little to do with the true marginals (Murphy et al., 1999); see also Figure 3.8 for the binary case. Another option would be to rely on BP to estimate the means, and estimate the precisions using Circular BP (see Equations (5.35) and (5.37)) separately, with appropriate amounts of loop correction (matrix $\boldsymbol{\alpha}$).

Proposing a neural implementation in the general case We proposed a neural model of Circular BP for very specific probability distributions: Markov Random Fields with pairwise interactions, composed of binary (in chapter 4) or Gaussian variables (in the current chapter). This is already more general than most articles on the topic; for instance, most previous propositions of neural implementations of BP considered binary variables only Ott and Stoop (2006); Steimer et al. (2009); Litvak and Ullman (2009). Because humans are nearly optimal at carrying out probabilistic problems involving more various distributions than simply binary of Gaussian, future work should investigate potential implementations for (more) general probability distributions, such as the so-called exponential family of probability distributions. This will allow for even more plausibility of (extended) Circular BP as a possible model of how the brain performs probabilistic inference.

One problem of particular interest is the explaining-away problem, which is not covered by the current work because explaining-away involves probabilistic interactions which are more than pairwise. Note that the neural network implementing the explaining away problem would require additional connections and complexity than simply mirroring the probabilistic graph, which was enough for probability distributions with at most pairwise interactions.

Chapter 6

Discussion

Thesis summary

Circular Inference (Jardri and Denève, 2013a) is a Bayesian model of psychiatric disorders, previously designed to account for clinical manifestations of schizophrenia and psychosis. Circular Inference relies on the Circular Belief Propagation algorithm, an approximate probabilistic inference algorithm that proposes an additional parameter compared to Belief Propagation, called the *loop correction factor*. This loop correction factor sets the amount of circularity in the inference and is seen as a proxy to the (local) level of excitation-inhibition balance in the brain network assumed to perform probabilistic inferences. According to this framework, circular reasoning and psychotic symptoms arise for lowered loop correction factor, which would mean, for low levels of inhibition compared to excitation.

The work presented in this thesis provides further evidence for Circular Inference as a model of pathological inferences (e.g., hallucinations and delusions), near-optimal inferences, and in between non-clinical suboptimal inferences, ranging from usual inference biases (exemplified by the bistable perception and the jumping to conclusions phenomena, and the general overconfidence) to sub-clinical behavior like believing in conspiracy theories despite contradicting evidence; see chapters 2 and 3.

Additionally, this thesis develops the Circular Inference model in different ways. First, **conceptually**, by providing the Circular BP algorithm with a theoretical foundation, which is done by relating it to existing algorithms such as Fractional BP (see chapter 3). Second, more **practically**, by proposing neural implementations (rate networks and spiking networks, for binary or Gaussian variables) and biologically-plausible learning mechanisms overall describing how probabilistic inferences could be carried out in the brain using this algorithm (see chapters 4 and 5). Finally, the model is expanded **theoretically**, by investigating the convergence properties of the algorithm, by writing Circular BP for more complex probability distributions than previously, and by generalizing the initial Circular BP into extended Circular BP (see chapters 3 and 5).

In what follows, we first provide some general perspectives to the work, which can be considered as "food for thought" and are not directly connected to each other. Next, we tackle open discussion points. Last, we open up on potential follow-ups to the work as well as future applications.

General perspectives of the work

Marr's levels of analysis and the Circular Inference model Marr (Marr, 1982) proposes that computational models might be used to investigate three levels of analysis: the computational level (*what* dose the brain compute and *why*?), the algorithmic level (which *representations and algorithms can describe these computations?*), and the physical level (*how* are these algorithms implemented in neural circuits?). We develop in this thesis a normative model of how the brain performs probabilistic inferences (the "what" question). We propose that neurons encode for (log-)parameters of the probability distribution, and carry out the Circular Belief Propagation algorithm (the "which" question: representation and algorithm). This naturally leads to a rate network implementation of inference where neurons fire proportionally to the variable they encode, or equivalently, to a spiking network composed of neurons whose membrane potential represent errors of prediction and spikes are emitted to reduce this error (the "how" question). Note that the Circular Inference model is one of the few computational psychiatry models to propose both a quantitative and mechanistic account of psychiatric symptoms (see Valton et al. (2017)).

But what is even more interesting in Circular Inference is its simplicity. More precisely, the parameter α both means something at the implementation level (loop correction factor, controlling for the reverberation of information) and the physical level (excitation-inhibition balance, or specifically in the implementations proposed, synaptic gain of the control unit). This allows for simple predictions. For instance from the computational level to the physical level: if people are suboptimal at carrying out a particular task, it implies that there is an imbalance between excitation and inhibition during or for this task. But also from physical level to computational level: a population which does not differ from another one in terms of the excitation-inhibition balance (measurable with brain data) is not expected to show more circular reasoning at the task (measurable with behavioral data).

Cyclic generative models and the structure of the recurrent brain In this work, we expanded the scope of the Circular Inference model by using cyclic graphs. Considering cyclic probabilistic graphs is an important step to model pathology using Circular Inference. First, because tasks that we perform on a daily basis are often associated with generative models having cyclic relationships. For instance, in the case of perception, fruits is a higher concept, "causing" the concepts of apple and strawberry, both of which "causing" the concept of color red, which is also caused by the concept of red peppers. Inferring whether there are red apples in what seems to be (because it's far) a fruit stand at the market given that you see a redish color from a distance and that the harvest season for strawberries is about to end, is a task involving cyclic dependencies. Obviously, hierarchical generative models (typically a chain of concepts) are appealing because of their simplicity and because the brain looks hierarchical at first glance. However, even though it is well accepted that the visual cortex has a hierarchical structure overall, it might not be the case in all cortices and at all scales. We believe that other Bayesian models which relate to BP would highly benefit from considering (tasks involving) cyclic generative models. The second reason why considering cyclic graphs is important is that such graphs can explain more strange types of inferences numerically. This allows for the modeling of specific symptoms without the need for additional complexity. For instance, we believe that the inability in Jardri and Denève (2013a) (and all the remaining work studying the Circular Inference) to account for hallucinations while using a symmetric loop correction factor (matrix $\boldsymbol{\alpha}$) - that is, without considering different treatments for the reverberation of sensory evidence versus prior - relies on the fact that the probabilistic graphs considered are acyclic. The solution to this issue was to selectively impair the so-called *ascending loops* (controlling for the reverberation of sensory

evidence exclusively) or *descending loops* (controlling for the reverberation of prior knowledge exclusively). We hypothesize that in the case where graphs are cyclic, the effects of ascending loops or descending loops could at least partly be explained using a global loop correction factor (that is, no differentiation between ascending and descending loops) with particular cyclic graph structures.

This raises the need to know (or at least make hypotheses on) the type of generative model associated to a task. For instance, in section 2.2, the bistable perception phenomenon was modeled using an acyclic probabilistic graph composed of three nodes. According to the proposed neural implementation of eCBP, if the generative model is known then it directly predicts the associated structure of the network performing the inference. Interestingly, the hypothesized difficulty (discussed in the following paragraph) of learning excitation-inhibition imbalance in dense recurrent networks could potentially mean that brain networks are not hierarchical by mistake. If the brain uses Circular BP to perform inference but cannot accurately learn a good-enough excitation-inhibition balance (that is, how to cancel properly the effects of cycles), then the brain must rather be reasoning over concepts organized hierarchically (acyclic generative models which take the form of trees) or nearly hierarchically (few cycles in a globally hierarchical generative model). More generally, an intriguing observation is that BP performs well on graph structures usually found in the brain (e.g. on small-world structures (Litvak et al., 2009) and on hierarchical structures) compared to random graphs. This might be linked to the need to make probabilistic inferences (and potentially other tasks) with high performance; see also Litvak et al. (2009) which develops the idea that many features of cortical networks foster good inferences, probably as a result of evolution.

Optimality and suboptimality of BP, Circular BP, and of our inferences Related to the previous paragraph, we point here at a logical interrogation arising from the results of chapters 2 and 3. In chapter 2, we explain that Circular BP is a good model of suboptimal inference. However, in chapter 3, we provide evidence that (extended) Circular BP performs near-optimal inferences given appropriate loop correction factors. This raises the following question: if the brain is able to perform approximate inference of high enough quality by using eCBP, then why are our decisions suboptimal? Note that this piece of criticism can also be made to other models and does not mean that eCBP is not carried out by the brain. A first option is that the brain does not manage to find the perfect local excitation/inhibition ratios (that is, the loop correction factors) ensuring a good quality of inference. More work needs to be done to investigate the unsupervised learning rule provided in section 4.5, which might not find a solution as good as with supervised learning and lead to suboptimal inferences. Another remark is that we proposed a neural implementation of inference and learning in the binary case only, and even if the inference and learning were shown to be good in this case, it is highly possible that the results do not transpose to more general cases, for example to graphs with higher order potentials. A second option of the reason why our inferences are not optimal is that the brain does not only seek for optimality while doing inference but instead needs to trade-off between the quality of inference and other factors. These factors include the need for fast inferences (e.g. the necessary to react quickly while detecting dangerous predators), which is fostered by circularity (by considering for instance that one should run away as soon as there the chances of having a predator are at least 10%). A plausible intuition is that these other components (such as the need for quick inferences) bias the quality of inference by purposefully adding circularity at the expense of the sole quality of inference.

The suboptimality of BP in cyclic graphs (see chapter 3) interrogates the fact that Circular BP is implemented in the brain. We discuss here briefly an alternative to the view presented in the thesis, namely, that BP, not Circular BP, is carried out in the brain. According to this view,

the difference between performing exact and suboptimal/pathological inferences can be explained by the structure of the generative problem, not by the inference algorithm which is fixed (BP). In acyclic graphs, BP is exact and therefore explains optimal inferences. In contrast, in very cyclic graphs, BP can be extremely inaccurate (extremely fallacious inferences, potential absence of convergence) and thus could be a model of psychosis, as pointed out by Valton et al. (2017). In other words, differences between a person with and without schizophrenia and would not be based on the inference mechanism, but instead on the structure and weights of the generative network. However, this hypothesis seems highly unlikely. Indeed, BP performs *extremely* poorly in many cases, which is at odds with the substantial amount of evidence showing that inferences made by humans are usually close to optimality (see Introduction chapter). In the simple case of section 2.3 where BP was indeed considered as a model of normal brain functioning, but we had to select carefully the parameters of the cyclic generative model in order for BP to indeed perform well (e.g., by restricting the amplitude of the graphs weights, that is, of the interaction strength between variables). If BP can already perform pathologically for such tiny small-world graphs, humans would have never been able to develop as a species if their brains indeed carried out BP, because we would not be able to treat more than a very limited range of probabilistic problems yet arising on a daily basis.

If our brains implement (extended) Circular BP to carry out near-optimal inferences, then finding the right amount of correction to be applied to counter the effects of cycles is equivalent to balancing excitation and inhibition in the network, as stated in section 4.5. That means that moving away from BP - that is, perform "circular reasoning" ($\alpha \neq 1$) in the initial sense of Jardri et al. (2017) - is necessary for the quality of our inferences, as long as we move away from BP in a direction opposite to the effects of the loops (*anti* circular reasoning).

Decision-making and overconfidence Until now, the only tasks considered to differentiate people with more circular reasoning from people with less circular reasoning were the Necker cube task (bistable perception), and (variants of) the beads task. Note that in all these tasks, we suppose that participants report their true value of belief, meaning that we do not have a particular decision model. Decision-making, similarly, is affected by the presence of circular reasoning. However, decision-making also depends on possible additional biases at the moment of making the decision. Vinckier et al. (2016) reports for instance an altered decision stochasticity in controls under the influence of ketamine, which was proposed as a pharmacological model of schizophrenia and/or the transition to psychosis. However, even by hypothesizing that there is no decision noise, i.e., that the decision truly depends on the belief (marginal probability), the picture is not so clear about the consequences of circular reasoning on decision-making as it highly depends on the type of the task.

