Codage prédictif dans les transformations visuo-motrices

Laurent PERRINET

Habilitation à Diriger des Recherches

Équipe Inference in Visual Behaviour (InViBe)
Institut de Neurosciences de la Timone
UMR 7289, CNRS / Aix-Marseille Université
27, Bd. Jean Moulin, 13385 Marseille Cedex 5, France
http://invibe.net/LaurentPerrinet
Laurent.Perrinet@univ-amu.fr

Jury composé de Messieurs :
Prof. Laurent Madelain
Dr. Alain Destexhe
Prof. Gustavo Deco
Dr. Guillaume Masson
Dr. Viktor Jirsa
Prof. Jean-Louis Mege
Université Lille III
Université Paris XI (Rapporteur)
Universitat Pompeu Fabra, Barcelona (Rapporteur)
Aix-Marseille Université
Aix-Marseille Université (Rapporteur)
Aix-Marseille Université

Document de thèse - Avril 2014
Abstract

Ce manuscrit présente une synthèse de mon projet scientifique afin de préparer le diplôme d’Habilitation à Diriger des Recherches (HDR). Il est l’occasion de faire le point sur un travail scientifique depuis ma soutenance de thèse en 2003.

La première partie résume ce parcours : le Chapitre 1 présente mon CV détaillé, tandis que le Chapitre 2 comprend une synthèse du travail accompli, puis développe mes thèmes actuels de recherche.

La seconde partie du manuscrit (Chapitres 3 à 6) présente de manière plus précise mon activité de recherche telle qu’elle est résumée au Chapitre 2. Cette partie est rédigée en anglais. Nous conclurons dans la dernière partie par une description des perspectives de cette démarche scientifique.
II Activité de recherche détaillée

3 Motion-based prediction explains the role of tracking in motion extrapolation

3.1 Introduction ................................................. 48
    3.1.1 Problem statement ..................................... 48
    3.1.2 Different types of motion extrapolation .......... 51
    3.1.3 Experimental evidence of motion extrapolation .... 52
    3.1.4 Existing theories on motion extrapolation ......... 53
    3.1.5 Motion extrapolation and motion-based prediction .. 54
    3.1.6 Objectives and outline ............................... 56
3.2 Model & methods ............................................ 56
    3.2.1 Probabilistic detection of motion ................... 56
    3.2.2 Luminance-based detection of motion ............... 57
    3.2.3 Motion-based predictive coding ..................... 58
    3.2.4 Numerical simulations ............................... 60
    3.2.5 Experimental procedure .............................. 61
3.3 Results .................................................... 62
    3.3.1 Extrapolation of motion information in a blanked trajectory .... 63
    3.3.2 Predictive tracking in the presence of varying noise levels 65
    3.3.3 Motion extrapolation in noisy blanked trajectory .. 67
3.4 Discussion ................................................ 69

4 Statistics of edge co-occurrences are sufficient to categorize natural images

4.1 Introduction ................................................. 74
4.2 Results .................................................... 76
    4.2.1 Statistics of edges .................................. 76
    4.2.2 Edge co-occurrences are sufficient to categorize natural images ................. 79
    4.2.3 Robustness to noise, translation and rotations ...... 81
4.3 Discussion ................................................ 81
4.4 Methods .................................................. 84
    4.4.1 Linear representation of edges ..................... 84
    4.4.2 Sparse coding and validation of the edge extraction method ..................... 85
    4.4.3 Histogram of edge co-occurrences and geometrical symmetries ................ 88
5 Anisotropic connectivity implements motion-based prediction in a spiking neural network

5.1 Introduction

5.1.1 Problem statement: Codage prédicatif dans les transformations visuo-motrices

5.1.2 Neuro-physiological correlates of prediction for motion extrapolation

5.1.3 Existing neuromorphic models for prediction

5.1.4 Our approach: an anisotropic connectivity pattern implementing motion-based prediction

5.1.5 Objectives and outline

5.2 Methods

5.2.1 Neuron parameters

5.2.2 Tuning properties

5.2.3 Input Stimulus

5.2.4 Network connectivity

5.2.5 Choice of parameters

5.2.6 Prediction readout

5.3 Results

5.3.1 Anisotropic diffusion transports information during the blank

5.3.2 Reading-out the population response

5.3.3 Motion-based predictive anisotropic diffusion solves motion extrapolation

5.3.4 Conclusion

5.4 Discussion

5.4.1 Summary and comments

5.4.2 Context of existing models

5.4.3 Outlook and future work

6 Active inference, eye movements and oculomotor delays

6.1 Introduction

6.1.1 Existing solutions and the proposed hypothesis

6.1.2 Outline

6.2 From predictive coding to active inference

6.2.1 From free energy to generalized filtering

6.2.2 Hierarchical form of the generative model

6.2.3 Perception and predictive coding

6.2.4 Action

6.2.5 Summary

6.3 Active inference with sensorimotor delays

6.3.1 Summary

6.4 Results: pursuit initiation

6.4.1 Generative model of pursuit initiation
Part I

Présentation générale
Chapter 1

Curriculum Vitæ détaillé

1.1 Présentation rapide

41 ans, né le : 23-02-1973 à Talence (Gironde).

- Affiliation: Chargé de Recherche (CR1, CNRS), Équipe Inference in Visual Behaviour (InViBe) - Institut de Neurosciences de la Timone (UMR 7289, CNRS / Aix-Marseille Université)
- Adresse: 27, Bd. Jean Moulin, 13385 Marseille Cedex 5, France
- E-mail: Laurent.Perrinet@univ-amu.fr
- Téléphone: 04 91 32 40 44
- URL: http://invibe.net/LaurentPerrinet

Intérets de Recherche

Mon objectif de recherche est d’étudier l’hypothèse selon laquelle on peut comprendre les liens entre la structure neurale et la fonction des systèmes sensoriels comme l’optimisation de leur adaptation aux statistiques des scènes naturelles.

Cet objectif vise à étendre la compréhension des facultés sensorielles et cognitives sous la forme de modèles de réseaux de neurones impulsionnels qui réalisent de façon efficace des algorithmes de la perception visuelle. En effet, les brèves impulsions du potentiel de membrane se propageant au fil des neurones sont une caractéristique universelle des systèmes nerveux et permettent de construire des modèles événementiels efficaces de traitement dynamique de l’information. Dans un but fonctionnel, je désire notamment implanter dans ces modèles des stratégies d’inference grâce à des mécanismes d’apprentissage auto-organisés fixant les relations spatio-temporelles entre les neurones. Dans le cadre du projet BrainScaleS, nous envisageons la création de nouveaux types d’algorithmes basés sur ces recherches.
Mots clés

1.2 Diplômes & titres universitaires

Diplôme d’ingénieur SUPAÉRO, Toulouse, France. 1993-1998
Spécialisation dans le traitement du signal et de l’image et en particulier dans les techniques des réseaux de neurones artificiels.

- Projets individuels sur la perception visuelle, la reconnaissance de locuteur, la compression de la parole et sur la reconnaissance de caractères.


- Assistant de recherche, JET PROPULSION LABORATORY (NASA), Pasadena, Californie. Département des Sciences de la Terre, Laboratoire d’imagerie radar, INTERFÉROMÉTRIE RADAR SAR APPLIQUÉE À LA GÉOPHYSIQUE
  - Programmation d’un processus de traitement d’images radar interférométriques SAR comprenant des corrections géographiques, une série de filtrages et un traitement d’interférométrie.
  - Étude et programmation d’un générateur de carte topographique.
  - Traitement des images obtenues pour surveiller la déformation de la croûte terrestre. Étude des tremblements de terre de Landers (Californie) et de Gulan (Chine).

- Assistant de recherche, CerCo (CNRS, UMR5549), Toulouse (stage de fin d’études d’ingénieur). Développement d’un réseau de neurones asynchrone appliqué à la reconnaissance de caractères.
  - Programmation du code du réseau de neurones asynchrones.
  - Étude et utilisation des statistiques non-paramétriques pour l’utilisation d’un code basé sur le rang d’activation des neurones.
  - Implantation d’une nouvelle règle d’apprentissage du réseau de neurones.
DEA de Sciences cognitives
1998-1999
Univ. Paris VII, P. Sabatier, EHESS, Polytechnique, mention TB. Allocataire d’une bourse de DEA.

- Assistant de recherche, ONERA/DTIM (Département de Traitement de l’Image et de Modélisation), Toulouse (stage de DEA). 3/1999-7/1999
  - Étude de l’apprentissage de type Hebbien de réseaux de neurones basés sur un codage par rang.
  - Application à la reconnaissance de textures visuelles.

- Assistant de recherche, USAFB (Rome, NY) / University of San Diego in California (États-Unis). Étude de l’apprentissage autonome dans un système complexe de type automate cellulaire.

Doctorat de Sciences cognitives ONERA/DTIM, Toulouse 1999-2003

Titre : Comment déchiffrer le code impulsionnel de la Vision? Étude du flux parallèle, asynchrone et épars dans le traitement visuel ultra-rapide.
Allocataire d’une bourse MENRT, accueil à l’ONERA/DTIM.

- Cette thèse a été initiée par les résultats de la collaboration pendant le stage de DEA. Elle a été dirigée par Manuel Samuelides (professeur à Supaëro et chargé de recherche à l’ONERA/DTIM) et co-dirigée par Simon Thorpe (directeur de recherche au CerCo)

- Participation et présentations à de nombreux colloques internationaux (IJCNN99, NeuroColt00, CNS00, CNS01, LFTNC01, ESANN02, NSI02). Participation aux écoles d’été “EU Advanced Course in Computational Neuroscience” à Trieste (Italie) et “Telluride Neuromorphic Workshop” au Colorado (États-Unis).

- En parallèle, j’ai participé à l’organisation d’une conférence sur les réseaux de neurones dynamiques (DyNN*2000). Je me suis aussi impliqué dans des activités d’enseignement (initiation à la programmation sous Matlab et théorie de la probabilité) pour des élèves de première et deuxième année d’école d’ingénieur (à Supaëro et à l’ENSICA, Toulouse) et des travaux dirigés de robotique (Traitement de l’image et reconnaissance d’objets au laboratoire d’Informatique et d’Automatique de Supaëro).

- La thèse de doctorat a été soutenue le 7 février 2003 à l’université Paul Sabatier avec la mention ”Très honorable avec les félicitations du jury”. Le jury était présidé par Michel Imbert (Prof. Université P. Sabatier) et constitué par Yves Burnod (Directeur de recherche à l’INSERM U483) et Jeanny Hérault (Professeur à l’INPG, Grenoble).
**Post-doctorat**, Redwood Neuroscience Institute (RNI), États-Unis 2004
Sous la conduite de Bruno Olshausen, j’ai comparé des modèles standards d’apprentissage avec une méthode originale centrée sur les potentiels d’action. Notamment, j’ai développé une méthode générique évaluant l’importance des processus homéostatiques dans l’apprentissage non-supervisé, qui a conduit à une publication dans le journal Neural Computation (référence A20-[Per10b]). J’ai ensuite étendu ce modèle à l’apprentissage spatio-temporels dans des flux vidéo.

1.3 Situation actuelle

**Chargé de Recherche grade 1**, INT/CNRS, Marseille 2004-2014
Sous la conduite de Guillaume Masson à l’Institut de Neurosciences de la Timone (UMR 7289, CNRS / Aix-Marseille Université ) à Marseille, j’étudie des modèles spatio-temporels d’inférence dans des scènes naturelles en application de la compréhension des mouvements oculaires. Au 1er janvier 2012, notre équipe a intégré l’Institut des Neurosciences de la Timone, Marseille (direction Guillaume Masson).

**Mission longue** Karl Friston’s theoretical neurobiology group (The Wellcome Trust Centre for Neuroimaging, University College London, London, UK)). Collaboration avec Karl Friston sur l’application de modèles d’énergie libre aux mouvements oculaires.