If we consider the MAP problem, that is, if the task requires to report the most probable possibility, then circular reasoning has a small effect for binary variables, and a potentially strong effect for Gaussian variables. For instance, in the example of Figure 1.3, most beliefs remain on the same side of 0.5 except from one node. In the Gaussian case, however, means vary with α, meaning that the MAP answer is different as soon as α is changed (more analyses would be needed to estimate the strength of the variation with α).

On the contrary, for any other decision criterion than the MAP problem, the presence of circular reasoning highly modifies the decision, depending on the specific task. An example of other decision criterion is to answer A if the belief is above 80%, and no otherwise.

Future directions and applications

Although the work presented in this thesis is mostly conceptual and theoretical, it has several potential practical applications.

Fitting brain imaging data A very interesting direction would be to fit resting-state brain imaging data (EEG) to our rate implementation of Circular BP, where the generative model associated to the rate network (and defines, among others, its structure) is determined by anatomical-functional connections measured in brain networks. Whole-brain atlases of the brain or *connectomes* are typically composed of around 100 nodes (Yeh et al., 2018) , which makes it possible to run Circular BP on. There are a number of potential technical issues which would make impractical the fitting, but finding a correlation between the amount of circularity from brain imagining data only and the severity of symptoms (or with the amount of circularity fitted from behavioural data only) would be a giant leap. An collaboration on this topic is starting with Pierre Yger, Sophie Denève and Renaud Jardri.

Fitting several probabilistic tasks to the same subject The field of computational psychiatry is nowhere near explaining all symptoms associated to mental disorders. Eventually, models must aim at understanding the heterogeneity of symptoms, of their time evolution, and of responses to treatment among individuals. Until now (Jardri et al., 2017; Simonsen et al., 2021; Chrysaitis et al., 2021), each participant to the study took one task only. Instead, we suggest that having the same participants go through several tasks is a step in that direction. Indeed, the amount of circularity as measured by α depends on the task (intuitively, because it represents a local measure of the E/I ratio in the brain circuit performing the task). Therefore, we believe that different tasks reflect potentially different components of the disorder. For instance, a person with hallucinations might not be different from a person with delusions on a given probabilistic task, but behave differently on another task. If this is indeed the case, then probabilistic tasks as different as possible need to be imagined, in order to capture several components of the disorder and possibly differenciate the subpopulations thanks to the fitted amounts of circularity.

Fitting more complex probabilistic tasks Currently, the data fitted using the Circular Inference model come from very simple probabilistic tasks composed of a "prior" and a "sensory evidence" (Jardri et al., 2017; Simonsen et al., 2021); Seriès (2021), however, points out that it is not clear how the high-level social cue from Simonsen et al. (2021) compares to the prior cue from Jardri et al. (2017). Designing tasks where subjects need to reason on variables with cyclic dependencies seems crucial. There are several reasons for that. First, it will ensure that subjects are indeed performing inference, and instead do not have another simple strategy: intuitively, the more complex the task is, the more intuitive the behavior is and therefore beliefs will truly be the ones computed by the brain. The second reason is that an interesting prediction comes from Circular BP in cyclic graphs: the fact that for certain tasks (that is, certain cyclic generative models), the schizophrenia population will perform better than the control population. Indeed, there are possible compensatory effects between the natural circularity arising from the graph (which can be positive or negative) and the circular reasoning on this graph (see chapter 3). Showing such an effect would undoubtedly be a big step for the Circular Inference model.

Social cognition and Circular Inference In terms of potential broader impact, this work could help to understand the impact of the propagation of false information in terms of the

creation of potentially wrong beliefs in the general population. An example of this is social networks, online or offline: exchange of information between people can be seen as a message-passing procedure. The self-amplification of beliefs in local communities/societies (local networks) could be seen as a consequence of information not being properly removed, as explained throughout this text. The experiment described in Simonsen et al. (2021), in which the subject reports his confidence level based on the sensory evidence and the information given by 4 other people paves the way for such work. A related point is that Circular Inference could be used to understand the impairments in social cognition in schizophrenia and in other mental disorders (Seriès, 2021).

Considering social cognition leads to the following broader questions: which evidence should be shown to us (for instance on a web page) to create the least biased beliefs possible? In fact, the formation and maintenance of false beliefs is probably at least equally impacted by the context of clustered social groups (including online social networks) which tend to reinforce their beliefs, than by the false reasoning of an individual of this group provided with contradictory sources of information. Indeed, not only corrective factors can be chosen to minimize the reverberation of given external evidence, but external evidence could also be chosen to minimize reverberation for a given amount of corrective factors. In other words, whereas we propose in chapter 3 that loop corrective factors could act to cancel the effects of cycles in probabilistic graphs, carefully chosen evidence presented to the network could limit the effects of circularity in probabilistic graphs. An example of experiment testing circular reasoning in a group would be the following: some evidence about a particular statement (e.g., it will rain tomorrow), more or less reliable (e.g., is provided to certain random people of the group. Each person can only interact with his given "neighbors", in the form of a single number. We let communication take place in a synchronized manner (similarly to message-passing algorithms, including BP) until convergence, and people are asked to report their beliefs after each new time they talk to all of their neighbors.

Using Circular Belief Propagation in real life: future leads for machine learning research We have proposed with the Circular BP algorithm an alternative to Fractional BP to perform approximate probabilistic inference. We also proposed a way of improving further the power of Circular BP with the "Circular BP with memory" model (see section 3.6), which could not have been possible with Fractional BP. Such algorithms could be applied to inference problems in real life, replacing BP or algorithms based on BP. Note that eCBP has a similar complexity to BP, and thus it could be implemented as efficiently, but with increased performance. Of course, that requires to learn (in a supervised or unsupervised way) the parameters of the algorithm adapted to the task to be performed, and fix these parameters once for all. note that this is already the case for other algorithms, for instance artificial neural networks encoded in our phones to process images from the camera. A reliable unsupervised learning rule for eCBP would even allow an inference system to learn on-line a new task without being reprogrammed (as supervised learning is in general a lot more time- and resource-consuming).

It is crucial for the prediction of an inference system to be (nearly) exact: such systems are used in many key domains including medical diagnosis, and having biased or overconfident predictions could cause death (for instance, if an algorithm decides to not accept a person to the hospital based on its mild symptoms although the person needed treatment). However, better methods like extended Circular BP could help reducing such systematic biases. Future work will need to determine whether the error made by eCBP can always be reduced to an "acceptable" threshold value, whatever the probabilitiy distribution is. More generally, the type and amplitude of the error made by eCBP needs to be further investigated.

Appendix A

Theoretical background: From Gibbs free energy approximation to BP and its variants

Here we provide the theoretical background underlying Belief Propagation and its variants (Fractional BP, Circular BP, extended Fractional BP, extended Circular BP, and all their special cases). Starting from an approximation of the Gibbs free energy, we derive the expression of extended Fractional BP , before considering the special case of binary variables and pairwise factors.

From Gibbs free energy approximation to messages

In what follows, we derive the message-passing update equations of the extended Fractional BP algorithm (and its special cases BP, Fractional BP, Power EP and α-BP, among others) based on the following approximation of the Gibbs Free Energy G_{approx}:

$$\begin{aligned} G_{\text{approx}} = &\sum_{(i,j)} \hat{\beta}_{ij} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log\left(\frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)}\right) - \sum_{(i,j)} \beta_{ij} \sum_{(x_i,x_j)} b_{ij}(x_i,x_j) \log\Big(\psi_{ij}(x_i,x_j)\Big) \\ &+ \sum_i \hat{\gamma}_i \sum_{x_i} b_i(x_i) \log\Big(b_i(x_i)\Big) - \sum_i \gamma_i \sum_{x_i} b_i(x_i) \log\Big(\psi_i(x_i)\Big) \end{aligned} \tag{A.1}$$

where $(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}})$ are called *counting numbers* (see Yedidia et al. (2005)), with *entropic counting numbers* $(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}})$ and *average energy counting numbers* $(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}})$. In the following, we define $\boldsymbol{\kappa} = 1/\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\alpha} = 1/\hat{\boldsymbol{\beta}}$, for consistency with the main text.

Special cases of this approximation lead to different algorithms. BP corresponds to $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$, Fractional BP, Power EP, and α-BP to $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$, nodal Fractional BP to $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1})$ and extended BP for $\boldsymbol{\beta} = \mathbf{1}$, among others. Circular BP and extended Circular BP do not correspond to a special case of $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. Instead, as explained below in the text, (extended) CBP is defined as a approximation of the message update equation of (extended) FBP, not of the Gibbs Free Energy.

The message update equations are simply fixed-point equations of G_{approx}, meaning that the beliefs computed by eFBP are stationary points of the approximate Gibbs free energy G_{approx}. This demonstration is similar to the one for BP (Yedidia et al., 2001, 2003), with the additional

parameters ($\boldsymbol{\kappa}$, $\boldsymbol{\alpha}, \boldsymbol{\gamma}$, $\boldsymbol{\beta}$). BP is closely linked to the Bethe free energy G_{Bethe}: fixed points of Loopy BP are stationary points of the Bethe free energy (Yedidia et al., 2001), and stable fixed points of Loopy BP correspond to minima of the Bethe free energy (Heskes, 2002). eFBP is similarly linked to G_{approx} (which generalizes the Bethe free energy G_{Bethe}).

The goal is to minimize the Gibbs Free Energy, so we form a Lagrangian to take the constraints into account. The Lagrangian is formed by adding Lagrange multipliers (μ, λ) to G_{approx}. Lagrange multipliers μ_i corresponds to the normalization constraint $\sum_{x_i} b_i(x_i) = 1$, while $\lambda_{i\to j}(x_j)$ corresponds to the marginalization constraint $\sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j)$. The Lagrangian is equal to:

$$\begin{aligned}\mathcal{L} = G_{\text{approx}} &+ \sum_i \mu_i \Big(\sum_{x_i} b_i(x_i) - 1 \Big) \\ &+ \sum_{(i,j)} \sum_{x_j} \lambda_{i\to j}(x_j) \Big(\sum_{x_i} b_{ij}(x_i, x_j) - b_j(x_j) \Big) \\ &+ \sum_{(i,j)} \sum_{x_i} \lambda_{j\to i}(x_i) \Big(\sum_{x_j} b_{ij}(x_i, x_j) - b_i(x_i) \Big) \end{aligned} \tag{A.2}$$

The partial derivatives of the Lagrangian are:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial b_i(x_i)} = -\sum\limits_{j\in\mathcal{N}(i)} \dfrac{1}{\alpha_{ij}} + \dfrac{1}{\kappa_i} + \dfrac{1}{\kappa_i} \log(b_i(x_i)) - \gamma_i \log(\psi_i(x_i)) + \mu_i - \sum\limits_{j\in\mathcal{N}(i)} \lambda_{j\to i}(x_i) \\ \dfrac{\partial \mathcal{L}}{\partial b_{ij}(x_i, x_j)} = \dfrac{1}{\alpha_{ij}} + \dfrac{1}{\alpha_{ij}} \log\Big(\dfrac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \Big) - \beta_{ij} \log(\psi_{ij}(x_i, x_j)) + \lambda_{j\to i}(x_i) + \lambda_{i\to j}(x_j) \end{cases}$$

It comes, by cancelling the partial derivatives of the Lagrangian, the following expression for the unitary beliefs:

$$b_i(x_i) \propto \psi_i(x_i)^{\kappa_i \gamma_i} \prod_{k\in\mathcal{N}(i)} \exp\Big(\kappa_i \lambda_{k\to i}(x_i) \Big) \tag{A.4}$$

and the pairwise beliefs:

$$\begin{aligned} b_{ij}(x_i, x_j) \propto &\psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \psi_i(x_i)^{\kappa_i\gamma_i} \psi_j(x_j)^{\kappa_j\gamma_j} \prod_{k\in\mathcal{N}(i)\setminus j} \exp\Big(\kappa_i \lambda_{k\to i}(x_i) \Big) \prod_{k\in\mathcal{N}(j)\setminus i} \exp\Big(\kappa_j \lambda_{k\to j}(x_j) \Big) \\ &\times \exp\Big(\lambda_{j\to i}(x_i) \Big(\kappa_i - \alpha_{ij} \Big) \Big) \exp\Big(\lambda_{i\to j}(x_j) \Big(\kappa_j - \alpha_{ij} \Big) \Big) \end{aligned} \tag{A.5}$$

Now defining the messages as a function of the Lagrange multipliers $m_{j\to i}(x_i) \equiv \exp(\lambda_{j\to i}(x_i))$, the approximate marginals $b_i(x_i)$ and approximate pairwise marginals $b_{ij}(x_i, x_j)$ can be written simply as:

$$b_i(x_i) \propto \Big(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i) \Big)^{\kappa_i} \tag{A.6}$$

$$\begin{aligned} b_{ij}(x_i, x_j) \propto &\ \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \psi_i(x_i)^{\kappa_i\gamma_i} \psi_j(x_j)^{\kappa_j\gamma_j} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i)^{\kappa_i} \\ &\times \prod_{k\in\mathcal{N}(j)\setminus i} m_{k\to j}(x_j)^{\kappa_j} m_{j\to i}(x_i)^{\kappa_i - \alpha_{ij}} m_{i\to j}(x_j)^{\kappa_i - \alpha_{ji}} \end{aligned} \tag{A.7}$$

Eventually, thanks to the constraint $\sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j)$, we obtain the relation between the messages m:

$$m_{i\to j}(x_j)^{\kappa_j} \propto \left(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i} \right)^{\kappa_i} \right) \left(m_{i\to j}(x_j) \right)^{\kappa_j - \alpha_{ij}} \tag{A.8}$$

$$\Leftrightarrow m_{i\to j}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i} \right)^{\kappa_i} \right)^{1/\alpha_{ij}} \tag{A.9}$$

The eFBP algorithm consists of running iteratively the fixed-point equation (A.9):

$$m^{\text{new}}_{i\to j}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m^{\text{old}}_{k\to i}(x_i) m^{\text{old}}_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i} \right)^{\kappa_i} \right)^{1/\alpha_{ij}} \tag{A.10}$$

Note that one could also use directly Equation (A.8) instead of (A.9) to define the eFBP algorithm:

$$m^{\text{new}}_{i\to j}(x_j) \propto \left(\sum_{x_i} \psi_{ij}(x_i, x_j)^{\alpha_{ij}\beta_{ij}} \left(\psi_i(x_i)^{\gamma_i} \prod_{k\in\mathcal{N}(i)\setminus j} m^{\text{old}}_{k\to i}(x_i) m^{\text{old}}_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i} \right)^{\kappa_i} \right)^{1/\kappa_j} \left(m^{\text{old}}_{i\to j}(x_j) \right)^{1-\alpha_{ij}/\kappa_j} \tag{A.11}$$

In fact, Equations (A.11) and (A.10) correspond respectively to the damped (with a particular damping value) versus undamped update equation; see section 4.2.1. Fractional BP (Wiegerinck and Heskes, 2002), which is derived similarly to extended Fractional BP, uses Equation (A.11) (with $(\boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$) rather than Equation (A.10). There is no absolute better choice: fixed points obtained are identical in both cases, and damping might provide better convergence properties but might slow down the system in cases where the algorithm would have converged without damping (see section 4.2.1).

Special case of extended FBP: Binary case, pairwise factors

Until now we simply hypothesized that factors were at most pairwise. In this paragraph we also consider variables x_i to be binary. We place ourselves in the log domain and we define $M_{i\to j} \equiv \frac{1}{2}(\log(m_{i\to j}(+1)) - \log(m_{i\to j}(-1)))$ and $B_i \equiv \frac{1}{2}(\log(b_i(+1)) - \log(b_i(-1)))$. We also write the factor ψ_{ij} as the following the 2x2 non-negative matrix:

$$\psi_{ij}(x_i, x_j) \equiv \begin{pmatrix} \psi_{ij}^{0,0} & \psi_{ij}^{0,1} \\ \psi_{ij}^{1,0} & \psi_{ij}^{1,1} \end{pmatrix} \tag{A.12}$$

We obtain from Equations (A.6) and (A.10):

$$\begin{cases} M_{i\to j} = g_{ij}\left(B_i - \alpha_{ij} M_{j\to i} \right) & \text{(A.13a)} \\ B_i = \kappa_i \left(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i} \right) & \text{(A.13b)} \end{cases}$$

where

$$g_{ij}(x) = \frac{1}{2\alpha_{ij}} \log \left(\frac{\left(\psi_{ij}^{1,1}\right)^{\alpha_{ij}\beta_{ij}} e^{2x} + \left(\psi_{ij}^{1,0}\right)^{\alpha_{ij}\beta_{ij}}}{\left(\psi_{ij}^{0,1}\right)^{\alpha_{ij}\beta_{ij}} e^{2x} + \left(\psi_{ij}^{0,0}\right)^{\alpha_{ij}\beta_{ij}}} \right) \tag{A.14}$$

Equations (A.13a) and (A.13b) correspond respectively to Equations (3.20a) and (3.20b) from the main text. However, the expression of function g_{ij} is different from the main text, which only considers the Ising model case: see below. Equation (A.14) described a sigmoidal function with parameters 4 parameters $\psi_{ij}^{0,0}, \psi_{ij}^{0,1}, \psi_{ij}^{1,0}, \psi_{ij}^{1,1}$.

Even more special case of extended FBP: Ising model

The Ising model, which we consider in the main text, is itself a particular case of the case considered just above. In the general case described above, pairwise interactions can be described by a non-negative 2x2 matrix with independent coefficients. In an Ising model, pairwise factors take a specific form: $\psi_{ij}(x_i, (x_j) \propto \exp(J_{ij} x_i x_j)$.

$$\psi_{ij}(x_i, x_j) \equiv \begin{pmatrix} e^{J_{ij}} & e^{-J_{ij}} \\ e^{-J_{ij}} & e^{J_{ij}} \end{pmatrix} \tag{A.15}$$

$$\implies g_{ij}(x) = \frac{1}{2\alpha_{ij}} \log \left(\frac{e^{\alpha_{ij}\beta_{ij}J_{ij}+2x} + e^{-\alpha_{ij}\beta_{ij}J_{ij}}}{e^{-\alpha_{ij}\beta_{ij}J_{ij}+2x} + e^{\alpha_{ij}\beta_{ij}J_{ij}}} \right)$$

It comes after a few manipulations, using $\log(x) = 2\phi^{-1}\left(\frac{x-1}{x+1}\right)$ (where $\phi = \tanh$) and $\phi(x) = 2\sigma(2x) - 1$:

$$g_{ij}(x) = \frac{1}{\alpha_{ij}} \phi^{-1} \left(\phi\Big(\alpha_{ij}\beta_{ij}J_{ij}\Big) \phi\big(x\big) \right) \tag{A.16}$$

We thus recover the expression of g_{ij} given in the main text. g_{ij} is a sigmoidal function depending on J_{ij} (strength of the interaction between variables x_i and x_j) and α_{ij} (inverse entropic counting number associated to the edge (i, j)).

Eventually, we recover the expression of the eFBP algorithm given in the main text:

$$\begin{cases} M_{i\to j} = g_{ij}\Big(B_i - \alpha_{ij} M_{j\to i}\Big) & \text{(A.17a)} \\ B_i = \kappa_i \left(\sum_{j \in \mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i} \right) & \text{(A.17b)} \end{cases}$$

with

$$g_{ij}(x) = \frac{1}{\alpha_{ij}} \phi^{-1} \left(\phi\Big(\alpha_{ij}\beta_{ij}J_{ij}\Big) \phi\big(x\big) \right) \tag{A.18}$$

Relation to other generalized BP algorithms (including Fractional BP, Power EP, alpha BP) and to Circular BP

Here we show the equivalence between nodal Fractional BP with $\gamma = 1$, Fractional BP, Power EP, and α-BP, if we consider a damped version of the algorithms.

In this work (see Equation (3.2)) we tackle the particular case where the counting numbers associated to the average energy are equal to 1: $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1})$. We use the same regions as in

the Bethe approximation, with modified entropic counting numbers $(1/\boldsymbol{\alpha}, 1/\boldsymbol{\kappa})$ but with average energy counting numbers equal to 1.

The special case of Belief Propagation is recovered for $(\boldsymbol{\kappa}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$.

Fractional BP (Wiegerinck and Heskes, 2002), Power EP (Minka, 2004) and α-BP (Liu et al., 2019) correspond to $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$, and more particularly use the damped message update equation (A.8) rather than its undamped version (A.9) (see damping in section 4.2.1).

Circular BP (Jardri and Denève, 2013a) does not correspond to any choice of $(\boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\beta}$, or $\boldsymbol{\gamma})$, but can be seen as an approximation of the case $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$ as explained below. Its message update equation is:

$$m_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{i\to j}} \tag{A.19}$$

(see Equation (1.3) for the BP equivalent). This leads in the binary case to $M_{i\to j} = f_{ij}(B_i - \alpha_{i\to j} M_{j\to i})$ where function f_{ij} is:

$$f_{ij}(x) = \frac{1}{2}\log\left(\frac{\psi_{ij}^{1,1} e^{2x} + \psi_{ij}^{1,0}}{\psi_{ij}^{0,1} e^{2x} + \psi_{ij}^{0,0}}\right) \tag{A.20}$$

in the general case, and

$$f_{ij}(x) = \phi^{-1}\big(\phi(J_{ij})\phi(x)\big) \tag{A.21}$$

for an Ising model. Because functions g_{ij} and f_{ij} are similar:

$$g_{ij}(x) \equiv \frac{1}{\alpha_{ij}}\phi^{-1}\left(\phi\Big(\alpha_{ij}\beta_{ij}J_{ij}\Big)\phi\big(x\big)\right) \approx \phi^{-1}\left(\phi\Big(\beta_{ij}J_{ij}\Big)\phi\big(x\big)\right) \equiv f_{ij}(x) \tag{A.22}$$

, then the message update equation of Circular BP approximates the update equation corresponding to $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{1}, \mathbf{1}, \mathbf{1})$.

Likewise, we can propose an extended Circular BP algorithm, defined by an approximation of eFBP given in Equation (A.17):

$$\begin{cases} M_{i\to j} = f_{ij}\Big(B_i - \alpha_{ij} M_{j\to i}\Big) & \text{(A.23a)} \\ B_i = \kappa_i\left(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\right) & \text{(A.23b)} \end{cases}$$

as

$$g_{ij}(x) = \frac{1}{\alpha_{ij}}\phi^{-1}\left(\phi\Big(\alpha_{ij}\beta_{ij}J_{ij}\Big)\phi\big(x\big)\right) \approx \phi^{-1}\left(\phi\Big(\beta_{ij}J_{ij}\Big)\phi\big(x\big)\right) = f_{ij}(x) \tag{A.24}$$

This can be rewritten:

$$\begin{cases} M_{i\to j} = \phi^{-1}\big(\phi(\beta_{ij}J_{ij})\phi(B_i - \alpha_{ij}M_{j\to i})\big) & \text{(A.25a)} \\ B_i = \kappa_i\Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big) & \text{(A.25b)} \end{cases}$$

instead of the initial system (extended Fractional BP):

$$\begin{cases} M_{i\to j} = \dfrac{1}{\alpha_{ij}}\phi^{-1}\big(\phi(\alpha_{ij}\beta_{ij}J_{ij})\phi(B_i - \alpha_{ij}M_{j\to i})\big) & \text{(A.26a)} \\ B_i = \kappa_i\Big(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + \gamma_i M_{\text{ext}\to i}\Big) & \text{(A.26b)} \end{cases}$$

Bibliography

Abbott, L. F., DePasquale, B., and Memmesheimer, R.-M. (2016). Building functional networks of spiking model neurons. *Nature neuroscience*, 19(3):350–5.

Adams, R. A., Napier, G., Roiser, J. P., Mathys, C. D., and Gilleen, J. (2018). Attractor-like Dynamics in Belief Updating in Schizophrenia. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 38(44):9471–9485.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, 4:47.