1.3.1 Contrats et collaborations

- Le renforcement de la modélisation au sein de l’équipe passe par son insertion dans un réseau efficace de collaborations: J’ai contribué développer un réseau dense de collaborations effectives, que ce soit au niveau local (Laboratoire de Mathématique, LATP; LMP, Equipe Biorobotique de N. Franceschini), régional (équipe Odysée, INRIA Sophia Antipolis), national (UNIC, Gif-sur-Yvette; CerCo, Toulouse; LPP, Paris; LoRRA, Nancy) ou encore international (Bernstein Center for Computational Neuroscience, Münich, Freiburg; Instituto de Optica, Madrid; Heiko Neumann - Universität Ulm).

- Régionalement, nous avons développé des interfaces avec les Mathématiques Appliquées pour l’étude des systèmes dynamiques complexes qui sont empruntés des probabilités, du calcul stochastique et de la physique statistique. Nous avons développé cet axe par une collaboration régionale (avec Bruno Cessac, Université de Nice, Emanuel Daucé, Ecole Central de Marseille et Manuel Samuelides, SUPAERO) grâce à une ACI ("Temps et Cerveau") qui a conduit
à la publication dans la revue "European Physics Journal" à des articles longs de revue sur le sujet [Per07]. Nous avons bénéficié d’un financement ANR (MAPS) pour pérenniser cette collaboration avec de nouveaux partenaires français (F. Alexandre au LoRIA, L. Goffart à l’INCM).

- Au niveau international nous avons bénéficié d’un contrat important, le projet intégré européen FACETS (EC IP project FP6-015879, 600Kuros) entre 2005 et 2010. Il a permis de financer la bourse d’un doctorant (Jens Kremkow, en co-direction avec Ad Aertsen, Freiburg) pour 3 ans et d’un post-doctorant (Nicole Voges) pour 18 mois. Une extension de ce projet sous la forme d’un réseau d’apprentissage intégré a permis de financer une bourse de thèse (Mina Khoei, FACETS-ITN, EU funding grant number 237955). EN 2010, nous avons bénéficié d’un nouveau contrat, sous la forme d’un projet intégré, BrainScaleS (EU funding, grant number FP7-269921, 800Kuros). Ce contrat permet de structurer nos collaborations (par exemple en participant et en organisant des ateliers thématiques) tout en renforçant certain liens existant (notamment UNIC, Freiburg, Technische Universität Graz). Il permet ainsi de donner une portée internationale au travaux inscrits dans ce projet.

1.3.2 Enseignement & formation

- Encadrement projet Master BIM à Luminy,
- Cours magistraux de Neurosciences Computationnelles (3ème cycle, Luminy),
- Encadrements passés : 2 stagiaires Mastere (2008 et 2010), 1 PhD (Jens Kremkow, soutenue en 2010), 1 PhD invité (Sylvain Fischer, soutenue en 2008), 2 post docs (Nicole Voges et Julian Budd dans le cadre du contrat FACETS),
- Encadrements en cours: 1 PhD (Mina Khoei, en cours, financement FACETS-ITN), 1 Post-Doc (Wahiba Taouali, financement BrainScaleS).

1.3.3 Animation scientifique

• Responsable de deux « work-packages » dans le projet européen Brain-ScaleS ; organisation d’un atelier (« demo 1 Stockholm meeting » en 2012), animation des collaborations (dissémination des codes produits dans la structure "NeuralEnsemble" associée au projet) et participation à l’atelier "Using the ESS + Neuromorphic hardware Workshop" (TU Dresden, Germany, 2011) pour porter nos algorithmes sur les circuits intégrés. Accueil d’un étudiant en thèse (Bernhard Kaplan, KTH) pour renforcer notre collaboration avec Anders Lansner et proposer un modèle.

• Membre du réseau "NeuroComp":


  – Organisation d’un workshop annuel en neurosciences computationnelles à Marseille.


1.3.4 Promotion et diffusion de la culture scientifique

• Participation à des activités grand public: conférence invitée “Peut-on parler d’intelligence mécanique?” organisée par l’Association Science Technologie Société.

• Publication destinée au public scolaire “Qui créera le premier calculateur intelligent?”, Doc-Sciences (voir http://www.docsciences.fr/).


1.4 Liste des publications

1.4.1 Articles de revues internationales à comité de lecture

A32 Mina A. Khoei, Guillaume S. Masson, Laurent U. Perrinet. Motion-2013


A28 Laurent U. Perrinet and Guillaume S. Masson. Motion-based prediction is sufficient to solve the aperture problem. *Neural Computation*, pages 1–25, June 2012. ISSN 1530-888X.


1.4.2 Chapitres d’ouvrage à comité de lecture


1.4.3 Thèse


1.4.4 Actes de conférences internationales à comité de lecture


34. Claudio Simoncini et al. “Different pooling of motion information for perceptual speed discrimination and behavioral speed estimation”. In: Vision Science Society. 43.503. 2010


31. Laurent Perrinet. “Probabilistic models of the low-level visual system: the role of prediction in detecting motion”. In: LADISLAV TAUC & GDR MSPC NEUROSCIENCES CONFERENCE, From Mathematical Image Analysis to Neurogeometry of the Brain. 2010


16


Chapter 2

Résumé de mon activité de recherche

2.1 Introduction: codage prédictif et code neural

Le but de mon activité de recherche est de déchiffrer le “code neural”, c'est-à-dire de révéler dans la structure dynamique de l'activité nerveuse des algorithmes fonctionnels de traitement de l'information. Plus particulièrement je m'intéresse à comprendre comment le système visuel peut exploiter les régularités statistiques des scènes naturelles pour traiter le flux sensoriel de façon la plus efficace possible. Ce thème de recherche s'intègre donc plus généralement au problème de notre compréhension entre la structure du système nerveux central et de sa fonction.

À ce titre, l'étude de l'intégration spatio-temporelle de l'information sensorielle est primordiale. En effet, les neurones présentent des contraintes physiologiques qui font que l'information sensorielle est locale aux premiers étages de captation du signal, alors qu'elle doit devenir globale et unique au niveau de la réponse comportementale. De plus, au fur et à mesure qu'ils montent les voies sensorielles jusqu'à la réponse motrice, ces signaux subissent de nombreuses transformations et subissent différentes sources de bruit. Ces problèmes se révèlent de façon saillante dans le système oculomoteur: Devant une réponse visuelle, comme l'image d'un prédateur pour une proie (et inversement), il est primordial à la survie du “sujet voyant” de pouvoir orienter son regard de façon efficace vers l'objet d'intérêt et de programmer une réponse adaptée le plus rapidement possible.

Dans ce cadre, le système oculomoteur procure un excellent modèle pour mettre à jour des processus de codage prédictif dans le code neural. En particulier, quel codage neural est le plus efficace pour l'intégration de l'information? Comment intégrer différentes sources d'information (montante, associative, descendante)? Quel est le meilleur compromis entre rapidité de la réponse et sa précision?
Afin de répondre à ces questions fondamentales pour les neurosciences, nous allons les aborder suivant les niveaux d'études suggérés par Marr [Mar83]. Ceux-ci structurent mes axes de recherche suivant différentes approches:

**Fonctionnelle** Quelles fonctions sont à la source de nos capacités visuelles et en particulier nous aident à comprendre la perception visuelle du mouvement?

**Algorithmique** Comment exploiter le parallélisme et la dynamique des réseaux neuronaux de façon efficace par rapport à la représentation de l'information visuelle? Existe-t-il des algorithmes d'apprentissage?

**Computationnelle** Comment implanter ces algorithmes dans la circuiterie neurale? Quels enseignements ces modèles nous donnent pour déchiffrer le code neural de la perception visuelle?

Cette dichotomie est bien sûr *a priori* arbitraire et constitue plutôt une grille de lecture pour aborder ces problèmes complexes.

Dans cette partie, nous résumerons donc mon activité de recherche en suivant cette grille. Nous débuterons par mes travaux de thèse dans la section 2.2 et en particulier par l’étude du lien entre une propriété fonctionnelle (le codage ultra-rapide d’images naturelles) et son corrélat neural (un code neural utilisant la latence de décharge d’un neurone) tout en détaillant des algorithmes faisant le lien entre ces deux niveaux. Ensuite nous détaillerons dans la section 2.3 le travail accompli au sein de l’équipe DyVA, dirigée par Guillaume Masson à l’INCM et en particulier les modèles que nous avons développés autour du système oculomoteur comme modèle dynamique d’intégration sensorielle. Je détaillerai enfin dans la section 2.4 mes travaux actuels de recherche au sein de l’équipe INVibe dirigée par Frédéric Chavane à l’INT en les focalisant sur les liens entre codage neural et dynamique de la réponse oculaire. Cette section résumera ainsi la partie suivante qui détaille mes axes actuels de recherche.

### 2.2 Rappels sur mon travail de thèse

Le but de mon travail de thèse était d’étendre la compréhension de modèles des facettes cognitives sous la forme de réseaux de neurones réalisant des algorithmes de la perception visuelle. En effet, j’avais développé sous la direction de Manuel Samuelidès, professeur de mathématiques à Supaéro, des algorithmes novateurs de traitement de l’image basé sur des réseaux de neurones. Ces algorithmes sont basés sur des réseaux de neurones impulsionnels. En effet, les brèves impulsions du potentiel de membrane se propageant au fil des neurones sont une caractéristique universelle des systèmes nerveux et permettent de construire des modèles événementiels efficaces de traitement dynamique de l’information. Grâce à la rencontre de ce dernier avec Simon
Thorpe, chercheur au CerCo, j’ai pu développer la démarche inverse, c’est-à-dire de comprendre le fonctionnement du code neural à partir de contraintes fonctionnelles. En effet, Simon Thorpe a démontré qu’il était possible de catégoriser des images, par exemple contenant ou non un animal, de façon très rapide avec une latence d’environ 150 ms [TFM96]. Des algorithmes neuromimétiques réalisant une telle prouesse, en supposant qu’ils sont basés sur un certain nombre de couche de traitements, doivent nécessairement effectuer de tels traitement avec un nombre minimal d’impulsions. Partant de ce fait, l’objectif de ma thèse était d’étudier des codes neuronaux plausibles qui ne nécessitent pour chaque image que d’une impulsion par neurone.

2.2.1 Codage par rang

Tout d’abord, j’ai étudié l’influence de la forme de l’information axonale émise par les neurones sur les propriétés fonctionnelles du code neural. En effet, par sa nature événementielle et parallèle, l’information contenue dans l’activité neuronale est radicalement différente de formes classiques de représentation de l’information. Dans le formalisme que nous avons choisi, seul la latence de décharge de la première impulsion par neurone importe. Au niveau de la population, ce sont les rapports entre les valeurs qui importent plutôt que leur valeur analogique: Ainsi, nous pouvons dans une large mesure catégoriser une image indépendamment du contraste. L’hypothèse que nous avons alors retenue prédit que la valeur analogique est codée par son rang plutôt que par une valeur analogique, telle la fréquence de décharge. Cette solution est à la fois économique (elle peut être implémentée de façon physiologique de façon simple [Del+01a]) et robuste. En particulier, j’ai alors montré le lien entre la transformation d’une valeur analogique en un rang et le processus d’égalisation de l’histogramme [Per99], qui est caractéristique du fonctionnement neural [Lau81]. Un tel rapprochement a été ensuite exploité pour élaborer un modèle de rétine [RT01] et étendu au décodage de valeurs de rang pour l’optimisation de la reconstruction [Per10b; PST04a].