Adler, C. M., Malhotra, A. K., Elman, I., Goldberg, T., Egan, M., Pickar, D., and Breier, A. (1999). Comparison of Ketamine-Induced Thought Disorder in Healthy Volunteers and Thought Disorder in Schizophrenia. *American Journal of Psychiatry*, 156(10):1646–1649.

Ahmadian, Y. and Miller, K. D. (2021). What is the dynamical regime of cerebral cortex? *Neuron*, 0(0).

Aitchison, L. and Lengyel, M. (2016). The Hamiltonian Brain: Efficient Probabilistic Inference with Excitatory-Inhibitory Neural Circuit Dynamics. *PLOS Computational Biology*, 12(12):e1005186.

Aitchison, L. and Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227.

Albert, S., Schmack, K., Sterzer, P., and Schneider, G. (2017). A hierarchical stochastic model for bistable perception. *PLOS Computational Biology*, 13(11):e1005856.

Alderson-Day, B., Lima, C. F., Evans, S., Krishnan, S., Shanmugalingam, P., Fernyhough, C., and Scott, S. K. (2017). Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. *Brain*, 140(9):2475–2489.

Ankan, A. and Panda, A. (2015). pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer.

Anticevic, A., Murray, J. D., and Barch, D. M. (2015). Bridging Levels of Understanding in Schizophrenia Through Computational Modeling.

Arnold, D. H. (2011). Why is Binocular Rivalry Uncommon? Discrepant Monocular Images in the Real World. *Frontiers in Human Neuroscience*, 5(OCTOBER).

Averbeck, B. B., Evans, S., Chouhan, V., Bristow, E., and Shergill, S. S. (2011). Probabilistic learning and inference in schizophrenia. *Schizophrenia research*, 127(1-3):115.

Barlow, H. B. (1969). Pattern recognition and the responses of sensory neurons. *Annals of the New York Academy of Sciences*, 156(2):872–881.

Baum, G. L., Ciric, R., Roalf, D. R., Betzel, R. F., Moore, T. M., Shinohara, R. T., Kahn, A. E., Vandekar, S. N., Rupert, P. E., Quarmley, M., Cook, P. A., Elliott, M. A., Ruparel, K., Gur, R. E., Gur, R. C., Bassett, D. S., and Satterthwaite, T. D. (2017). Modular Segregation of Structural Brain Networks Supports the Development of Executive Function in Youth. *Current Biology*, 27(11):1561–1572.e8.

Baumeister, D., Sedgwick, O., Howes, O., and Peters, E. (2017). Auditory verbal hallucinations and continuum models of psychosis: A systematic review of the healthy voice-hearer literature. *Clinical Psychology Review*, 51:125–141.

Baxter, R. J. (1982). *Exactly solved models in statistical mechanics*.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, 53:370–418.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., and Pouget, A. (2012). Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, 74(1):30–39.

Beck, J. M. and Pouget, A. (2007). Exact inferences in a neural implementation of a hidden markov model. *Neural Computation*, 19:1344–1361.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science (New York, N.Y.)*, 331(6013):83–7.

Berrou, C., Glavieux, A., and Thitimajshima, P. (1993). Near shannon limit error-correcting coding and decoding: Turbo-codes. 1. volume 2, pages 1064 – 1070 vol.2.

Bertolero, M. A., Yeo, B. T., Bassett, D. S., and D'Esposito, M. (2018). A mechanistic model of connector hubs, modularity and cognition.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bitzer, S., Bruineberg, J., and Kiebel, S. J. (2015). A Bayesian Attractor Model for Perceptual Decision Making. *PLOS Computational Biology*, 11(8):e1004442.

Blake, R. (1989). A neural theory of binocular rivalry. *Psychological review*, 96(1):145–167.

Blake, R. and Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience 2001 3:1*, 3(1):13–21.

Boerlin, M. and Denève, S. (2011). Spike-Based Population Coding and Working Memory. *PLoS Comput Biol*, 7(2).

Boerlin, M., Machens, C. K., and Denève, S. (2013). Predictive Coding of Dynamical Variables in Balanced Spiking Networks. *PLoS Computational Biology*, 9(11):e1003258.

Bouttier, V., Duttagupta, S., Denève, S., and Jardri, R. (2021). Circular inference predicts nonuniform overactivation and dysconnectivity in brain-wide connectomes. *Schizophrenia Research*.

Brandl, F., Avram, M., Weise, B., Shang, J., Simões, B., Bertram, T., Ayala, D. H., Penzel, N., Gürsel, D. A., Bäuml, J., Wohlschläger, A. M., Vukadinovic, Z., Koutsouleris, N., Leucht, S., and Sorg, C. (2019). Specific Substantial Dysconnectivity in Schizophrenia: A Transdiagnostic Multimodal Meta-analysis of Resting-State Functional and Structural Magnetic Resonance Imaging Studies. *Biological Psychiatry*, 85(7):573–583.

Brascamp, J., Blake, R., and Knapen, T. (2015a). Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. *Nature Neuroscience 2015 18:11*, 18(11):1672–1678.

Brascamp, J., Klink, P., and Levelt, W. (2015b). The 'laws' of binocular rivalry: 50 years of Levelt's propositions. *Vision research*, 109(Pt A):20–37.

Brascamp, J., Sohn, H., Lee, S.-H., and Blake, R. (2013). A monocular contribution to stimulus rivalry. *Proceedings of the National Academy of Sciences*, 110(21):8337–8344.

Brascamp, J., Sterzer, P., Blake, R., and Knapen, T. (2018). Multistable Perception and the Role of the Frontoparietal Cortex in Perceptual Inference. *Annual review of psychology*, 69:77–103.

Brascamp, J., van Ee, R., Pestman, W., and van den Berg, A. (2005). Distributions of alternation rates in various forms of bistable perception. *Journal of vision*, 5(4):287–298.

Braunstein, A. and Zecchina, R. (2004). Survey propagation as local equilibrium equations. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(06):P06007.

Brendel, W., Bourdoukan, R., Vertechi, P., Machens, C. K., and Denève, S. (2020). Learning to represent signals spike by spike. *PLoS Computational Biology*, 16(3):e1007692.

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. *PLoS Computational Biology*, 7(11):e1002211.

Bullier, J., Hupé, J., James, A., and Girard, P. (2001). The role of feedback connections in shaping the responses of visual cortical neurons. *Progress in brain research*, 134:193–204.

Bullmore, E. T. and Bassett, D. S. (2011). Brain Graphs: Graphical Models of the Human Brain Connectome. *http://dx.doi.org/10.1146/annurev-clinpsy-040510-143934*, 7:113–140.

Canitano, R. and Pallagrosi, M. (2017). Autism Spectrum Disorders and Schizophrenia Spectrum Disorders: Excitation/Inhibition Imbalance and Developmental Trajectories. *Frontiers in psychiatry*, 8(MAY).

Carpenter, R. H. S. and Williams, M. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, 377:59.

Carrera, E. and Tononi, G. (2014). Diaschisis: past, present, future. *Brain*, 137(9):2408–2422.

Cascella, N. G., Schretlen, D. J., and Sawa, A. (2009). Schizophrenia and epilepsy: is there a shared susceptibility? *Neuroscience research*, 63(4):227–235.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Chettih, S. N. and Harvey, C. D. (2019). Single-neuron perturbations reveal feature-specific competition in V1. *Nature*, 567(7748):334–340.

Chrysaitis, N. A., Jardri, R., Denève, S., and Seriès, P. (2021). No increased circular inference in autism or autistic traits. *bioRxiv*, page 2021.04.28.441748.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science.

Corlett, P. R., Frith, C. D., and Fletcher, P. C. (2009). From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4):515–530.

Corlett, P. R., Honey, G., and Fletcher, P. C. (2007). From prediction error to psychosis: ketamine as a pharmacological model of delusions. *Journal of Psychopharmacology*, 21(3):238–252.

Corlett, P. R., Honey, G. D., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Absalom, A. R., Lee, M., Pomarol-Clotet, E., Murray, G. K., McKenna, P. J., Robbins, T. W., Bullmore, E. T., and Fletcher, P. C. (2006). Frontal responses during learning predict vulnerability to the psychotogenic effects of ketamine: Linking cognition, brain activity, and psychosis. *Archives of General Psychiatry*, 63(6):611–621.

Corlett, P. R., Honey, G. D., and Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology*, 30(11):1145–1155.

Corlett, P. R., Honey, G. D., Krystal, J. H., and Fletcher, P. C. (2011). Glutamatergic Model Psychoses: Prediction Error, Learning and Inference. *Neuropsychopharmacology*, 36(1):294–315.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., and Powers, A. R. (2019). Hallucinations and Strong Priors. *Trends in cognitive sciences*, 23(2):114–127.

Crossley, N. A., Mechelli, A., Ginestet, C., Rubinov, M., Bullmore, E. T., and Mcguire, P. (2016). Altered hub functioning and compensatory activations in the connectome: A meta- Analysis of functional neuroimaging studies in schizophrenia. *Schizophrenia Bulletin*, 42(2):434–442.

Crossley, N. A., Mechelli, A., Scott, J., Carletti, F., Fox, P. T., Mcguire, P., and Bullmore, E. T. (2014). The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain*, 137(8):2382–2395.

Cseke, B. and Heskes, T. (2011). Properties of Bethe Free Energies and Message Passing in Gaussian Models. *Journal of Artificial Intelligence Research*, 41:1–24.

Ćurčić-Blake, B., Ford, J. M., Hubl, D., Orlov, N. D., Sommer, I. E., Waters, F., Allen, P., Jardri, R., Woodruff, P. W., David, O., Mulert, C., Woodward, T. S., and Aleman, A. (2017). Interaction of language, auditory and memory brain networks in auditory verbal hallucinations. *Progress in Neurobiology*, 148:1–20.

Dayan, P. (1998). A hierarchical model of binocular rivalry. *Neural computation*, 10(5):1119–1135.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Denève, S. (2005). Bayesian inference in spiking neurons. In *Advances in neural information processing systems*.

Denève, S. (2008). Bayesian Spiking Neurons I: Inference. *Neural Computation*, 20(1):91–117.

Denève, S., Duhamel, J.-R., and Pouget, A. (2007). Optimal Sensorimotor Integration in Recurrent Cortical Networks: A Neural Implementation of Kalman Filters. *Journal of Neuroscience*, 27(21):5744–5756.

Denève, S. and Jardri, R. (2016). Circular inference: mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences*, 11:40–48.

Denève, S., Latham, P., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature neuroscience*, 4:826–31.

Denève, S. and Machens, C. K. (2016). Efficient codes and balanced networks. *Nature Neuroscience*, 19(3):375–382.

Diaconescu, A. O., Litvak, V., Mathys, C. D., Kasper, L., Friston, K. J., and Stephan, K. E. (2017). A computational hierarchy in human cortex. *arXiv*.

Douglas, R., Koch, C., Mahowald, M., Martin, K., and Suarez, H. (1995). Recurrent excitation in neocortical circuits. *Science (New York, N.Y.)*, 269(5226):981–985.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. The MIT Press.

Dudley, R., Taylor, P., Wickham, S., and Hutton, P. (2016). Psychosis, Delusions and the "Jumping to Conclusions" Reasoning Bias: A Systematic Review and Meta-analysis. *Schizophrenia bulletin*, 42(3):652–65.

Echeveste, R., Aitchison, L., Hennequin, G., and Lengyel, M. (2020). Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, 23(9):1138–1149.

Eliasmith, C. and Trujillo, O. (2014). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology*, 25:1–6.

Evans, S. L., Averbeck, B. B., and Furl, N. (2015). Jumping to conclusions in schizophrenia. *Neuropsychiatric disease and treatment*, 11:1615–24.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, 1(1):1–47.

Finlayson, N., Zhang, X., and Golomb, J. (2017). Differential patterns of 2D location versus depth decoding along the visual hierarchy. *NeuroImage*, 147:507–516.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–30.

Fletcher, P. C. and Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58.

Forney, G. (2001). Codes on graphs: normal realizations. *IEEE Transactions on Information Theory*, 47(2):520–548.

Fornito, A., Zalesky, A., and Breakspear, M. (2015). The connectomics of brain disorders.

Foss-Feig, J. H., Adkinson, B. D., Ji, J. L., Yang, G. J., Srihari, V. H., McPartland, J. C., Krystal, J. H., Murray, J. D., and Anticevic, A. (2017). Searching for Cross-Diagnostic Convergence: Neural Mechanisms Governing Excitation and Inhibition Balance in Schizophrenia and Autism Spectrum Disorders. *Biological psychiatry*, 81(10):848–861.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Essen, D. C. V., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678.

Frey, B. J. and MacKay, D. (1998). A Revolution: Belief Propagation in Graphs with Cycles. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Friston, K. (2008). Hierarchical Models in the Brain. *PLOS Computational Biology*, 4(11):e1000211.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1456):815–36.

Friston, K. J. (2020). Bayesian Dysconnections. *https://doi.org/10.1176/appi.ajp.2020.20091421*, 177(12):1110–1112.

Friston, K. J., Brown, H. R., Siemerkus, J., and Stephan, K. E. (2016). The dysconnection hypothesis (2016). *Schizophrenia Research*, 176(2-3):83–94.

Friston, K. J., Parr, T., and de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, 1(4):381–414.

Gallager, R. (1962). Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1):21–28.

Gao, R. and Penzes, P. (2015). Common Mechanisms of Excitatory and Inhibitory Imbalance in Schizophrenia and Autism Spectrum Disorders. *Current molecular medicine*, 15(2):146.

Garety, P. A., Hemsley, D. R., and Wessely, S. (1991). Reasoning in deluded schizophrenic and paranoid patients - biases in performance on a probabilistic inference task. *Journal of Nervous and Mental Disease*, 179(4):194–201.

Geisler, W. S. and Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience 2002 5:6*, 5(6):508–510.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

George, D. and Hawkins, J. (2009). Towards a Mathematical Theory of Cortical Micro-circuits. *PLoS Computational Biology*, 5(10):e1000532.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, 24(1):1–24.

Gigante, G., Mattia, M., Braun, J., and Giudice, P. D. (2009). Bistable Perception Modeled as Competing Stochastic Integrations at Two Levels. *PLoS Computational Biology*, 5(7):1000430.

Glöckner, A. and Moritz, S. (2008). A Fine-grained Analysis of the Jumping to Conclusions Bias in Schizophrenia: Data-Gathering, Response Confidence, and Information Integration. *Discussion Paper Series of the Max Planck Institute for Research on Collective Goods*.

Gold, J. I. and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5(1):10–16.

Guo, S., Palaniyappan, L., Yang, B., Liu, Z., Xue, Z., and Feng, J. (2014). Anatomical distance affects functional connectivity in patients with schizophrenia and their siblings. *Schizophrenia Bulletin*, 40(2):449–459.

Hadley, J. A., Kraguljac, N. V., White, D. M., Ver Hoef, L., Tabora, J., and Lahti, A. C. (2016). Change in brain network topology as a function of treatment response in schizophrenia: a longitudinal resting-state fMRI study using graph theory. *npj Schizophrenia 2016 2:1*, 2(1):1–7.

Hennequin, G., Aitchison, L., and Lengyel, M. (2014). Fast sampling-based inference in balanced neuronal networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Heskes, T. (2002). Stable Fixed Points of Loopy Belief Propagation Are Minima of the Bethe Free Energy. In *Advances in neural information processing systems*.

Heskes, T. (2004). On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11):2379–2413.

Heskes, T., Albers, K., and Kappen, B. (2002). Approximate inference and constrained optimization. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, page 313–320, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hilgetag, C. C. and Goulas, A. (2020). 'Hierarchy' in the organization of brain networks. *Philosophical Transactions of the Royal Society B*, 375(1796).

Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3):687–701.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106.

Huguet, G., Rinzel, J., and Hupé, J. (2014). Noise and adaptation in multistable perception: noise drives when to switch, adaptation determines percept choice. *Journal of vision*, 14(3):1–24.

Hupé, J., James, A., Payne, B., Lomber, S., Girard, P., and Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, 394(6695):784–787.

Huq, S. F., Garety, P. A., and Hemsley, D. R. (1988). Probabilistic Judgements in Deluded and Non-Deluded Subjects. *The Quarterly Journal of Experimental Psychology Section A*, 40(4):801–812.

Ihler, A. and McAllester, D. (2009). Particle belief propagation. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 256–263, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Ihler, A. T. (2007). Accuracy Bounds for Belief Propagation. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, pages 183–190, Arlington, Virginia, USA. AUAI Press.

Ihler, A. T., Fisher III, J. W., and Willsky, A. S. (2005). Loopy Belief Propagation: Convergence and Effects of Message Errors. In *Journal of Machine Learning Research*, volume 6, pages 905–936.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31(1):253–258.

Jardri, R. and Denève, S. (2013a). Circular inferences in schizophrenia. *Brain*, 136(11):3227–3241.

Jardri, R. and Denève, S. (2013b). Computational Models of Hallucinations. In *The Neuroscience of Hallucinations*, pages 289–313. Springer New York, New York, NY.

Jardri, R., Duverne, S., Litvinova, A. S., and Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, 8:14218.

Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., Smailes, D., Sterzer, P., Corlett, P. R., Leptourgos, P., Debbané, M., Cachia, A., and Denève, S. (2016). Are Hallucinations Due to an Imbalance Between Excitatory and Inhibitory Influences on the Brain? *Schizophrenia Bulletin*, 42(5):1124–1134.

Jardri, R., Pouchet, A., Pins, D., and Thomas, P. (2011). Cortical activations during auditory verbal hallucinations in schizophrenia: A coordinate-based meta-analysis. *American Journal of Psychiatry*, 168(1):73–81.

Karch, S., Segmiller, F., Hantschk, I., Cerovecki, A., Opgen-Rhein, M., Hock, B., Dargel, S., Leicht, G., Hennig-Fast, K., Riedel, M., and Pogarell, O. (2012). Increased γ oscillations during voluntary selection processes in adult patients with attention deficit/hyperactivity disorder. *Journal of psychiatric research*, 46(11):1515–1523.

Kayser, M. S. and Dalmau, J. (2016). Anti-NMDA receptor encephalitis, autoimmunity, and psychosis. *Schizophrenia research*, 176(1):36–40.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, 55:271–304.

Klink, P. C., van Ee, R., and van Wezel, R. J. A. (2008). General Validity of Levelt's Propositions Reveals Common Computational Mechanisms for Visual Rivalry. *PLOS ONE*, 3(10):e3473.

Knapen, T., Brascamp, J., Pearson, J., van Ee, R., and Blake, R. (2011). The Role of Frontal and Parietal Brain Areas in Bistable Perception. *The Journal of Neuroscience*, 31(28):10293.

Knierim, J. and van Essen, D. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of neurophysiology*, 67(4):961–980.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Knoll, C. and Pernkopf, F. (2017). On Loopy Belief Propagation - Local Stability Analysis for Non-Vanishing Fields. In *Proceedings of the conference on Uncertainty in Artificial Intelligence*.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press.

Körding, K. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427:244–247.

Kornmeier, J., Ehm, W., Bigalke, H., and Bach, M. (2007). Discontinuous presentation of ambiguous figures: How interstimulus-interval durations affect reversal dynamics and ERPs. *Psychophysiology*, 44(4):552–560.

Krystal, J. H., Karper, L. P., Seibyl, J. P., Freeman, G. K., Delaney, R., Bremner, J. D., Heninger, G. R., Bowers, M. B., and Charney, D. S. (1994). Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. Psychotomimetic, perceptual, cognitive, and neuroendocrine responses. *Archives of general psychiatry*, 51(3):199–214.

Krystal, J. H., Murray, J. D., Chekroud, A. M., Corlett, P. R., Yang, G., Wang, X.-J., and Anticevic, A. (2017). Computational Psychiatry and the Challenge of Schizophrenia. *Schizophrenia Bulletin*, 43(3):473–475.

Krystal, J. H., Perry, E., Gueorguieva, R., Belger, A., Madonick, S., Abi-Dargham, A., Cooper, T., Macdougall, L., Abi-Saab, W., and D'Souza, D. (2005). Comparative and interactive human psychopharmacologic effects of ketamine and amphetamine: implications for glutamatergic and dopaminergic model psychoses and cognitive function. *Archives of general psychiatry*, 62(9):985–995.

Kschischang, F., Frey, B., and Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

Kuck, J., Chakraborty, S., Tang, H., Luo, R., Song, J., Sabharwal, A., and Ermon, S. (2020). Belief propagation neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 667–678. Curran Associates, Inc.

Lago-Fernández, L. F. and Deco, G. (2002). A model of binocular rivalry based on competition in IT. *Neurocomputing*, 44-46:503–507.

Laing, C. R. and Chow, C. C. (2002). A Spiking Neuron Model for Binocular Rivalry. *Journal of Computational Neuroscience 2002 12:1*, 12(1):39–53.

Lam, N. H., Borduqui, T., Hallak, J., Roque, A. C., Anticevic, A., Krystal, J. H., Wang, X.-J., and Murray, J. D. (2017). Effects of Altered Excitation-Inhibition Balance on Decision Making in a Cortical Circuit Model. *bioRxiv*, page 100347.

Lanillos, P., Oliva, D., Philippsen, A., Yamashita, Y., Nagai, Y., and Cheng, G. (2020). A review on neural network models of schizophrenia and autism spectrum disorder. *Neural Networks*, 122:338–363.

Lapicque, L. (1907). Recherches quantitatives sur l'excitation electrique des nerfs traitée comme une polarization. *Journal de Physiologie et Pathologie General*, 9:620–635.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 20(7):1434–1448.

Lehky, S. R. (1995). Binocular rivalry is not chaotic. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 259(1354):71–76.

Leisenberger, H., Knoll, C., Seeber, R., and Pernkopf, F. (2021). Convergence behavior of belief propagation: Estimating regions of attraction via lyapunov functions. In *Uncertainty in Artificial Intelligence*. 37th Conference on Uncertainty in Artificial Intelligence : UAI 2021 ; Conference date: 27-07-2021 Through 29-07-2021.

Leopold, D., Wilke, M., Maier, A., and Logothetis, N. (2002). Stable perception of visually ambiguous patterns. *Nature neuroscience*, 5(6):605–609.

Leptourgos, P., Bouttier, V., Denève, S., and Jardri, R. (2021). From hallucinations to synaesthesia: a circular inference account of unimodal and multimodal erroneous percepts in clinical and drug-induced psychosis. *PsyArXiv*.

Leptourgos, P., Bouttier, V., Jardri, R., and Denève, S. (2020a). A functional theory of bistable perception based on dynamical circular inference. *PLoS Computational Biology*, 16(12).

Leptourgos, P., Denève, S., and Jardri, R. (2017). Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Current Opinion in Neurobiology*, 46:154–161.

Leptourgos, P., Notredame, C.-E., Eck, M., Jardri, R., and Denève, S. (2020b). Circular inference in bistable perception. *Journal of Vision*, 20(4).

Levelt, W. (1967). Note on the distribution of dominance times in binocular rivalry. *British journal of psychology (London, England : 1953)*, 58(1):143–145.

Levelt, W. J. M. (1966). The alternation process in binocular rivalry. *British Journal of Psychology*, 57(3-4):225–238.

Lewis, D. A., Curley, A. A., Glausier, J., and Volk, D. W. (2012). Cortical Parvalbumin Interneurons and Cognitive Dysfunction in Schizophrenia. *Trends in Neurosciences*, 35(1):57.

Lewis, D. A. and Gonzalez-Burgos, G. (2006). Pathophysiologically based treatment interventions in schizophrenia. *Nature medicine*, 12(9):1016–1022.

Lewis, D. A., Hashimoto, T., and Volk, D. W. (2005). Cortical inhibitory neurons and schizophrenia. *Nature Reviews Neuroscience 2005 6:4*, 6(4):312–324.