En outre, j’ai montré pendant ma thèse le lien entre le codage par rang tel qu’il était proposé par Simon Thorpe et son équipe et des tests statistiques classiques. En effet, le résultat d’un test de corrélation de type Wilcoxon pouvait être rapproché de la dynamique d’un neurone utilisant un codage par rang. Grâce à un tel rapprochement et aux résultats du théorème central limite permutational, j’ai ainsi pu démontrer de façon analytique la dynamique de la distribution de la densité de probabilité de l’activité d’un neurone pour une entrée aléatoire [Per03]. Grâce à de tels résultats, nous avons pu prédire les seuils qui sont optimaux pour atteindre un certain compromis entre vitesse et précision, un ingrédient qui est particulièrement important pour la classification mais aussi par exemple en cours d’apprentissage de poids synaptiques.
2.2.2 Plasticité dépendant de la latence de décharge neurale

Connaissant ainsi de façon complète le comportement d’un neurone à codage par rang, j’ai pu implanter des algorithmes d’apprentissage pour ce type de réseaux. Nous avons alors exploité la mise en évidence récente d’un phénomène de potentiation ou de dépotentiation des synapses dépendant de l’ordre d’arrivée des potentiels d’action [BP98; Mar+97] (ou STDP). Celle-ci a été alors formalisée dans un cadre physiologique par des modèles génératifs [Per+01; PS00]. Nous avons ainsi montré que cette règle pouvait conduire à l’émergence de champs récepteurs réalistes de l’aire primaire visuelle [Del+01b].

Pour étendre la compréhension de tels mécanismes, nous avons étendu cette approche phénoménologique en essayant de comprendre pourquoi une telle règle d’apprentissage était efficace. En nous basant sur un coût basé sur la précision de la détection d’une vague synchrone de potentiels d’action, nous avons établi une règle d’apprentissage — similaire en nature mais modifiée par rapport à la règle phénoménologique — qui permettait de détecter des structures cohérentes dans les entrées pré-synaptiques [PS02a]. De tels travaux ont été récemment étendus à des modèles physiologiques encore plus réalistes [Mas12]. Ils correspondent à des principes d’optimalité qui ont été étendus à des problèmes d’apprentissage plus complexes [HPM13].

2.2.3 Codage épars

Pour étendre ce type d’architectures à des entrées plus réalistes, comme des images naturelles, j’ai ensuite développé l’architecture proposée par Rullen and Thorpe [RT01]. En effet, celle-ci était basée sur une approximation d’une base d’ondelettes pour lequel nous avons montré qu’elle peut être optimisée en découplant base de décomposition et base de synthèse [PST04a]. Une fois ce modèle de rétine optimisé, j’ai voulu l’étendre et modéliser l’aire visuelle primaire qui se caractérise par un plus grand nombre de filtres sélectifs à différentes orientations. Toutefois, nous avons alors observé qu’en augmentant le nombre de filtres, de telle sorte que la base devient sur-complète, le code neural devient redondant et perd de son efficacité. Afin de lever cette source d’inefficacité, j’ai implanté une méthode de propagation "en-avant" d’un signal de décorrélation utilisant des connections latérales. Nous avons alors mis en évidence [PS02b; PS02c; PST04b] le parallèle entre une telle approche et l’algorithme de Matching Pursuit [MZ93].

J’ai alors appliqué ce modèle à une architecture simplifiée de l’aire visuelle primaire. Les résultats ont montré qu’une telle représentation était efficace et qu’elle répliquait le caractère épars du code neural dans l’aire visuelle primaire (V1). En effet, par rapport à un modèle classique (tel que celui de [RT01]), l’activité neurale telle qu’elle est mesurée physiologiquement est plus épars, c’est-à-dire qu’on observe moins de potentiels d’action que la
prévision linéaire. Un tel principe peut s’expliquer en terme d’économie de moyen (on code le même signal avec moins de potentiels d’action) ou plus généralement en terme d’efficacité car on ne code que les parties les plus significatives du signal. Par ailleurs, cette règle a été utilisée pour la définition d’un coût de représentation qui permet d’expliquer la formation de champs récepteurs dans l’aire visuelle primaire [OF96]. Nous avons alors considéré un tel principe et montré des résultats similaires pour V1 [PST03] et qui ont été ensuite généralisés à des conditions expérimentales plus génériques [Per10b]. De plus, j’ai montré qu’une telle représentation conduisait à une grande régularité des coefficients analogiques en fonction de leur rang, une propriété essentielle pour leur utilisation dans des réseaux de neurones tels que nous les étudions [PST04a].

2.2.4 Synthèse

Pour résumer, ces travaux de thèse ont permis de réaliser l’objectif initial et de proposer des solutions novatrices pour comprendre des aspects essentiels du codage neural dans les aires visuelles de bas niveau. À partir de l’architecture événementielle et parallèle du code neural dans l’aire V1, nous avons alors mis en évidence l’importance du caractère épars du code neural, aussi bien pour optimiser l’efficacité de la représentation d’un image (le codage) que pour implanter des algorithmes efficaces d’apprentissage non-supervisé.

En résumé, ces travaux sur l’étude du flux parallèle, asynchrone et épars dans le traitement visuel ultra-rapide m’ont permis de développer des modèles tout en les confrontant à des applications au traitement de l’image comme la compression d’image ou la reconnaissance d’objets. Nous avons ainsi développé un formalisme original de représentation optimale par des réseaux de neurones impulsionnels de l’information visuelle pour des images statiques. Ceux-ci comprennent aussi bien des applications “bas-niveau” (compression d’image, denoising) que “haut-niveau” (détectection, segmentation). Ces résultats ont été synthétisés dans un article de revue publié dans le “European Physical Journal” [Per07].

Toutefois, ils comportent des limites. Tout d’abord, ces modèles étaient le plus souvent limités aux modèles à codage par rang implanté dans le laboratoires et manquaient de généralité par rapport à des modèles neuro-mimétiques. Ensuite, les entrées que nous considérions étaient le plus souvent constituées d’images statiques. Enfin, les activités neuronales sont sensées représenter des valeurs analogiques, mais de telles représentations ne peuvent pas explicitement coder pour des dimensions essentielles de l’information, comme l’incertitude d’une mesure. Mes travaux en post-doctorat et en tant que chercheur m’ont ensuite permis d’étendre de tels modèles à des entrées dynamiques.
2.3 Travail accompli

En effet, à mon arrivée dans l’équipe DyVA (Dynamique de la Vision et de l’Action) dirigée par Guillaume Masson à l’INCM, j’ai étendu les modèles développés durant ma thèse tout en les ouvrant à de nouveaux axes de recherche. En particulier, un objectif majeur a été :

1. de baser la représentation de l’information sur les solides fondations théoriques de la théorie de la probabilité,

2. d’élargir les modèles à des bases neurophysiologiques plus plausibles,

3. mais aussi de valider les modèles en lien direct avec les expériences comportementales et physiologiques qui étaient conduites dans le laboratoire.

Tout d’abord, je vais les placer dans leur contexte tant au niveau de leur intégration dans les travaux de l’équipe qu’aux niveaux des différents contrats que nous avons obtenus pour les réaliser. Ces travaux ont été développés entre les années 2004 et 2010 (date du déménagement du laboratoire dans un nouveau site), notamment dans le cadre du projet européen “FACETS” et exposés dans de nombreuses conférences internationales et revues (Neural Computation, Vision Research, ...). Je vais dans cette section détailler ces différents points.

2.3.1 Place du thème dans le laboratoire

L’équipe DyVA conduisait des programmes de recherche portant sur la dynamique des processus fondamentaux de la perception visuelle et du contrôle visuo-moteur. Les travaux sont menés chez le primate éveillé, l’adulte sain et des patients atteints de lésions rétiniennes. Ces questions sont abordées au niveau comportemental et théorique comme au niveau neurophysiologique (électrophysiologie et imagerie) grâce à la forte pluridisciplinarité permise par la taille de l’équipe (25 personnes fin 2010) et la diversité d’origine des chercheurs et enseignant-chercheurs qui la composent. C’était une jeune équipe (4 chercheurs en 2003), qui s’est considérablement renforcée en quelques années grâce à 4 recrutements CNRS et qui l’ouverture au Service d’Ophtalmologie du CHU Timone.

Ces techniques font appel à des technologies de plus en plus lourdes pour l’analyse et la modélisation de ces données. La modélisation tient à ce titre un rôle fédérateur dans l’équipe puisqu’elle permettait de rapprocher des disciplines traditionnellement relativement éloignées comme la psychophysique et la neurophysiologie. En centrant le thème de recherche autour de l’inférence statistique, on peut ainsi regrouper les résultats comportementaux mais aussi les réponses neurales comme des réponses de systèmes décisionnels et
comparer leur efficacité par rapport à des résultats théoriques (l’observateur idéal).

Toutefois l’intégration de ces techniques des neurosciences computationnelles nécessite un soin particulier pour sa bonne intégration. En effet, elle introduisent de nouveaux concepts et une manière différente d’interpréter les données. Ces contraintes nous ont amenés à réfléchir à la construction au sein de l’équipe d’une plate-forme pour les neurosciences computationnelles. Nous avons en particulier monté un projet qui nous a permis le financement d’un ferme de serveurs de calcul de 20 nœuds qui nous a permis de mettre à un niveau international notre capacité de calcul.


2.3.2 Modélisation des fonctions de la perception visuelle du mouvement

Décoder des populations par un Observateur Idéal (avec G. Masson et F. Barthélemy)

Lors de l’analyse d’une scène naturelle, et afin d’extraire une information utile (efficace et rapide) de mouvement, le système visuel est confronté à une multitude de problèmes de haute complexité computationnelle. En effet, cette information de mouvement est par nature multi-modal, bruitée et potentiellement ambiguë et son identification est un problème pour lequel il n’existe pas d’algorithme séquentiel simple (il est NP-complet). Une approche de modélisation de la boucle sensori-motrice utilise l’inference statistique pour la perception du mouvement [WSA02]. Nous avons rapproché par cette méthode les données aussi bien neurophysiologiques que comportementales pour proposer un décodage des stratégies computationnelles utilisées par le système visuel. Il peut alors être considéré dans son ensemble comme un réseau adaptatif et auto-régulé d’agents répondant de façon optimale [Per+05]. Une question majeure reste à savoir: 1) quelles sont les règles régissant l’intégration des information multi-modales et 2) comment des décisions sont produites à partir de ces informations distribuées. Cette dynamique se manifeste dans nos capacités cognitives comme par exemple la divergence dynamique des flux d’information: comment séparer dans le temps les composantes rapides et grossières des composantes plus lentes mais plus fines.

Ce projet implique une interaction forte avec l’expérimentation pour à la fois dessiner les expériences pertinentes grâce à l’identification des composantes
Figure 2.1: Modélisation de l’effet de la taille du stimulus sur le mouvement de l’œil pendant le réflexe de poursuite oculaire. Grâce à la modélisation du champ récepteur comportemental, on peut prévoir que les changement de comportement de la réponse de l’œil en fonction du diamètre du stimulus (à gauche) pour différentes fréquences spatiales (à droite). Le gain d’accélération sature rapidement pour toutes les courbes alors que pour une fréquence spatiale plus haute, nous prévoyons une super-saturation de ces réponses.

indépendantes portant l’information de mouvement mais aussi pour l’analyse des réponses oculaires [Per+05; Per+06; Per+09b; PM08b].

Le champ récepteur comportemental (avec G. Masson et F. Barthélemy)

Nous avons poursuivi ainsi nos efforts de modélisation pour comprendre comment un décodage probabiliste de l’activité des différentes sous-populations neurales peut rendre compte de ce “champ récepteur comportemental”. Du point de vue théorique, ceci revient à tenter de reconstruire les propriétés de chacune des sous-populations à partir des distributions de probabilités obtenues comportementalement et de baser une fonction de décision sur un observateur idéal pouvant extraire vitesse et direction à partir d’une représentation distribuée probabiliste. Ce type de modèle bayesien contraint l’espace de paramètres à explorer aussi bien comportementalement que pour les modèles: Il permet en particulier de valider l’efficacité du système visuel par rapport à un ”observateur idéal”. Nous avons poursuivi nos travaux comportementaux chez l’homme et chez le singe visant à décrire les propriétés spatio-temporelles et leurs dynamiques du champ récepteur comportemental sous-jacent à l’initiation des réponses oculaires réflexes. Il est à noter que cette définition du champ récepteur est en quelque sorte un retour aux sources puisque c’est Sherrington [She06] qui la formula en premier sous la forme suivante : “the whole set of points of skin surface from which the scratch-reflex
can be elicited”. Nous avons exploré les interactions entre populations de neurones telles qu’il est possible de les mesurer en titrant la réponse à un stimulus en fonction de son contexte, spatialement recouvrant ou non. La structure spatio-temporelle du champ peut ainsi être cartographiée au moyen de la technique de corrélation inverse (ou classification d’image en psychophysique) adaptée pour les réponses oculaires réflexes. Ces travaux donnent ainsi une image complète de ce “champ sensorimoteur” au moyen d’un ensemble d’opérateurs définis à partir d’un modèle inférentiel de décodage des populations neurales sous-jacentes [PBM06; Per+05; Per+09a; PM07; PM08a]. Ce travail a notamment abouti à une publication de revue dans la revue *Neuroscience & Biobehavioral Reviews* [MP12].

Modélisation inférentielle dynamique (avec A. Montagnini, G. Masson, F. Barthélemy, P. Mamassian / Laboratoire de Psychophysique de la Perception, CNRS, Université Paris V)

Ces travaux se rapprochent d’une modélisation bayésienne de l’intégration des signaux de mouvements : la décision perceptive est élaborée à partir de représentations distribuées, probabilistes, des différents signaux ambigus issus de l’image mais aussi des connaissances à priori que possède le système visuel sur les régularités de l’environnement. Cette approche permet de saisir dans un même cadre théorique le traitement de l’information visuelle à différents niveaux d’analyse (neuronal, comportemental). Un problème théorique majeur est cependant l’aspect statique de ces modélisations. Les connaissances a priori sont statiques et seules les représentations probabilistes de l’image peuvent évoluer dans le temps. Nous nous sommes attaqués à ce problème à partir de nos travaux antérieurs sur la poursuite oculaire. Sur le plan théorique, nous avons étudié différents modèles dynamiques comme la mise à jour du prior ou la propagation spatiale d’inférence. Cette approche s’est accompagnée d’un travail expérimental spécifique pour définir les différents paramètres du modèle (estimation des variances des signaux 1D et 2D; paramètres de l’évolution temporelle du prior). Les conséquences de cette approche pour la modélisation d’un système moteur simple ont été étudiées en prenant en compte la dynamique de la boucle sensorimotrice elle-même [Mon+07a; Mon+07b; Mon+07c; Mon+07d]. Enfin, du point de vue oculomoteur, ce système dynamique peut s’apparenter à un filtre de Kalman et le contrôle de la boucle sensori-motrice implique de comprendre comment le système dans son ensemble peut contrôler de façon optimale les différentes stratégies de mouvement de l’œil [Bog+11; Fle+11].
2.3.3 Algorithmique efficace des représentations visuelles

Codage épars et robuste: quelle signification au niveau de la population? (Jens Kremkow avec F. Chavane, G. Masson, Y. Frégnac /UNIC, CNRS - Gif-sur-Yvette)

L’approche bayésienne permet de définir précisément les fonctionnalités du système qui sont autant de problèmes pour lesquels il est intéressant d’étudier des algorithmes les résolvant de façon optimale comme l’approximation donnée par la Poursuite d’Inférence (ou Matching Pursuit) [PST04b]. En s’inspirant de l’architecture du système nerveux central, nous nous sommes concentrés en particulier sur des algorithmes adaptés à des architectures dynamiques et parallèles [Per04a]. En effet, à l’image des cartes neurales, les réseaux bayesiens peuvent s’implanter grâce à des représentations explicites des probabilités d’inférence par des populations de neurones sur des cartes représentant de façon la plus continue possible l’ensemble des valeurs d’une caractéristique du mouvement [Per05]. L’intégration de l’information s’apparente alors à une diffusion sur ces cartes similaire aux algorithmes d’Équations aux Dérivées Partielles (ÉDP, collaboration avec l’équipe d’Olivier Faugeras à l’INRIA). Nous avons approfondi cette similarité en étendant les représentations utilisées pour notre problème et en nous concentrant en particulier sur les applications potentielles en traitement de l’image grâce à l’utilisation de représentations sur-complètes par ondelettes. En particulier, si on adapte alors le système aux statistiques des scènes naturelles, on obtient un codage épars et robuste dont nous avons étudié plus précisément les propriétés.

Ondélettes et extraction automatique des contours (avec S. Fischer)

Plusieurs inconvénients des algorithmes traditionnels en traitement de l’image peuvent être levés par les techniques neurales que nous avons développées. Notamment, l’utilisation courante de bases (ou sous-bases orthogonales d’un dictionnaire sur-complet) dans le traitement multi-résolution par ondelettes introduit des problèmes quant à la robustesse de la représentation aux transformations usuelles des images (translations, rotations, dilations) [Per03; PST04a]. Les techniques exposées ci-dessus permettent de donner des solutions algorithmiques efficaces et de complexité algorithmiques faibles sur des architectures parallèles. Elles permettent en particulier de résoudre des problèmes comme la détection efficace et parcimonieuse de contours quand le dictionnaire d’ondelettes est sur-complet. C’est par exemple le cas pour l’analyse multi-résolution développée par S. Fischer, basée sur une fonction log-Gabor inspirée du champ récepteur des cellules du cortex visuel primaire [Fis+07a]. Nous avons développé des algorithmes originaux de détection de contour, de compression mais aussi de séparation du bruit.
Figure 2.2: Apprentissage de représentation efficaces par codage parcimonieux. Nous montrons un montage des 324 champs récepteurs (entourés d’une bande noire) obtenus par apprentissage de type hebbien sur notre représentation parcimonieuse. Par rapport à la solution classique de Bruno Olshausen (à gauche), notre solution (à droite) produit des filtres similaires mais dont les propriétés computationnelles se montrent meilleurs (scripts téléchargeables depuis http://invibe.net/LaurentPerrinet/SparseHebbianLearning).

Apprentissage de représentation efficaces

Enfin, un problème majeur de la modélisation de l’architecture corticale et que celle-ci semble constitué de circuits élémentaires et qu’il faut donc inclure un processus d’apprentissage afin de parvenir à des représentations efficaces. Ce problème a été résolu en étendant des travaux précédents [Per04b; PST03] grâce au lien avec l’inférence statistique. En effet, l’apprentissage peut alors se comprendre comme une optimisation progressive du système en adaptant dans les réseaux que nous avons créé la forme des champs récepteurs ainsi que des interactions latérales qui leur sont liées. Ce travail a été développé lors d’une visite et d’une collaboration avec Bruno Olshausen (RNI, San Francisco) et à permis de montrer la supériorité de cette approche sur la méthode analytique classique [Per06]. Cette approche a été aussi

étendue à une architecture de type ondelettes et montre l’émergence de champs récepteurs similaires à ceux trouvés dans le cortex visuel primaire qui se révèlent être particulièrement efficace pour la représentation d’images naturelles [Per06; Per08a; Per08b; Per10b] (voir Figure 2.2).

2.3.4 Neurosciences Computationnelles: déchiffrer le code neuronal de la perception visuelle du mouvement

L’équipe DyVA s’est largement impliquée dans le développement d’outils de modélisation afin de mieux comprendre les problèmes liés à l’intégration spatio-temporelle d’information pour la perception visuelle. Ces outils se définissent

1. en premier lieu dans un cadre théorique de compréhension des mécanisme des processus neuronaux. Ceux-ci sont alors définis en utilisant les outils mathématiques pour l’étude des systèmes dynamiques complexes qui sont empruntés des probabilités, du calcul stochastique et de la physique statistique [DP10].

2. en interface avec les résultats neuro-physiologiques et comportementaux.

Nous avons construit des stratégies de simulation qui permettent de répondre à ces critères en utilisant une interface commune (basée sur Python) qui s’interface avec différents simulateurs [Dav+07; Dav+08] (Collaboration avec A. Davison, UNIC - Gif-sur-yvette, Eilif Mueller, Heidelberg).

Implémentations neurales: diversité des réponses neurales et fonctionnalités visuelles (avec J. Kremkow et N. Voges)

Une stratégie complémentaire aux deux précédentes est de confronter les connaissances neuro-physiologiques à une implémentation de ce type d’algorithme dans un cadre de modélisation neurale. Nous simulons des réseaux de neurones récurrents de large taille afin de simuler des effets de populations qui ne peuvent pas être mis en évidence dans des systèmes plus simples [Kre05]. Ces réseaux sont utilisés traditionnellement afin d’étudier leur propriétés complexes de système dynamique [Kre+07a] et nous les avons étendu pour qu’ils implantent les fonctionnalités et algorithmes désirés (collaboration avec Ad Aertsen, Freiburg). Nous étudions en particulier l’importance de certaines caractéristiques neurales pour l’implémentation de ces réseaux:

1. caractère impulsionnel de l’information neurale: les neurones génèrent des signaux prototypiques (Potentiels d’Action, PA) qui favorisent la détection d’événements neuronaux synchronisés. Nous étudions comment les PAs peuvent simplifier l’implémentation de fonctionnalités neurales et en particulier le rôle particulier de la balance active entre excitation et inhibition [Kre+07b; Kre+08a; Kre+08b; Kre+09; Kre+10; VKP08].
2. stratification des connexions corticales horizontales: le cortex révèle une connectivité par sauts et une connectivité prototypique qui permet une représentation explicite de l’information utilisée par le système inférrentiel. Nous étudions pourquoi cette architecture est nécessaire à une implémentation de stratégie de contrôle prédictif et ses implications quant à la dynamique neurale [Per+09a; VKP08; Vog+10; VP08; VP09a; VP09b; VP12].

Ces études nécessitent un important ensemble de moyens humains et informatiques pour implanter ces algorithmes et produire des simulations de tailles réalistes (millions de neurones, milliards de connections, centaines de conditions et paramètres expérimentaux) qui nécessitent des collaborations (dans le cadre de FACETS) et des moyens conséquents. Une telle méthodologie a permis des collaborations innovantes, comme par exemple en mettant en évidence des comportements non-linéaires en adéquation avec les observations en Imagerie Optique [Per+09b]. Toutefois, nous avons rencontré des limites dans cette approche, notamment dans le manque d’interactions entre les différentes disciplines (modélisation, physiologie, comportement) et c’est pourquoi j’ai participé à l’élaboration d’un nouvel axe de recherche lors du déménagement de notre équipe sur le site de la Timone.

2.4 Activités de recherche actuelles

En effet, suivant la restructuration des neurosciences sur les différents site du CNRS à Marseille, le laboratoire INCM a intégré un nouveau site sur le campus de la faculté de médecine de la Timone, l’Institut de Neurosciences de la Timone (INT). Ce déménagement a pris place fin 2010 et a donné lieu à une restructuration des équipes arrivantes pour assurer la cohésion de l’ensemble. Notamment, l’équipe DyVA est devenue l’équipe “inference and visual behavior" (inViBe). Tout en gardant de nombreux points communs avec les thèmes et méthodes développés à l’INCM, la formation de cette nouvelle équipe a permis de redéfinir son champ d’action. Notamment, l’accent a été mis sur l’intégration des pistes de recherche développées individuellement dans l’équipe et je vais développer dans cette section les principales contributions que j’ai pu apporter.