Li, P., Fan, T.-T., Zhao, R.-J., Han, Y., Shi, L., Sun, H.-Q., Chen, S.-J., Shi, J., Lin, X., and Lu, L. (2017). Altered Brain Network Connectivity as a Potential Endophenotype of Schizophrenia. *Scientific Reports 2017 7:1*, 7(1):1–9.

Li, Q. and Pehlevan, C. (2020). Minimax Dynamics of Optimally Balanced Spiking Networks of Excitatory and Inhibitory Neurons. In *Advances in Neural Information Processing Systems*, volume 33, pages 4894–4904.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., and Stephan, K. E. (2013). Modelling Trial-by-Trial Changes in the Mismatch Negativity. *PLOS Computational Biology*, 9(2):e1002911.

Lisman, J. (2012). Excitation, inhibition, local oscillations, or large-scale loops: What causes the symptoms of schizophrenia?

Litvak, S., Karlinsky, L., and Ullman, S. (2009). Properties of Cortical Networks Improve Inference in Highly Interconnected Graphical Models.

Litvak, S. and Ullman, S. (2009). Cortical Circuitry Implementing Graphical Models. *Neural Computation*, 21(11):3010–3056.

Liu, D., Moghadam, N. N., Rasmussen, L. K., Huang, J., and Chatterjee, S. (2019). α Belief Propagation as Fully Factorized Approximation. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5.

Liu, D., Vu, M. T., Li, Z., and Rasmussen, L. K. (2020). α Belief propagation for approximate inference. *arXiv*.

Liu, G. (2004). Local structural balance and functional interaction of excitatory and inhibitory synapses in hippocampal dendrites. *Nature neuroscience*, 7(4):373–379.

Lochmann, T. and Denève, S. (2011). Neural processing as causal inference. *Current Opinion in Neurobiology*, 21(5):774–781.

Loh, M., Rolls, E. T., and Deco, G. (2007). A Dynamical Systems Hypothesis of Schizophrenia. *PLoS Computational Biology*, 3(11):2255–2265.

Lord, L.-D., Stevner, A. B., Deco, G., and Kringelbach, M. L. (2017). Understanding principles of integration and segregation using whole-brain computational connectomics: implications for neuropsychiatric disorders. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 375(2096):20160283.

Lucas-Meunier, E., Monier, C., Amar, M., Baux, G., Frégnac, Y., and Fossier, P. (2009). Involvement of Nicotinic and Muscarinic Receptors in the Endogenous Cholinergic Modulation of the Balance between Excitation and Inhibition in the Young Rat Visual Cortex. *Cerebral Cortex*, 19(10):2411–2427.

Lumer, E. D., Friston, K. J., and Rees, G. (1998). Neural Correlates of Perceptual Rivalry in the Human Brain. *Science*, 280(5371):1930–1934.

Luscher, B., Shen, Q., and Sahir, N. (2011). The GABAergic Deficit Hypothesis of Major Depressive Disorder. *Molecular psychiatry*, 16(4):383.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Maier, A., Wilke, M., Logothetis, N. K., and Leopold, D. A. (2003). Perception of Temporally Interleaved Ambiguous Patterns. *Current Biology*, 13(13):1076–1085.

Malioutov, D. M. (2008). Approximate Inference in Gaussian Graphical Models. *Thesis*, page 169.

Mamassian, P. and Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, 38(18):2817–2832.

Manita, S., Suzuki, T., Homma, C., Matsumoto, T., Odagawa, M., Yamada, K., Ota, K., Matsubara, C., Inutsuka, A., Sato, M., Ohkura, M., Yamanaka, A., Yanagawa, Y., Nakai, J., Hayashi, Y., Larkum, M., and Murayama, M. (2015). A Top-Down Cortical Circuit for Accurate Sensory Perception. *Neuron*, 86(5):1304–1316.

Maraš, M. (2019). *Learning efficient signal representation in sparse spike-coding networks*. Theses, Université Paris sciences et lettres.

Marín, O. (2012). Interneuron dysfunction in psychiatric disorders. *Nature Reviews Neuroscience*, 13(2):107–120.

Markicevic, M., Fulcher, B. D., Lewis, C., Helmchen, F., Rudin, M., Zerbi, V., and Wenderoth, N. (2020). Cortical Excitation:Inhibition Imbalance Causes Abnormal Brain Network Dynamics as Observed in Neurodevelopmental Disorders. *Cerebral cortex (New York, N.Y. : 1991)*, 30(9):4922–4937.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

Mastrogiuseppe, F. (2017). *From dynamics to computations in recurrent neural networks*. Theses, Université Paris sciences et lettres.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., and Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in human neuroscience*, 8:825.

Miao, A., Liu, Q., Li, Z., Liu, W., Wang, L., Ge, J., Yu, C., Wang, Y., Huang, S., Yu, Y., Shi, Q., Sun, J., and Wang, X. (2020). Altered cerebral blood flow in patients with anti-NMDAR encephalitis. *Journal of Neurology 2020 267:6*, 267(6):1760–1773.

Minka, T. (2001a). The EP Energy Function and Minimization Schemes. Technical report.

Minka, T. (2004). Power EP. Technical report.

Minka, T. (2005). Divergence measures and message passing. Technical report.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 352–359, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Minka, T. P. (2001b). Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, volume 17, pages 362–369.

Mongillo, G. and Deneve, S. (2008). Online Learning with Hidden Markov Models. *Neural Computation*, 20(7):1706–1716.

Mooij, J. M. (2010). libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173.

Mooij, J. M. and Kappen, H. (2009). Bounds on marginal probability distributions. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Mooij, J. M. and Kappen, H. J. (2004). Validity estimates for loopy Belief Propagation on binary real-world networks. In *Advances in neural information processing systems*.

Mooij, J. M. and Kappen, H. J. (2005). On the properties of the Bethe approximation and loopy belief propagation on binary networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11012—-P11012.

Mooij, J. M. and Kappen, H. J. (2007a). Loop corrections for approximate inference on factor graphs. *J. Mach. Learn. Res.*, 8:1113–1143.

Mooij, J. M. and Kappen, H. J. (2007b). Sufficient conditions for convergence of the sum–product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437.

Moreno-Bote, R. and Drugowitsch, J. (2015). Causal Inference and Explaining Away in a Spiking Network. *Scientific Reports*, 5(1):17531.

Moreno-Bote, R., Rinzel, J., and Rubin, N. (2007). Noise-induced alternations in an attractor network model of perceptual bistability. *Journal of neurophysiology*, 98(3):1125–1139.

Moreno-Bote, R., Shpiro, A., Rinzel, J., and Rubin, N. (2010). Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. *Journal of vision*, 10(11):1.

Moritz, S. and Woodward, T. S. (2005). Jumping to conclusions in delusional and non-delusional schizophrenic patients. *The British journal of clinical psychology*, 44(Pt 2):193–207.

Muldoon, S., Bridgeford, E., and Bassett, D. S. (2016). Small-World Propensity and Weighted Brain Networks. *Scientific reports*, 6.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 467–475, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Murray, J. D. and Anticevic, A. (2017). Toward understanding thalamocortical dysfunction in schizophrenia through computational models of neural circuit dynamics. *Schizophrenia Research*, 180:70–77.

Murray, J. D., Anticevic, A., Gancsos, M., Ichinose, M., Corlett, P. R., Krystal, J. H., and Wang, X.-J. (2014). Linking Microcircuit Dysfunction to Cognitive Impairment: Effects of Disinhibition Associated with Schizophrenia in a Cortical Working Memory Model. *Cerebral Cortex*, 24(4):859–872.

Nawrot, M. and Blake, R. (1989). Neural integration of information specifying structure from stereopsis and motion. *Science*, 244(4905):716–718.

Newcomer, J. W., Farber, N. B., Jevtovic-Todorovic, V., Selke, G., Melson, A. K., Hershey, T., Craft, S., and Olney, J. W. (1999). Ketamine-Induced NMDA Receptor Hypofunction as a Model of Memory Impairment and Psychosis. *Neuropsychopharmacology*, 20(2):106–118.

Noest, A., van Ee, R., Nijs, M., and van Wezel, R. (2007). Percept-choice sequences driven by interrupted ambiguous stimuli: a low-level neural model. *Journal of vision*, 7(8):1–14.

Notredame, C.-E., Pins, D., Denève, S., and Jardri, R. (2014). What visual illusions teach us about schizophrenia. *Frontiers in integrative neuroscience*, 8:63.

Okun, M. and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature Neuroscience 2008 11:5*, 11(5):535–537.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *The Journal of Machine Learning Research*, 6:2177–2204.

Orbach, J., Ehrlich, D., and Heath, H. A. (1963). Reversibility of the Necker Cube: I. An Examination of the Concept of "Satiation of Orientation". *Perceptual and motor skills*, 17:439–458.

Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543.

Ott, T. and Stoop, R. (2006). The Neurodynamics of Belief Propagation on Binary Markov Random Fields. In *Advances in neural information processing systems*, volume 30.

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5(5):1–1.

Panagiotaropoulos, T. I., Kapoor, V., Logothetis, N. K., and Deco, G. (2013). A Common Neurodynamical Mechanism Could Mediate Externally Induced and Intrinsically Generated Transitions in Visual Awareness. *PLOS ONE*, 8(1):e53833.

Pantano, P., Baron, J. C., Samson, Y., Bousser, M. G., Derouesne, C., and Comar, D. (1986). Crossed cerebellar diaschisis. Further studies. *Brain*, 109(4):677–694.

Parenti, A., Jardri, R., and Geoffroy, P. A. (2016). How Anti-NMDAR Encephalitis Sheds Light on the Mechanisms Underlying Catatonia: The Neural Excitatory/Inhibitory Imbalance Model. *Psychosomatics*, 57(3):336–338.

Parr, T. and Friston, K. J. (2018). The anatomy of inference: Generative models and brain structure. *Frontiers in Computational Neuroscience*, 12:90.

Parr, T., Markovic, D., Kiebel, S. J., and Friston, K. J. (2019). Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific Reports*, 9(1):1889.

Pastukhov, A. and Braun, J. (2011). Cumulative history quantifies the role of neural adaptation in multistable perception. *Journal of vision*, 11(10):12–12.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Paus, T., Keshavan, M., and Giedd, J. N. (2008). Why do many psychiatric disorders emerge during adolescence? *Nature Reviews Neuroscience 2008 9:12*, 9(12):947–957.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearson, J. and Brascamp, J. (2008). Sensory memory for ambiguous vision. *Trends in cognitive sciences*, 12(9):334–341.

Peer, M., Prüss, H., Ben-Dayan, I., Paul, F., Arzy, S., and Finke, C. (2017). Functional connectivity of large-scale brain networks in patients with anti-NMDA receptor encephalitis: an observational study. *The Lancet Psychiatry*, 4(10):768–774.

Peters, E., Joseph, S., Day, S., and Garety, P. (2004). Measuring Delusional Ideation: The 21-Item Peters et al. Delusions Inventory (PDI). *Schizophrenia Bulletin*, 30(4):1005–1022.

Peterson, C. and Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019.

Pitkow, X. and Angelaki, D. E. (2017). Inference in the Brain: Statistics Flowing in Redundant Population Codes. *Neuron*, 94(5):943–953.

Plarre, K. and Kumar, P. (2004). Extended message passing algorithm for inference in loopy gaussian graphical models. *Ad Hoc Networks*, 2(2):153 – 169.

Poirazi, P., Brannon, T., and Mel, B. W. (2003). Pyramidal Neuron as Two-Layer Neural Network. *Neuron*, 37(6):989–999.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–8.

Powers, A. R., Mathys, C. D., and Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351):596–600.

Price, C. J., Warburton, E. A., Moore, C. J., Frackowiak, R. S. J., and Friston, K. J. (2001). Dynamic Diaschisis: Anatomically Remote and Context-Sensitive Human Brain Lesions. *Journal of Cognitive Neuroscience*, 13(4):419–429.

Probst, D., Petrovici, M. A., Bytschok, I., Bill, J., Pecevski, D., Schemmel, J., and Meier, K. (2015). Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons. *Frontiers in Computational Neuroscience*, 9:13.