Une étape importante dans la structuration du thème de recherche a été la recherche de nouvelles sources de financement et de nouvelles collaborations. En particulier, nous avons obtenu un financement important grâce au projet BrainScaleS (commission européenne, contrat numéro FP7-269921), qui nous a permis d’envisager l’élaboration de nouveaux types d’algorithmes basés sur ces recherches. Une autre étape importante a été la collaboration avec Karl Friston à l’University College de Londres qui a permis d’étendre la portée théorique des modèles probabilistes que nous utilisons. Cette collaboration a pris la forme d’une mission longue de 14 mois (d’octobre 2010 à février 2012)
Figure 2.3: Thématiques de recherche de l’équipe inViBe: L’équipe inViBe, créée en 2010 vise à étudier différentes approches sur les mécanismes inférentiels qui sous-tendent la réponse comportementale dans le système oculomoteur. Ce schéma représente les différents aspects qui sont étudiés de la psychophysique [San+12] (cube en haut à gauche), aux modèles [PM12a; TMK10] (à droite) et à la physiologie par enregistrements extra-cellulaires de neurones de V1 analysés par Giacomo Benvenuti (en bas à gauche). Les modèles inférentiels de type bayesien ici représentés par le modèle de Montagnini et al. [Mon+07b], constituent une synergie pour rassembler les différentes facettes présentes dans l’équipe.
sous l’invitation de Karl Friston et a permis l’établissement de nombreuses collaborations dans Londres et nationalement (notamment Jim Bednar à Edinburgh). Ces différents facteurs ont contribué à la structuration du projet de recherche que je présente aujourd’hui.

Dans cette section, je vais résumer les principaux axes de recherches qui ont été développés durant cette période. En particulier, tout en gardant une lecture proche des niveaux d’étude de Marr, nous allons progressivement les dépasser pour mettre en avant les collaborations entre différents niveaux. Pour cela nous allons d’abord étudier une approche héritée de l’ingénierie des systèmes pour caractériser le système oculomoteur (Sec. 2.4.1), pour ensuite étudier le rôle fonctionnel des interactions latérales dans l’intégration spatiodimensionnelle, et en particulier le rôle du codage prédictif (Sec. 2.4.2). Afin de confronter de tels modèles avec des données physiologiques et comportementales, nous allons enfin montrer des modèles de réseaux neuronaux impulsionnels à grande échelle, tout en formalisant une théorie de décodage de cette activité neurale en terme d’information visuelle. Enfin, nous synthéterons ces différentes approches en présentant le modèle de minimisation de l’énergie libre présenté par Karl Friston et son application à l’unification des différentes théories en cours en neurosciences computationnelles.

2.4.1 Caractérisation fonctionnelle du système oculomoteur

Mesurer la vitesse et la direction d’un objet en translation est une étape computationnelle cruciale pour bouger nos yeux, nos mains dans l’environnement, attraper un objet ainsi que percevoir l’organisation de la scène visuelle et de ses éléments. Par exemple, alors que nous avons une bonne connaissance des mécanismes perceptifs et neuronaux de l’encodage et du décodage de l’information de direction ainsi que des algorithmes biologiquement plausibles utilisés dans différentes espèces, comment le cerveau traite et représente l’information de vitesse (la norme du vecteur) reste largement incompris. Des neurones sélectifs à la vitesse ont été identifiés à différents niveaux hiérarchiques des voies visuelles chez l’homme et le singe mais comment cette sélectivité est construite est encore mystérieux. Ceci explique l’absence de modèle consensuel sur cette question. Une hypothèse de travail est que ces mécanismes neuronaux, et leurs corrélates perceptifs, combine de façon non-lineaire l’information locale de mouvement extraite à travers plusieurs filtres spatiotemporels, prenant avantage de la structure multi-échelle des images naturelles. De plus, l’organisation perceptive de la scène et de ses parties doivent être pris en compte pour une intégration contextuelle et tâche-dépendante. Enfin, le code de population sous-jacent à la perception de la vitesse reste lui aussi largement mystérieux et donc nous sommes loin de comprendre comment l’information de vitesse est décodé pour contrôler des réponses (oculo)motrices et des jugements perceptifs.

Une étape fondamentale a été franchie en important dans nos méthode
Figure 2.4: **Motion Clouds.** Les *Motion Clouds* constituent un ensemble de stimuli visant à explorer de manière systématique la réponse fonctionnelle d’un système sensoriel à un stimulus en mouvement de type naturel. Ceux-ci sont optimisées pour décrire un mouvement de translation pure en plein champ et sont par construction des textures synthétisées à partir de “patches” élémentaires de mouvement semblable placés au hasard dans l’espace. L’objet d’une telle entreprise est de tester systématiquement un système en variant les paramètres de telles textures sur les dimensions perceptives principales (vitesse moyenne, direction, l’orientation spatiale et fréquence). Nous montrons ici un *espace de stimuli* comme une grille tri-dimensionnelle dont les nœuds correspondent à des stimuli et les axes des paramètres du mouvement: bande passante pour la vitesse (panneau de gauche), fréquence (panneau de droite) et orientation (partie supérieure). Chaque nœud contient un cube élémentaire qui représente le film correspondant au stimulus, avec le temps qui s’écoule du coin inférieur gauche au coin en haut à droite dans les facettes à droite et en haut. Nous avons superposés en couleur une teinte qui représente une mesure de la réponse sensorielle (ici un modèle d’énergie du mouvement) dans cet espace de stimuli. Ce genre de caractérisation permet une étude systématique du système (ici oculomoteur) qui est étudié.
une approche héritée de l’ingénierie des systèmes. En effet, il est usuel pour caractériser le système oculomoteur d’utiliser des stimuli visuel simples comme des points, lignes ou des réseaux et de varier les paramètres de ces stimuli (contraste, orientation, direction, vitesse) pour en déduire la réponse comportementale. L’avantage de cette méthode est clairement la simplicité des stimuli. Toutefois, celle-ci s’accompagne paradoxalement avec le désavantage de créer des stimuli pour lesquels l’information peut être distribuée à différents niveaux de complexité structurelle. Ainsi une ligne en mouvement apporte un signal simple de mouvement, mais inclut intrinsèquement aussi des informations de haut niveau, comme l’alignement des différentes informations locales de mouvement. Une approche inverse est d’utiliser des stimuli écologiques en utilisant cette fois-ci des scènes naturelles. Le désavantage de ces stimuli est cette fois-ci que la complexité du stimulus est trop grande alors que l’on ne contrôle pas le contenu informationnel. Une solution pour caractériser le système oculomoteur est plutôt de faire l’hypothèse qu’il infère le mouvement d’un objet à partir d’un modèle interne de ce mouvement. En paramétrisant ce modèle, on peut générer grâce au modèle direct des stimuli qui seront optimaux pour caractériser le système —sous réserve des hypothèses formulées.

J’ai formalisé un tel modèle de génération de texture qui a ensuite été implanté pour l’étude de la détection du mouvement [San+12], les *Motion Clouds*. Il s’appuie sur la formalisation la plus simple d’un détecteur élémentaire de mouvement, le modèle “Motion Energy” [AB85]. Ce même modèle peut être de la même façon considéré comme la solution du problème inverse au modèle de conservation de la luminosité qui est souvent utilisé en vision par ordinateur [ADK99]. Nous avons ensuite formulé ce modèle sous la forme d’une texture à phases aléatoires [GGM10] en paramétrisant des axes perceptifs saillants (vitesse, direction, orientation) ainsi que les largeurs de bande (variabilité) qui leur sont associées (voir Fig. 2.4). On obtient alors des stimuli aux statistiques proches des images naturelles, avec un jeu de paramètres à contrôler et avec une implantation simple1. Grace à ce type de stimuli, nous avons pu par exemple caractériser la réponse oculomotrice en fonction de la richesse du contenu fréquentiel. Cette étude, parue récemment dans *Nature Neuroscience*, nous a permis de dissocier les différents processus non-linéaires en jeu dans une tache décisionnelle ou perceptive: L’estimation du mouvement est intrinsèquement un problème multi-échelle et tâche dépendant [Sim+12] (voir Fig. 2.5). Ces Motion Clouds constituent une base pour l’intégration de différentes études aux niveaux de la modélisation (pour valider les résultats théoriques), et aux niveaux physiologiques et comportementaux.

En particulier, nous développons autour de cet ensemble de stimuli

---

1Le code de cet algorithme de génération de textures est disponible sur https://github.com/NeuralEnsemble/MotionClouds.
différents axes de recherche. Dans une première étude, en collaboration avec Andrew Meso et Guillaume Masson, nous étudions l’estimation de la vitesse en fonction du contenu fréquentiel de la texture. Cette tâche est importante car elle nous permet de dissocier les contributions indépendantes des différents canaux dans la hiérarchie des voies visuelles et ainsi de caractériser finement la dynamique de l’intégration spatio-temporelle. Les résultats préliminaires indiquent l’importance d’une information a priori telle que prédite par des modèles bayesiens [SS06]. De façon plus surprenante, nous observons aussi des phénomènes de sur-estimation de la vitesse qui peuvent être expliqués en complétant ce dernier modèle [Mes+13]. Dans une seconde étude, en collaboration avec Claudio Simoncini, Anna Montagnini, Laurent Goffart et Guillaume Masson, nous avons étudié le rôle du contenu fréquentiel des Motion Clouds sur les saccades de fixation. En effet, l’origine fonctionnelle de ces mouvements de faibles amplitudes est encore largement inconnu. En mesurant ces mouvements avec différentes classes de motion clouds, nous avons par exemple mis en évidence que certains aspects de ces mouvements corrélat avec la statistique des textures, indiquant une possible fonctionnalité exploratoire des mouvements micro-saccadiques [Sim+13]. Cette approche est donc novatrice et pertinente pour étudier les propriétés non-linéaires de l’intégration de mouvement, dans un contexte de réponse motrice ou de jugement perceptif sur la vitesse.

Ce projet réuni des psychophysiciens, des spécialistes du contrôle oculomoteur chez l’homme et des modélisateurs pour caractériser le système oculomoteur. Notre but est d’étendre le travail élaboré ensemble ces dernières années pour comprendre comment mouvements de poursuite et perception visuelle tirent avantage d’un traitement multi-échelle pour estimer le mouvement d’une cible. Nous poursuivrons notre travail de conception mathématique de stimuli de haute dimensionnalité grâce à notre modèle génératif des images naturelles. Grâce à eux, nous rechercherons comment la vitesse est encodée grâce à l’extraction de l’énergie de mouvement dans différents filtres spatiotemporels. En analysant les réponses motrices et perceptives, nous mettrons en évidence les mécanismes non-linéaires (dépendance au contraste, superposition, supra-linéarité, précision...) sous-jacente à l’intégration des sorties de ces filtres et nous modéliserons ces mécanismes dans une nouvelle version de notre modèle computationnel. De plus, nous testerons notre hypothèse que dans les scènes naturelles, ces mécanismes non-linéaires augmentent la précision des réponses et diminuent leur variabilité d’un essai à l’autre, ce qui conduit à des réponses motrices optimales. En comparant ces réponses motrices avec les jugements perceptifs, nous testerons une seconde hypothèse de travail: ces calculs non-linéaires sont dépendant de la tâche et du contexte sensoriel ou sensori-moteur. En particulier, nous verrons dans quelle mesure les structures géométriques des scènes visuelles sont décisives pour la perception, au-delà du seul calcul de l’énergie de mouvement qui est utilisée par les mouvements oculaires.
Figure 2.5: Construction de stimuli de type Motion Clouds pour différentes tâches comportementales. Pour montrer que l’estimation du mouvement est intrinsèquement un problème multi-échelle et tâche dépendant, nous avons construit le protocole suivant. (a) Dans l’espace représentant la distribution spatio-temporelle de fréquence, chaque ligne passant par l’origine correspond à des stimuli se déplaçant à la même vitesse. Un réseau en translation pure est simple et correspond à un seul point dans cet espace. Nos stimuli de texture en mouvement ont énergie distribuée dans une ellipse allongée le long d’une ligne de vitesse donné, en gardant constantes les fréquences spatiales et temporelles moyennes. La bande passante spatio-temporelle a été manipulée par des largeurs de bandes différentes suivant ces axes, comme illustré par les cubes représentant les stimuli. La performance humaine a été mesurée pour deux tâches différentes, gérées en blocs parallèles. (b) Pour le suivi oculaire, les stimuli de mouvement ont été présentés pour une courte durée (200ms) à la suite d’un centrage saccadique, visant à contrôler à la fois l’attention et états de fixation. (c) Pour la discrimination de la vitesse, des stimuli de test et de référence ont été présentés successivement pour la même durée et les sujets ont été invités à indiquer si le stimulus de test a été perçu comme lente ou plus rapide que la référence. Les résultats présentés dans [Sim+12] montrent alors des effets opposés de la largeur de bande dans ces deux tâches.
2.4.2 Rôle fonctionnel des interactions latérales dans l’intégration spatio-temporelle

Les Motion Clouds ne sont qu’un début dans la construction d’une approche systémique de la caractérisation du système oculomoteur. Ceux-ci sont en effet construits sur des hypothèses simples de synthèse du mouvement et vont nous servir de contrôle: Les scènes naturelles se caractérisent en effet par de nombreuses régularités statistiques qu’il faut alors introduire dans le modèle génératif de synthèse du mouvement. En particulier, il est plus probable que la trajectoire de l’image d’un objet suive une trajectoire continue plutôt que dis-continue. Il est remarquable de noter qu’alors les modèles qui considèrent une indépendance conditionnelle entre les mesures de mouvements voisines considèrent par conséquent qu’une trajectoire dis-continue est aussi probable qu’une trajectoire continue.