Raju, R. V. and Pitkow, Z. (2016). Inference by Reparameterization in Neural Population Codes. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Rao, R. P. (2004). Bayesian Computation in Recurrent Neural Circuits. *Neural Computation*, 16(1):1–38.

Rao, R. P. (2005a). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, 16(16):1843–8.

Rao, R. P. (2005b). Hierarchical Bayesian Inference in Networks of Spiking Neurons. In *Advances in neural information processing systems*.

Rao, R. P. (2006). Neural Models of Bayesian Belief Propagation. In Doya, K., Ishii, A., Pouget, A., and Rao, R. P., editors, *The Bayesian Brain: Probabilistic Approaches to Neural Coding*, pages 235–260. MIT Press.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, 20(4):260–281.

Reichert, D., Series, P., and Storkey, A. J. (2011). Neuronal Adaptation for Sampling-Based Probabilistic Inference in Perceptual Bistability. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. D. (2010). The asynchronous state in cortical circuits. *Science*, 327(5965):587–590.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591 vol.1.

Rohe, T. and Noppeney, U. (2015). Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception. *PLOS Biology*, 13(2):e1002073.

Rolls, E. T. and Deco, G. (2011). A computational neuroscience approach to schizophrenia and its onset.

Rolls, E. T., Joliot, M., and Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, 122:1–5.

Rolls, E. T., Loh, M., Deco, G., and Winterer, G. (2008). Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nature Reviews Neuroscience*, 9(9):696–709.

Rubenstein, J. L. and Merzenich, M. (2003). Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes, brain, and behavior*, 2(5):255–267.

Sakai, T., Oshima, A., Nozaki, Y., Ida, I., Haga, C., Akiyama, H., Nakazato, Y., and Mikuni, M. (2008). Changes in density of calcium-binding-protein-immunoreactive GABAergic neurons in prefrontal cortex in schizophrenia and bipolar disorder. *Neuropathology : official journal of the Japanese Society of Neuropathology*, 28(2):143–150.

Salvador, A., Arnal, L. H., Vinckier, F., Domenech, P., Gaillard, R., and Wyart, V. (2020). Premature commitment to uncertain beliefs during human NMDA receptor hypofunction. *bioRxiv*, page 2020.06.17.156539.

Schmack, K., Gòmez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., Heinz, A., Petrovic, P., and Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(34):13701–12.

Schmack, K., Rothkirch, M., Priller, J., and Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Human Brain Mapping*, 38(4):1767–1779.

Schmack, K., Schnack, A., Priller, J., and Sterzer, P. (2015). Perceptual instability in schizophrenia: Probing predictive coding accounts of delusions with ambiguous stimuli. *Schizophrenia Research: Cognition*, 2:72–77.

Schwartenbeck, P., FitzGerald, T. H., and Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, 125:578–586.

Schwemmer, M. A., Fairhall, A. L., Denève, S., and Shea-Brown, E. (2015). Constructing Precisely Computing Networks with Biophysical Spiking Neurons. *The Journal of Neuroscience*, 35(28):10112–34.

Selemon, L. D. (2013). A role for synaptic plasticity in the adolescent development of executive function. *Translational Psychiatry 2013 3:3*, 3(3):e238–e238.

Selimbeyoglu, A., Kim, C., Inoue, M., Lee, S., Hong, A., Kauvar, I., Ramakrishnan, C., Fenno, L., Davidson, T., Wright, M., and Deisseroth, K. (2017). Modulation of prefrontal cortex excitation/inhibition balance rescues social behavior in CNTNAP2-deficient mice. *Science translational medicine*, 9(401).

Selten, M., van Bokhoven, H., and Nadif Kasri, N. (2018). Inhibitory control of the excitatory/inhibitory balance in psychiatric disorders. *F1000Research*, 7:23.

Seriès, P., editor (2020). *Computational Psychiatry: A Primer*. MIT Press.

Seriès, P. (2021). The 'circular inference' model of schizophrenia gets pulled into the orbit of social cognition. *Brain*, 144(5):1293–1295.

Shadlen, M. N. and Newsome, W. T. (2001). Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology*, 86(4):1916–1936.

Shi, L. and Griffiths, T. (2009). Neural Implementation of Hierarchical Bayesian Inference by Importance Sampling. *Advances in Neural Information Processing Systems*, 22.

Shi, X., Schonfeld, D., and Tuninetti, D. (2010). Message error analysis of loopy belief propagation for the sum-product algorithm. *CoRR*, abs/1009.2305.

Shinn, A. K., Baker, J. T., Cohen, B. M., and Öngür, D. (2013). Functional connectivity of left Heschl's gyrus in vulnerability to auditory hallucinations in schizophrenia. *Schizophrenia Research*, 143(2-3):260–268.

Shon, A. P. and Rao, R. P. (2005). Implementing belief propagation in neural circuits. *Neurocomputing*, 65-66:393–399.

Shpiro, A., Curtu, R., Rinzel, J., and Rubin, N. (2007). Dynamical characteristics common to neuronal competition models. *Journal of neurophysiology*, 97(1):462–473.

Shpiro, A., Moreno-Bote, R., Rubin, N., and Rinzel, J. (2009). Balance between noise and adaptation in competition models of perceptual bistability. *Journal of computational neuroscience*, 27(1):37–54.

Shu, Y., Hasenstaub, A., and McCormick, D. A. (2003). Turning on and off recurrent balanced cortical activity. *Nature*, 423(6937):288–293.

Simonsen, A., Fusaroli, R., Petersen, M. L., Vermillet, A.-Q., Bliksted, V., Mors, O., Roepstorff, A., and Campbell-Meiklejohn, D. (2021). Taking others into account: combining directly experienced and indirect information in schizophrenia. *Brain*, 144(5):1603–1614.

Sohal, V. S. and Rubenstein, J. L. (2019). Excitation-inhibition balance as a framework for investigating mechanisms in neuropsychiatric disorders. *Molecular Psychiatry*, 24(9):1248–1257.

Sommer, I. E. C., Diederen, K. M. J., Blom, J.-D., Willems, A., Kushan, L., Slotema, K., Boks, M. P. M., Daalman, K., Hoek, H. W., Neggers, S. F. W., and Kahn, R. S. (2008). Auditory verbal hallucinations predominantly activate the right inferior frontal area. *Brain*, 131(12):3169–3177.

Speechley, W. J., Whitman, J. C., and Woodward, T. S. (2010). The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. *Journal of Psychiatry & Neuroscience*, 35(1):7.

Spratling, M. W. (2016). A neural implementation of Bayesian inference based on predictive coding. *Connection Science*, 28(4):346–383.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112:92–97.

Spratling, M. W. and Johnson, M. H. (2003). Exploring the functional significance of dendritic inhibition in cortical pyramidal cells. *Neurocomputing*, 52-54:389–395.

Steimer, A. and Douglas, R. (2013). Spike-Based Probabilistic Inference in Analog Graphical Models Using Interspike-Interval Coding. *Neural Computation*, 25(9):2303–2354.

Steimer, A., Maass, W., and Douglas, R. (2009). Belief propagation in networks of spiking neurons. *Neural Computation*, 21(219).

Stein, H., Barbosa, J., Rosa-Justicia, M., Prades, L., Morató, A., Galan-Gadea, A., Ariño, H., Martinez-Hernandez, E., Castro-Fornieles, J., Dalmau, J., and Compte, A. (2020). Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nature Communications 2020 11:1*, 11(1):1–11.

Stephan, K. E., Friston, K. J., and Frith, C. D. (2009). Dysconnection in Schizophrenia: From Abnormal Synaptic Plasticity to Failures of Self-monitoring. *Schizophrenia Bulletin*, 35(3):509–527.

Stephan, K. E. and Mathys, C. D. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, 25:85–92.

Sterzer, P., Adams, R. A., Fletcher, P. C., Frith, C. D., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P. J., Voss, M., and Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, pages 1–10.

Sterzer, P. and Kleinschmidt, A. (2007). A neural basis for inference in perceptual ambiguity. *Proceedings of the National Academy of Sciences*, 104(1):323–328.

Sterzer, P. and Rees, G. (2008). A neural basis for percept stabilization in binocular rivalry. *Journal of cognitive neuroscience*, 20(3):389–399.

Stuke, H., Stuke, H., Weilnhammer, V. A., and Schmack, K. (2017). Psychotic Experiences and Overhasty Inferences Are Related to Maladaptive Learning. *PLOS Computational Biology*, 13(1):e1005328.

Stuke, H., Weilnhammer, V. A., Sterzer, P., and Schmack, K. (2019). Delusion Proneness is Linked to a Reduced Usage of Prior Beliefs in Perceptual Decisions. *Schizophrenia Bulletin*, 45(1):80–86.

Sudderth, E. B., Ihler, A. T., Freeman, W. T., and Willsky, A. S. (2003). Nonparametric belief propagation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1.

Summerfield, C. and Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron*, 59(2):336–347.

Sun, C., Fei, Z., Cao, C., Wang, X., and Jia, D. (2019). Low complexity polar decoder for 5g embb control channel. *IEEE Access*, 7:50710–50717.

Sundareswara, R. and Schrater, P. R. (2008). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, 8(5):12–12.

Taga, N. and Mase, S. (2006). Error bounds between marginal probabilities and beliefs of loopy belief propagation algorithm. In *Proceedings of the 5th Mexican International Conference on Artificial Intelligence*, MICAI'06, page 186–196, Berlin, Heidelberg. Springer-Verlag.

Tatikonda, S. (2003). Convergence of the sum-product algorithm. In *Proceedings 2003 IEEE Information Theory Workshop (Cat. No.03EX674)*, pages 222–225.

Tatikonda, S. C. and Jordan, M. I. (2002). Loopy Belief Propagation and Gibbs Measures. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 493–500, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Tetzlaff, T., Helias, M., Einevoll, G. T., and Diesmann, M. (2012). Decorrelation of Neural-Network Activity by Inhibitory Feedback. *PLoS Computational Biology*, 8(8):e1002596.

Thalmeier, D., Uhlmann, M., Kappen, H. J., and Memmesheimer, R.-M. (2016). Learning Universal Computations with Spikes. *PLoS Computational Biology*, 12(6):e1004895.

Trakoshis, S., Martínez-Cañada, P., Rocchi, F., Canella, C., You, W., Chakrabarti, B., Ruigrok, A. N., Bullmore, E. T., Suckling, J., Markicevic, M., Zerbi, V., Baron-Cohen, S., Gozzi, A., Lai, M. C., Panzeri, S., and Lombardo, M. V. (2020). Intrinsic excitation-inhibition imbalance affects medial prefrontal cortex differently in autistic men versus women. *eLife*, 9:1–31.

Treiman, D. M. (2001). GABAergic mechanisms in epilepsy. *Epilepsia*, 42 Suppl 3(SUPPL. 3):8–12.

Uddin, L. Q. (2020). Bring the Noise: Reconceptualizing Spontaneous Neural Activity. *Trends in Cognitive Sciences*, 24(9):734–746.

Valton, V., Romaniuk, L., Douglas Steele, J., Lawrie, S. M., and Seriès, P. (2017). Comprehensive review: Computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews*, 83:631–646.

Van Den Heuvel, M. P., Sporns, O., Collin, G., Scheewe, T., Mandl, R. C., Cahn, W., Goni, J., Pol, H. E., and Kahn, R. S. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatry*, 70(8):783–792.

van Loon, A. M., Knapen, T., Scholte, H. S., St. John-Saaltink, E., Donner, T. H., and Lamme, V. A. (2013). GABA Shapes the Dynamics of Bistable Perception. *Current Biology*, 23(9):823–827.

Vattikuti, S., Thangaraj, P., Xie, H. W., Gotts, S. J., Martin, A., and Chow, C. C. (2016). Canonical Cortical Circuit Model Explains Rivalry, Intermittent Rivalry, and Rivalry Memory. *PLOS Computational Biology*, 12(5):e1004903.

Vinckier, F., Gaillard, R., Palminteri, S., Rigoux, L., Salvador, A., Fornito, A., Adapa, R., Krebs, M. O., Pessiglione, M., and Fletcher, P. C. (2016). Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. *Molecular Psychiatry*, 21(7):946–955.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory Plasticity Balances Excitation and Inhibition in Sensory Pathways and Memory Networks. *Science*, 334(6062):1569–1573.