Partant de ce constat, j’ai proposé un modèle d’intégration spatio-temporelle qui propose d’inclure l’information a priori que le mouvement d’un mouvement est continu et doit donc être pris en compte dans la dynamique globale du modèle. Nous avons mis en évidence la proximité de cette approche avec celle de Burgi, Yuille, and Grzywacz [BYG00] et aussi les limites de ce dernier modèle. En effet, établir des prédictions sur un espace de position et de vitesse entraîne une explosion du nombre combinatoire de prédictions possible. À l’inverse du modèle précédent qui considère une approximation grossière de l’espace topographique de position et de vitesse, nous avons utilisé une technique de traitement de l’image appliquée au suivi de contours, les filtres particulaires. J’ai implanté un tel algorithme qui fournit ainsi une plateforme de modélisation que nous appliquons à différents problèmes.

Une première application de cette méthode a consisté à résoudre le problème de l’ouverture. En effet, ce problème est remarquable car il souligne qu’une information locale (par exemple le mouvement d’une ligne infinie) n’est pas suffisant pour caractériser le mouvement global (le mouvement d’un segment fini dans une direction non perpendiculaire à son orientation). Classiquement, il est établi que ce problème est résolu par des mécanismes spécialisés qui déetectent soit le mouvement au centre du segment, soit le mouvement des fins de lignes. Alors, il est courant d’admettre que cette dernière information résout le problème de l’ouverture par un processus de compétition. Grâce au modèle de codage prédictif basé sur le mouvement, nous avons au contraire montré que ces mécanismes spécialisés étaient plutôt une propriété émergente du système: il est suffisant pour résoudre le problèmes de l’ouverture (voir Fig. 2.6).

Il est étonnant de constater qu’avec des hypothèses simples —le codage prédictif basé sur le mouvement, nous pouvons ainsi caractériser des propriétés du système oculomoteur attribuées classiquement à des mécanismes complexes et non-linéaires. Nous avons ainsi étendu, en collaboration avec
Figure 2.6: **Un codage prédicteur basé sur le mouvement est suffisant pour résoudre le problème de l’ouverture.**

Ce modèle est constitué d’un étage classique de mesure (estimation) et d’une couche de codage prédicteur. L’étape de mesure consiste en (A) inférer à partir de deux trames consécutives du flux d’entrée (B) une distribution de probabilité de mouvement. Cette distribution est ici représentée par des échantillons dans l’espace possible des positions et vitesses du mouvement (flèches noires). Cette couche interagit avec la couche prédicteur qui constitue (C) une étape de prédiction qui prédit l’état futur depuis l’estimation actuelle (des flèches grises aux flèches noires). Dans (D), l’étape d’estimation fusionne la prévision actuelle de mouvement avec la probabilité mesurée au même instant dans la couche précédente (B), comme représenté par les flèches noires. Dans [PM12a], nous avons montré qu’un tel modèle permet de résoudre le problème de l’ouverture.
Mina Khoei (en thèse FACETS-ITN) et Anna Montagnini (inViBe-INT), notre étude à un autre modèle classique pour l’oculomotricité: l’extrapolation du mouvement. Cette extension consiste à étudier le comportement du modèle lors d’une interruption transiente de l’entrée sensorielle. En effet, il est courant —par exemple lors d’un clignement de l’œil— que l’entrée sensorielle soit perturbée ou suspendue, et il est important pour le système de représenter une certaine continuité. Celle-ci se traduit à partir de certaines étapes dans la hiérarchie du système visuel par une activité neurale soutenue pendant l’interruption, comme montré dans le cortex inféro-temporal chez le macaque [AM95]. Nous avons mis en évidence que notre modèle pouvait répliquer un tel comportement et en particulier, nous avons apporté trois points: 1) une prédiction à la fois en position et en vitesse est nécessaire pour avoir un comportement robuste, 2) la représentation du mouvement perd progressivement de sa précision lors de l’interruption, et 3) le système doit avoir accumulé assez d’information pour être dans un mode de suivi. Nous détaillerons ces résultats dans le chapitre 3. Bien que ce modèle se base sur une conceptualisation de la propagation de l’information au sein d’une carte corticale (utilisant les filtres particulières), nous verrons dans la section suivante (Sec. 2.4.3 qu’elle admet une implantation neuromimétique.

Un autre axe d’exploration est d’implanter un prior d’association local similaire le codage prédicatif dans le temps basé sur le mouvement en étudiant les régularités statistiques des scènes naturelles dans l’espace pour un instant donné. Une telle tâche est similaire à l’identification d’un “champ associatif” qui connecterait des neurones sélectifs à des orientations locales suivant leur régularités [FHH93]. Ce concept est controversé car les études qui ont montré un corrélat neural pour une telle connectivité [Bos+97] sont souvent en contradiction avec des études physiologiques [Cha+11; HIG12]. En collaboration avec Jim Bednar (DTC, Edinburgh), j’ai utilisé des travaux précédents sur le représentation en ondelettes (section 2.3.3) pour quantifier un tel champ associatif. Il en ressort deux traits principaux: 1) quand on mesure la probabilité de cooccurrence de deux contours, leurs propriétés absolues (échelle, distance) sont indépendant de leur propriétés géométriques (azimuth, angle relatif); 2) les propriétés géométriques sont suffisantes pour caractériser des propriétés de ces images comme par exemple leur catégorie (animal/ non-animal ou naturel / artificiel). Nous détaillerons ces résultats dans le chapitre 4. Ces axes de recherche —prédiction sur une trajectoire et dans l’espace— sont complémentaires et nous verrons dans le dernier chapitre (voir partie III) qu’il est possible de les combiner.

2.4.3 Modélisation de réseaux de neurones impulsionnels

L’étude que nous menons sur les régularités statistiques —dans les trajectoires des objets et dans l’espace— se doivent d’être validés par les résultats expérimentaux. Pour ce faire, nous utilisons deux approches complémentaires.
La première consiste à utiliser des simulations à grande échelle des conductivités que nous avons mis en évidence pour comprendre si ces principes s’étendent tels quels à des réseaux de neurones. La deuxième approche consiste à utiliser les données neurales collectées au laboratoire et d’utiliser nos modèles pour décoder de l’activité neurale l’information visuelle pertinente.

Dans le premier axe, nous avons implanté en collaboration avec Bernhard Kaplan, Anders Lansner (KTH, Suède) et Frédéric Chavane (inViBe-INT), et dans le cadre de BrainScaleS, des simulations à grande échelle d’une aire corticale implantant le codage prédictif basé sur le mouvement. Cette simulation est basée sur le savoir-faire du KTH en la matière et nous a permis de valider le modèle probabiliste à l’échelle neuro-morphique. Nous avons utilisé comme contrôle le protocole d’extrapolation du mouvement (voir plus haut et chapitre 3). Les résultats ont montré des résultats similaires aux résultats théoriques (notamment les trois points évoqués plus haut), ainsi qu’une propriété reliée à l’implantation neurale. Durant l’interruption, la représentation probabiliste au niveau de la population de neurones reste la même, mais le niveau d’activité global (en termes de fréquence de décharge) diminue, en accord avec par exemple les mesures dans le cortex inféro-temporal chez le macaque [AM95]. Nous détaillerons ces résultats dans le chapitre 5. Il est important de noter que ce type d’implantation est basé sur une connectivité anisotropique qui n’avait jamais —à notre connaissance— été explorée. Ce modèle fait partie des modèles sélectionnés dans BrainScaleS pour être implantés finalement sur les micro-circuits neuromorphiques à grande échelle.

Un aspect complémentaire à la simulation est l’étude du décodage de l’activité neurale. En collaboration avec Giacomo Benvenutti (thèse BrainScaleS) et Frédéric Chavane (inViBe-INT), nous étudions des modèles statistiques qui nous permettent d’extraire l’information visuelle de populations de neurones. En premier lieu, une telle approche permet de consolider les bases théoriques qui permettent de caractériser l’activité neurale, tant au niveau de la statistique de fréquence de tir des neurones que pour la paramétrisation des courbes de sélectivité neurale, par exemple en fonction de l’orientation et de la direction. Aussi, une telle caractérisation nous permettent de valider des modèles de décodage et de représentation de l’information dans l’activité neurale (comme par exemple [JM06]) et ainsi de boucler le lien avec l’implantation neurale de tels processus.

2.4.4 Unification des théories computationnelles par la minimisation de l’énergie libre (MEL)

L’énergie libre est une mesure statistique qui quantifie la surprise d’échantillonnner des données (par exemple sensorielles), connaissant a priori un modèle génératif de la synthèse (cause) de ces données. L’hypothèse de minimisation de l’énergie libre (MEL) considère que tout agent qui résiste
à la tendance au désordre (tendance dictée par le second principe de la thermodynamique) développe alors nécessairement des stratégies de MEL, car celle-ci donne une borne supérieure mesurable de la surprise, c’est-à-dire de l’entropie liée à un modèle génératif. Dans la théorie développée par Karl Friston (UCL, Londres), appliquant ce principe aux neurosciences, il ajoute deux hypothèses supplémentaires :

- Le modèle génératif est hiérarchique et non-linéaire [Fri08],
- La représentation des données comme densité de probabilité est approchée par des lois normales (approximation laplacienne) dont les moments sont explicitement codés dans la décharge neurale [Fri09].

Une première conséquence de ce principe est qu’un système l’applicant modifie son état interne pour minimiser sa surprise. Cette minimisation s’établit aux différents niveaux du modèle hiérarchique et s’implémente par la communication de transferts d’informations, généralisant ainsi la théorie établie par Mumford [Mum92]. Cette jeune théorie, en plein développement théorique et applicatif, a engendré une unification —parfois décrite— de nombreux pan des approches computationnelles en neurosciences. Elle constitue à ce titre une avancée majeure dans les neurosciences dans les 10 dernières années.

J’ai initié une collaboration avec Karl Friston en 2010 pour appliquer ce principe à la modélisation du système oculomoteur. J’ai réalisé cette collaboration au cours d’une mission longue entre octobre 2010 et février 2012. Durant cette période, j’ai intégré son équipe de neurobiologie théorique, développé un modèle unificateur pour le système oculomoteur que nous avons décliné dans un premier temps en trois projets : dans le premier, nous avons étudié comment ce principe pouvait permettre de décrire les principes computationnelles de la recherche visuelle (visual search) en l’appliquant à un modèle de saccades oculaires [Fri+12]. En parallèle, nous avons étudié dans un second papier comment un modèle de la poursuite lente pouvait expliquer des phénomènes observés chez des patients schizophréniques [APF12]. Enfin, nous développons actuellement un modèle intégré du système oculomoteur qui sera détaillé dans le chapitre 6.