Voss, L. J., Hansson Baas, C., Hansson, L., Alistair Steyn-Ross, D., Steyn-Ross, M., and Sleigh, J. W. (2012). Investigation into the effect of the general anaesthetics etomidate and ketamine on long-range coupling of population activity in the mouse neocortical slice. *European Journal of Pharmacology*, 689(1-3):111–117.

Wacongne, C. (2016). A predictive coding account of MMN reduction in schizophrenia. *Biological Psychology*, 116:68–74.

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *The Journal of Neuroscience*, 32(11):3665–78.

Wainwright., M. J., Jaakkola, T., and Willsky, A. S. (2002). Tree-based reparameterization for approximate inference on loopy graphs. *Advances in Neural Information Processing Systems*.

Wainwright, M. J., Jaakkola, T., and Willsky, A. S. (2003). Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory*, 49(5):1120–1146.

Wainwright., M. J., Jaakkola, T., and Willsky, A. S. (2003). Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. *Workshop on Artificial Intelligence and Statistics*, 21:97.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Trans. Inf. Theor.*, 51(7):2313–2335.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Walker, P. (1975). Stochastic properties of binocular rivalry alternations. *Perception & Psychophysics 1975 18:6*, 18(6):467–473.

Wang, X.-J. (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron*, 36(5):955–968.

Wang, X.-J. (2006). Toward a Prefrontal Microcircuit Model for Cognitive Deficits in Schizophrenia. *Pharmacopsychiatry*, 39(S 1):80–87.

Wang, X.-J. and Krystal, J. H. (2014). Computational psychiatry. *Neuron*, 84(3):638–654.

Watanabe, Y. (2011). Uniqueness of Belief Propagation on Signed Graphs. In *Advances in neural information processing systems*.

Watanabe, Y. and Fukumizu, K. (2009). Graph zeta function in the bethe free energy and loopy belief propagation. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Waters, F., Blom, J. D., Dang-Vu, T. T., Cheyne, A. J., Alderson-Day, B., Woodruff, P., and Collerton, D. (2016). What Is the Link Between Hallucinations, Dreams, and Hypnagogic–Hypnopompic Experiences? *Schizophrenia Bulletin*, 42(5):1098–1109.

Wehr, M. and Zador, A. M. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature 2003 426:6965*, 426(6965):442–446.

Weilnhammer, V. A., Stuke, H., Hesselmann, G., Sterzer, P., and Schmack, K. (2017). A predictive coding account of bistable perception - a model-based fMRI study. *PLOS Computational Biology*, 13(5):e1005536.

Weilnhammer, V. A., Stuke, H., Sterzer, P., and Schmack, K. (2018). The Neural Correlates of Hierarchical Predictions for Perceptual Decisions. *The Journal of Neuroscience*, 38(21):5008–5021.

Weiss, Y. (1997). Interpreting Images by Propagating Bayesian Beliefs. In *Advances in neural information processing systems*, pages 908–914.

Weiss, Y. (2000). Correctness of Local Probability Propagation in Graphical Models with Loops. *Neural Computation*.

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in mrfs. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods: Theory and Practice*, pages 229–239. MIT Press.

Weiss, Y. and Freeman, W. T. (1999). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience 2002 5:6*, 5(6):598–604.

Weller, A., Tang, K., Sontag, D., and Jebara, T. (2014). Understanding the bethe approximation: When and how can it go wrong? In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 868–877, Arlington, Virginia, USA. AUAI Press.

Wiegerinck, W. and Heskes, T. (2002). Fractional Belief Propagation. In *Advances in Neural Information Processing Systems*, volume 15.

William Moreau, A., Amar, M., Le Roux, N., Morel, N., and Fossier, P. (2010). Serotoninergic Fine-Tuning of the Excitation–Inhibition Balance in Rat Visual Cortical Networks. *Cerebral Cortex*, 20(2):456–467.

Wilson, H. R. (2003). Computational evidence for a rivalry hierarchy in vision. *Proceedings of the National Academy of Sciences*, 100(24):14499–14503.

Wilson, H. R. (2007). Minimal physiological conditions for binocular rivalry and rivalry memory. *Vision Research*, 47(21):2741–2750.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, 12(1):1.

Wilson, R. and Finkel, L. (2009). A Neural Implementation of the Kalman Filter. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, 6:661–694.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882.

Xiang, Q., Xu, J., Wang, Y., Chen, T., Wang, J., Zhuo, K., Guo, X., Zeljic, K., Li, W., Sun, Y., Wang, Z., Li, Y., and Liu, D. (2019). Modular Functional-Metabolic Coupling Alterations of Frontoparietal Network in Schizophrenia Patients. *Frontiers in Neuroscience*, 0(FEB):40.

Xue, M., Atallah, B. V., and Scanziani, M. (2014). Equalizing excitation–inhibition ratios across visual cortical neurons. *Nature*, 511(7511):596–600.

Yang, G. J., Murray, J. D., Wang, X.-J., Glahn, D. C., Pearlson, G. D., Repovs, G., Krystal, J. H., and Anticevic, A. (2016). Functional hierarchy underlies preferential connectivity disturbances in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, 113(2):E219–E228.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Generalized Belief Propagation. *Advances in neural information processing systems*.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2003). *Understanding Belief Propagation and Its Generalizations*, page 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312.

Yeh, F. C., Panesar, S., Fernandes, D., Meola, A., Yoshino, M., Fernandez-Miranda, J. C., Vettel, J. M., and Verstynen, T. (2018). Population-averaged atlas of the macroscale human structural connectome and its network topology. *NeuroImage*, 178:57–68.

Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O'Shea, D. J., Sohal, V. S., Goshen, I., Finkelstein, J., Paz, J. T., Stehfest, K., Fudim, R., Ramakrishnan, C., Huguenard, J. R., Hegemann, P., and Deisseroth, K. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, 477(7363):171–178.

Yoon, J., Maddock, R., Rokem, A., Silver, M., Minzenberg, M., Ragland, J., and Carter, C. (2010). GABA concentration is reduced in visual cortex in schizophrenia and correlates with orientation-specific surround suppression. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(10):3777–3781.

Yoon, K., Liao, R., Xiong, Y., Zhang, L., Fetaya, E., Urtasun, R., Zemel, R., and Pitkow, X. (2018). Inference in probabilistic graphical models by graph neural networks. In *ICLR Workshop*.

Yu, Z., Chen, F., and Dong, J. (2017). Neural network implementation of inference on binary Markov random fields with probability coding. *Applied Mathematics and Computation*, 301:193–200.

Zalesky, A., Fornito, A., Seal, M. L., Cocchi, L., Westin, C.-F., Bullmore, E. T., Egan, G. F., and Pantelis, C. (2011). Disrupted Axonal Fiber Connectivity in Schizophrenia. *Biological Psychiatry*, 69(1):80–89.

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic Interpretation of Population Codes. *Neural Computation*, 10(2):403–430.

Zheng, Y., Jia, S., Yu, Z., Huang, T., Liu, J. K., and Tian, Y. (2020). Probabilistic inference of binary Markov random fields in spiking neural networks through mean-field approximation. *Neural Networks*, 126:42–51.

Zhou, Y., Gao, J., White, K., Merk, I., and Yao, K. (2004). Perceptual dominance time distributions in multistable visual perception. *Biological cybernetics*, 90(4):256–263.

TITRE

La propagation circulaire de croyances comme modèle d'inférences optimales et sous-optimales dans le cerveau : extension de l'algorithme et proposition d'implémentation neurale

RÉSUMÉ

Le modèle d'inférence circulaire est un modèle bayésien de troubles psychiatriques, initialement conçu pour rendre compte des manifestations cliniques de la schizophrénie et de la psychose. L'inférence circulaire repose sur l'algorithme de propagation circulaire de croyances, un algorithme d'inférence probabiliste approximative qui propose un paramètre additionnel comparé à l'algorithme de propagation de croyances ou Belief Propagation. Ce paramètre est appelé le facteur de correction des boucles. Il fixe la quantité de circularité dans l'inférence et est considéré comme représentant le niveau d'équilibre (local) entre les processus d'excitation et d'inhibition dans le réseau cérébral supposé réaliser les opérations d'inférence probabiliste. Dans ce cadre, les raisonnements circulaires et les symptômes psychotiques émaneraient d'une diminution du facteur de correction de boucles, c'est-à-dire d'un faible niveau d'inhibition comparé au niveau d'excitation.
Le travail présenté dans cette thèse permet d'appuyer le modèle d'inférence circulaire comme modèle d'inférences pathologiques (par exemple les hallucinations et les idées délirantes), d'inférences presque-optimales, et entre les deux d'inférences sous-optimales non cliniques, allant des biais usuels d'inférence (comme l'illustrent les phénomènes de perception bistable et de prise de décision hâtive, et la confiance excessive généralisée) aux comportements infracliniques comme le fait de croire en des théories du complot malgré des éléments contredisant ces théories.
De plus, cette thèse développe le modèle d'inférence circulaire de façons diverses. Premièrement, conceptuellement, en procurant à l'algorithme de propagation circulaire de croyances une fondation théorique, ce qui est réalisé en le reliant à des algorithmes existants comme la propagation fractionnaire de croyances. Deuxièmement, de façon plus pratique, en proposant des implémentations neurales (réseaux de neurones à rate ou à spikes, pour des variables binaires ou gaussiennes) et des mécanismes d'apprentissage biologiquement plausibles décrivant tous les deux comment les inférences probabilistes pourraient être réalisées dans le cerveau en utilisant cet algorithme. Enfin, le modèle est développé sur le plan théorique, en examinant les propriétés de convergence de l'algorithme de propagation circulaire, en formulant l'algorithme pour des distributions de probabilité plus complexes que précédemment, et en proposant une généralisation avec l'algorithme de propagation circulaire étendu.

MOTS CLÉS

psychiatrie computationnelle, neurosciences théoriques, inférence probabiliste, propagation des convictions, inférence circulaire, propagation circulaire des convictions, déséquilibre excitation-inhibition, schizophrénie, psychose

TITLE

Circular Belief Propagation as a model for optimal and suboptimal inference in the brain: extending the algorithm and proposing a neural implementation

ABSTRACT

Circular Inference is a Bayesian model of psychiatric disorders, previously designed to account for clinical manifestations of schizophrenia and psychosis. Circular Inference relies on the Circular Belief Propagation algorithm, an approximate probabilistic inference algorithm that proposes an additional parameter compared to Belief Propagation, called the *loop correction factor*. This loop correction factor sets the amount of circularity in the inference and is seen as a proxy to the (local) level of excitation-inhibition balance in the brain network assumed to perform probabilistic inferences. According to this framework, circular reasoning and psychotic symptoms arise for lowered loop correction factor, which would mean, for low levels of inhibition compared to excitation.
The work presented in this thesis provides further evidence for Circular Inference as a model of pathological inferences (e.g., hallucinations and delusions), near-optimal inferences, and in between non-clinical suboptimal inferences, ranging from usual inference biases (exemplified by the bistable perception and the jumping to conclusions phenomena, and the general overconfidence) to sub-clinical behavior like believing in conspiracy theories despite contradicting evidence.
Additionally, this thesis develops the Circular Inference model in different ways. First, conceptually, by providing the Circular BP algorithm with a theoretical foundation, which is done by relating it to existing algorithms such as Fractional BP. Second, more practically, by proposing neural implementations (rate networks and spiking networks, for binary or Gaussian variables) and biologically-plausible learning mechanisms overall describing how probabilistic inferences could be carried out in the brain using this algorithm. Finally, the model is expanded theoretically, by investigating the convergence properties of the algorithm, by writing Circular BP for more complex probability distributions than previously, and by generalizing the initial Circular BP into extended Circular BP.

KEYWORDS

computational psychiatry, theoretical neuroscience, probabilistic inference, belief propagation, circular inference, circular belief propagation, excitation-inhibition imbalance, schizophrenia, psychosis