La majeure innovation de ce principe est de considérer l’agent et son environnement de façon globale. Dans cette approche systémique, le système oculomoteur est considéré comme un système intégré plutôt que le chaînage de processus élémentaires de traitement comme ce qui est classiquement accepté [KL89; RGG86]. Appliqué à un agent, c’est-à-dire à un système pouvant agir sur son environnement pour minimiser son énergie libre, le principe de MEL conduit à l’inference active [FDK09]. Cette théorie permet d’unifier des modèles de natures différentes (probabilistes, modèle de contrôle, codage prédicteur, réseaux bayesiens, heuristiques sur des réseaux de neurones, ...) en proposant un langage commun [Fri10] (voir Fig. 2.7). Nous allons voir
Figure 2.7: Unification des théories computationnelles par la minimisation de l’énergie libre (MEL). Cette figure extraite de [Fri10] représente la place central du principe de MEL dans l’ensemble des théories computationnelles. En particulier, on peut noter que les principes que nous avons détaillés plus haut dans les chapitres précédents (réseaux de neurones heuristiques, principes d’optimisation, codage prédictif, ...) peuvent se rapporter à ce langage commun.

dans les prochains chapitres des exemples de ces différentes approches, pour conclure dans le chapitre 6 sur l’application de ce principe.
Part II

Activité de recherche détaillée
J’ai présenté dans la partie précédente mes axes de recherche en développant en particulier mes axes de recherche actuels. Je vais maintenant détailler ces axes en me focalisant sur quatre projets particuliers qui posent de façon détaillée les bases de mon projet de recherche:


2. Ensuite, dans le chapitre 4, nous étendrons le codage prédicatif en étudiant une quantification statistique du champ associatif. Pour montrer la puissance d’une telle représentation, nous l'utiliserons pour classifier des images, comme par exemple des images contenant des animaux ou non. Cette démarche permet de montrer que ce champ associatif —et donc une implantation utilisant un modèle avec des interactions latérales— permet de construire des algorithmes efficaces de traitement de l’information. Ils se posent donc comme une alternative novatrice aux modèles hiérarchiques et purement "en avant" qui sont couramment admis dans la littérature [SOP07].

3. Afin de valider à l’échelle neurale de tels algorithmes nous proposerons dans le chapitre 5 une implantation neurale du modèle décrit dans le chapitre 3. Les résultats montrent qu’une telle implantation n’est pas triviale et demande de définir de façon précise le micro-circuit qui implante au niveau de l’activité neurale les processus inférentiels décrits plus haut.

4. Enfin, dans le chapitre 6, je développerai la formulation de minimisation de l’énergie libre que nous avons abordé dans la section 2.4.4 en l’appliquant au système oculomoteur simplifié, mais pris dans sa globalité. Nous nous attacherons en particulier au problème des délais sensori-moteurs dans ce système et proposerons une méthodologie pour résoudre ce problème. Les résultats montreront l’émergence de comportement attribués classiquement à des systèmes complexes et que nous pouvons ici décrire de façon économique grâce à la théorie formulée par Karl Friston.

De part leur caractère détaillé, ces chapitres seront rédigés comme des publications indépendantes et en anglais, langage de référence de la communauté scientifique. Nous conclurons alors ce projet par des perspectives dans la partie III.
Chapter 3

Motion-based prediction explains the role of tracking in motion extrapolation

Abstract

During normal viewing, the continuous stream of visual input is regularly interrupted, for instance by blinks of the eye. Despite these frequent blanks (that is the transient absence of a raw sensory source), the visual system is most often able to maintain a continuous representation of motion. For instance, it maintains the movement of the eye such as to stabilize the image of an object. This ability suggests the existence of a generic neural mechanism of motion extrapolation to deal with fragmented inputs.

In this paper, we have modeled how the visual system may extrapolate the trajectory of an object during a blank using motion-based prediction. This implies that using a prior on the coherency of motion, the system may integrate previous motion information even in the absence of a stimulus. In order to compare with experimental results, we simulated tracking velocity responses.

We found that the response of the motion integration process to a blanked trajectory pauses at the onset of the blank, but that it quickly recovers the information on the trajectory after reappearance. This is compatible with behavioral and neural observations on motion extrapolation. To understand these mechanisms, we have recorded the response of the model to a noisy stimulus. Crucially, we found that motion-based prediction acted at the global level as a gain control mechanism and that we could switch from a smooth regime to a binary tracking behavior where the dot is tracked or lost.

Our results imply that a local prior implementing motion-based prediction is sufficient to explain a large range of neural and behavioral results at a more global level. We show that the tracking behavior deteriorates for
sensory noise levels higher than a certain value, where motion coherency and predictability fail to hold longer. In particular, we found that motion-based prediction leads to the emergence of a tracking behavior only when enough information from the trajectory has been accumulated. Then, during tracking, trajectory estimation is robust to blanks even in the presence of relatively high levels of noise. Moreover, we found that tracking is necessary for motion extrapolation, this calls for further experimental work exploring the role of noise in motion extrapolation.

3.1 Introduction

3.1.1 Problem statement

The continuous flow of information originating from the visual world is constantly fragmented by different sources of noise, occlusions or blanks. For instance, the path of a moving object can often be transiently blocked from the observer’s line of sight. However, one is still able to judge the current position of a moving object during such periods of occlusion as well as estimate its future trajectory at its reappearance. This ability to transform such fragmented sensory inputs into a correct continuous representation has been a major pressure in the evolution of visual systems because it leads to appropriate reactions matched to the physical evidences: It is vital to accurately follow the trajectory of a fleeing prey and stabilize its image onto the retina in order to catch it or, on the contrary, to escape from an approaching predator, despite the fact that it can transiently disappear from the line of sight [GM10]. The problem of motion occlusion is a particular case of a more general problem in neuroscience: motion extrapolation. In the absence of sensory input, the visual system can extrapolate the instantaneous position of a moving object from its past trajectory.

An essential clue to solve that problem is the prior knowledge that objects follow smooth, coherent trajectories. Following the first law of newtonian mechanics, the trajectory of an object is only perturbed by external forces. Since we know a priori that these forces are more likely to be small compared to the inertia of an object of relevance, the trajectory of objects in the physical world tend to follow smooth, straight trajectories. As such, the projection of these trajectories on the retinotopic space is such that the statistics of natural images also exhibit similar regularities regarding their visual trajectories. Such prior knowledge may be the basis of learning processes based on the prediction of the path of the trajectory. During transient blanking, it is most likely that such processes (along with the knowledge that the sensory input was indeed blanked and not definitively removed) are at the root of the mechanisms underlying motion extrapolation. Their behavioral consequences are well known. For instance, when a moving target disappears, smooth pursuit eye movements continue at the same velocity during the initial period.
of occlusion [BB03] and such a feat is only possible when observers have some knowledge on the path of motion [GWM03]. Therefore, there must be some underlying neural computations but it is yet not clear how this can be done efficiently and where it is implemented in the visual system.

This perceptual phenomenon provides invaluable tools with which we may study the mechanisms of motion detection and draw inferences about the properties of underlying neural populations. First, it is involved in different sensory modalities as sensory fragmentation exists in vision but also for instance in haptic tasks (hence in the somatosensory system). Second, it is a powerful mean to distinguish between the different computational steps of the visual motion system. Object motion information is extracted along a cascade of feedforward cortical areas, where area V1 extracts local motion information that is integrated in extra-striate middle temporal (MT) and medial superior temporal (MST) areas.

The middle temporal (MT) and medial superior temporal (MST) areas in the superior temporal sulcus (STS) process visual motion and oculomotor signals driving pursuit (see [Ilg97] for a review) and are therefore key elements in motion extrapolation. Early physiological studies in macaque monkey identified area MT as a specialized module for visual motion processing [AKL73; DZ71]. This involves extracting the speed and direction of the moving object. MT neurons respond selectively to visual motion and tuned for local speed and direction of luminance features moving in their receptive fields [MV83]. Pack and Born [PB01] have shown that the temporal dynamics of motion integration can be seen from time-varying firing rates. They showed that neuronal responses quickly progress from local to global motion direction in about 100 ms suggesting that such mechanisms are dynamical and progressive. These results pinpoint the key role of MT neurons in local motion analysis as well as global motion integration. However, these neurons respond only when the retinal image motion is present while MST neurons maintain their firing activity when there is no retinal image motion as during a transient image occlusion [NP88] or during tracking imaginary target covering the visual field outside of the receptive field currently recorded [IT03]. Similar sustained activity during target occlusion has been found in monkey posterior parietal cortex, and it is linked to an image motion prior to target disappearance [AM95]. In another study [SB08] have stimulated the retina of tiger salamander with a periodically flashing stimulus and have found various firing patterns when a flash is omitted. This sustained activity is known as “omitted stimulus response” (OSR) and is explained by a model based on tunable oscillators which extrapolate the response to the periodic stimulation even at times matched to the missing stimulus. OSR has also been reported in the flicker electroretinogram (ERG) of the human cone system [MA09].

What is the link between behavioral and neuronal signatures of motion extrapolation? Visual motion information is primarily used for gaze stabi-
Figure 3.1: The problem of fragmented trajectories and motion extrapolation. As an object moves in visual space (as represented here for commodity by the red trajectory of a tennis ball in a space-time diagram with an one-dimensional space on the vertical axis), the sensory flux may be interrupted by a sudden and transient blank (as denoted by the vertical, gray area and the dashed trajectory). How can the instantaneous position of the dot be estimated at the time of reappearance? This mechanism is the basis of motion extrapolation and is rooted on the prior knowledge on the coherency of trajectories in natural images. We show below the typical eye velocity profile that is observed during Smooth Pursuit Eye Movements (SPEM) as a prototypical sensory response. It consists of three phases: first, a convergence of the eye velocity toward the physical speed, second, a drop of velocity during the blank and finally, a sudden catch-up of speed at reappearance [BF85].
3.1.2 Different types of motion extrapolation

A classical way of studying motion extrapolation is by presenting a moving target that travels behind an occluder for a short period of time. A seminal study used timing estimation by asking participants to make a button press response at the time they judge the occluded target to have reached a particular point [Ros75]. Since then, this phenomenon has been studied at various levels (behavioral or neural), across species and modalities. For instance, motion extrapolation has been under study by focusing on various specific questions in physiology or behavior. In physiology, motion extrapolation was shown to occur in retina [GM10] and [SB08] or in higher cortical areas [AM95]. Behaviorally, motion extrapolation was studied in the context of target catching [Nij94], apparent motion [HCV08] and trajectory extrapolation for occluded or disappeared stimuli [MPE09], perceptual extrapolation of blurred visual target [FSD01], in audio visual targets [Wue+10], role of motion extrapolation in control of eye movements [MP11], blurred targets and behavior humans [FSD01]. Motion extrapolation can be carried out for lateral motion, with the target moving across the fronto-parallel plane, or for approach motion, when the object moves towards the observer [DeL04].

Herein, we investigate visual, lateral motion extrapolation as a generic paradigm to challenge prediction algorithms.

A tightly coupled phenomenon is motion inertia, which might be regarded as the perceptual equivalent of motion extrapolation for object identification. To put motion inertia in evidence, it has been shown in experiments that when one object moves and breaks into two trajectories, the trajectory that tends to be perceived as pursuing its motion is the one corresponding to the least perturbation (acceleration or curvature). Equivalently, if a moving object has been presented before, there is a strong perceptual tendency to continue seeing it in previous direction [RA83]. These findings also imply that the interactions between pairs of dots seen in sequence is affected by the history of their interactions, suggesting that probably the neurons responding to motion are directionally coupled in a feed forward way which facilitates the perception of unidirectional movement [AR87]. Assuming the existence of such a strategy, it needs to be clarified how such rules may be related to the spread of neural activity and how a neural system uses accumulated information from the trajectory of moving object in order to favor the detection of an unique, global motion. This was studied by looking at how people may extrapolate motion on a straight line [PCS92]. One can interpret that in a Bayesian way: as a prior, motion is temporally coherent, and motion inertia is a built in strategy of the visual system to respect this prior. As such motion inertia and motion extrapolation certainly share some common mechanisms, though here, we focus on the later.
3.1.3 Experimental evidence of motion extrapolation

The neural systems controlling smooth pursuit eye movements (SPEMs) are likely to be critically dependent upon motion extrapolation, in close synergy with saccades [XL07]. Several studies have shown that blanking a small moving target results in a very typical temporal profile of eye velocity (see Figure 3.1). Eckmiller and Mackeben [EM78] investigated monkey smooth pursuit behavior when a moving target briefly disappeared and then reappeared. They found that monkeys are able to continue pursuing when the target disappears for up to 800 ms. Using a similar paradigm, Becker and Fuchs [BF85] showed that humans maintain smooth pursuit up to 4 s after the disappearance of the target. They found that the eye velocity rapidly decreased about 200 ms after target disappearance. This deceleration phase lasted for about 280 ms and then the eye velocity stabilized at approximately 40 to 60% of the normal pursuit velocity. To develop an eye velocity related to the velocity of the target that preceded the extinction, the subjects needed to see the motion for at least 300 ms. Becker and Fuchs [BF85] referred to this phenomenon as predictive pursuit. This mechanism can also be at play during other open-loop responses such as anticipatory smooth tracking of a highly predictable target motion [BA91]. There is an ongoing debate of whether the origin of motion extrapolation is within the oculomotor control system [MP11] or rather occurs at the sensory level. Using event related potentials, Makin, Poliakoff, and El-Deredy [MPE09] have suggested on electrophysiological grounds that both systems may be contributing. To tease apart the relative contribution of retinal (i.e. image-driven) and extra-retinal (i.e. eye movements-driven) in the phenomenon of motion extrapolation is out of the scope of the present study and we will restrict ourselves herein to the open-loop, image-driven pursuit behavior.

Motion extrapolation seems to be a highly adaptable mechanism. We have already suggested that such behavior may be related to the regularities observed in natural scenes. One may then wonder how this may be affected by experimental conditions such as learning or reinforcement [MK03]. Becker and Fuchs [BF85] had already examined the effect of training on predictive pursuit and reported only a modest change, indicating that such a response could be under adaptive control. Using an operant conditioning procedure, Madelain and Krauzlis [MK03] found that human subjects instructed to track a small spot, tend to follow it even during the absence of sensory input. The speed decreased however to a smaller plateau value and subject often performed a catch-up saccade to track the object again. Crucially, their performance increased across sessions and subjects could pursue dots up to 4 seconds after the onset of a blank after intensive learning. One important aspect for prediction to occur is that target trajectories must be regular and clear. In another study, [BMM13] investigated used the aperture problem to probe the impact of visual motion information at target reappearance. A
moving tilted bar produces a small direction bias at pursuit initiation in the direction orthogonal to the bar’s orientation. They found a significant, albeit much smaller bias at target reappearance, as compared to pursuit initiation. Moreover, they put in evidence a strong difference in the amplitude of such a bias, depending on whether the blanking onset occurred in either the open- or closed-loop phase of pursuit. The tracking direction bias introduced by the aperture problem was significantly less in the late phase, suggesting that the oculo-motor system would switch from a preference for the sensory input (early phase) to an internal (motor-based) signal in the late phase. All these results raise the question of how we can model the different facets of motion extrapolation in a common framework.

3.1.4 Existing theories on motion extrapolation

There are a variety of models proposing different mechanisms underlying motion extrapolation. A first class of models are built upon control-like models of the visuo-oculomotor system [RGG86]. Such models were refined to specifically address the problem of motion extrapolation [CCL03] by including additional layers in a cascade model from Goldreich, Krauzlis, and Lisberger [GKL92]. These models may be subdivided into those where the predicted signal is based of some motor command [BB03] and those that specifically use the adaptation of an internal model [MK03]. Still, while these different behavioral models can fit some data very nicely, they lack a global explanation of the mechanisms underlying motion extrapolation.

Most of these models share a common mechanism: during blanking, information is inferred from past information using a smoothness constraint on possible trajectories. This is well formulated by smoothing the inferred velocity in control models with an internal positive feedback [KL89; RGG86; Rob73]. An engineering answer for such an adaptive system is a Kalman filter. It involves projecting the current estimate of the system based on the prior knowledge and correcting the predictions based on the measurement. A mix of measurement and prediction are used to estimate the current state based on their reliability reflected from their variances. Studies investigating sensory-motor transformation already suggest for a mix of measurement based signal and an internal signal based on reliability extracted from their respective uncertainties for an optimal performance in a motor task [BWH02]. Similarly, this may be expressed in as a Kalman filter, that is in a generic Bayesian framework with a clear hypothesis [WB95].

Following the idea of Kalman filter and extending the work of Montagnini et al. [Mon+07b], Bogadhi, Montagnini, and Masson [BMM13] proposed a hierarchical recurrent Bayesian framework to understand both motion integration as observed in smooth pursuit and also the predictive nature of pursuit. Probabilistic inference has been successful in explaining motion perception to a variety of stimuli [WSA02]. They are somewhat similar
to some of the engineering models proposed earlier [NS95] but allow for a more explicit formulation of the underlying hypothesis. Such a framework accommodates uncertainty in the motion information in the measurement likelihoods [HSS11; SS06; WSA02] and also expectation can be represented through the prior which can alter motion perception [SSS11]. Representing uncertainty in the measurements and prior expectation gives a simple, yet powerful framework to investigate predictive behavior of the system under investigation possibly to optimally adapt to changes in the measurements. As shown by Wuerger et al. [Wue+10] in a temporal localization task, the bias and variability show similar patterns for motion defined by vision, audition or both. Such optimal integration is consistent with a probabilistic representation of motion. The framework implements Bayesian estimation utilizing motion measurements and motion prediction. Measurements of observed input are interpreted probabilistically by a likelihood function. To detect straight trajectories with constant velocity, input motion can be temporally grouped and expressed in terms of a Bayesian generalization of a Kalman filtering [WB95], as standard Kalman filter models are not able to account for psychophysical data. A neural network model of described probabilistic framework shares interesting similarities with known properties of visual cortex and qualitatively accounts for psychophysical experiments on motion occluders and motion outliers. The approach from Bogadii, Montagnini, and Masson [BMM13] allows for a mix of prediction and measurement based on their reliability, as measured from their respective variances. The combined estimate is used to drive the pursuit response. The hierarchical framework allows to investigate the adaptive behavior of pursuit as well as the role of prediction on motion integration as observed in pursuit responses. However, this model may still be seen as an incremental refinement of previous results and does not yield a generic account on the motion extrapolation mechanism.

As we have seen, most theoretical efforts to study motion extrapolation is based on temporal coherency of motion. This assumption, as understood in a Bayesian framework, may be represented by defining a prior in the probabilistic representation of motion. This will then be integrated in the dynamical motion integration process: In a probabilistic representation of motion, posterior estimation of motion is the product of this prior and current sensory evidence (likelihood). An important question is therefore to know how to define this prior function.

3.1.5 Motion extrapolation and motion-based prediction

Yuille and Grzywacz [YG89] have shown that the efficiency of motion integration was highly dependent on the smoothness of the trajectory of the stimulus. Behavioral data showed that humans can detect a target dot moving in a smooth trajectory embedded in randomly moving dots, while the target dot is not distinguishable from noise in each frame separately.
This challenging detection task is called outlier detection and might be explained by a network of interconnected motion detectors [WMG95]. In such a network, every stimulated local motion detector sends a facilitatory signal to adjacent units. These are in turn stimulated and this sequence goes on, ultimately implementing a direction selective spatiotemporal integration. Signals from local motion detectors are made coherent in space and time and lower the threshold for detecting stimuli moving in smooth versus segmented trajectories [GWM95]. In the outlier detection case, distractor dots do not move coherently enough to accumulate information while for the target dot, precision increases gradually and as a consequence, the accuracy of velocity estimation is improved. During occlusion of target motion, that is without likelihood measurements, velocity estimation is degraded and probabilities are diffused in space and time. However, the model may still have enough momentum or motion inertia to propagate estimations of target dot’s position. This process will break down if the occluder gets too long but the motion inertia effect of target motion on distractors is visible [WMG95]. As a consequence, an important aspect of this prior is a motion-based prediction, that is, including both the position and velocity from the trajectory of motion.

Such a prior on the temporal coherency of motion can be defined in a probabilistic framework. This was formulated theoretically by [BYG00] but their neural network implementation lacked the precision needed to work on realistic input sequences. In our earlier work [PM12b], we implemented efficiently such a prior to investigate different aspects of spatiotemporal motion integration. Particularly, this model focused on the aperture problem and proposed that motion-based predictive coding is sufficient to infer global motion from all local ambiguous signals. The aperture problem is a challenging problem to study integration of local motion information [Cas+93; Lor+92; PB01]. The model proposed that instead of specific mechanisms such as line-endings detectors, the gradual spatio-temporal integration of motion. It accounts for the properties of physiological and behavioral responses to the aperture problem. First, the temporal dynamics of the solution to the aperture problem and its dependence on several properties of input such as contrast or bar length can be represented. Second, end stop cells emerge from the dynamics of the model instead of having ad hoc rules such as line-ending detectors.

The hypothesis of independence of motion signals in neighboring parts of visual space results in the failure of feedforward models in accounting for temporal dynamic of global motion integration. In those models, local measurement of global motion is the same everywhere independent of position. In motion-based prediction, the retinotopic position of motion is an essential piece of information to be represented. By explicitly including the interdependence of local motion signals between neighboring times and positions knowing the current speed along a smooth trajectory, incoherent features are explained away, while coherent information is progressively in-
This context-dependent, anisotropic diffusion in the probabilistic representation of motion also results in the formation of a tracking behavior favoring temporally coherent features. Herein, we will challenge such a model to account for the different properties of motion extrapolation.

3.1.6 Objectives and outline

This paper has been prepared in following order: In Section 3.2 we develop the same probabilistic modeling framework as the one proposed for the solution to the aperture problem [PM12b]. Moreover, we include details on the structure and implementation of the model but also details on the experimental and numerical aspects of the model. Then, we report in Section 3.3 results from experiments where we studied motion extrapolation under three different conditions for a horizontally moving dot: moving in a blanked trajectory, moving in presence of high background noise and moving in a blanked trajectory with high background noise. In the first condition, extrapolation of motion information during a blank has been studied compared to a control stimulus without blank. To stress on the role of prediction in motion extrapolation, we have done all experiments under three configuration of the model which correspond to motion estimations with and without prediction in position or velocity of stimulus. In the second condition, we have surveyed motion extrapolation by looking at states of motion tracking and its stability. In the last condition, we predict that motion extrapolation is dependent on noise and propose a behavioral experiment to test this prediction.

Finally in the discussion (Section 3.4), we will interpret these results in the light of current knowledge on probabilistic inference and dynamical systems and we will discuss the limitations of the current study along with suggestions for future work.

3.2 Model & methods

3.2.1 Probabilistic detection of motion

First, we define a generic probabilistic framework for studying motion integration. The translation of an object in the planar visual space at a given time is fully given by the probability distribution of its position and velocity, that is, as a distribution of our value of belief among a set of possible positions and velocities. It is usual to define motion probability at any given location. If one particular velocity is certain, its probability becomes 1 while other probabilities are 0. The more the measurement is uncertain (for instance when increasing noise), the more the distribution of probabilities will be spread around this peak. This type of representation can be successfully used to solve a large range of problems related to visual motion detection. These
problems belong to the more general framework of the optimal detection of a signal perturbed by different sources of noise and ambiguity.

In such a framework, Bayesian models make explicit the optimal integration of sensory information with prior information. These models may be decomposed in three stages. First, one defines likelihoods as a measure of belief knowing the sensory data. This likelihood is based on the definition of a generative model. Second, any prior distribution, that is, any information on the data that is known before observing it, may be combined to the likelihood distribution to compute a posterior probability using Bayes’ rule. The prior defines generic knowledge on the generative model over a set of inputs, such as regularities observed in the statistics of natural images or behaviorally relevant motions. Finally, a decision can be made by optimizing a behavioral cost dependent on this posterior probability. An often used choice is to choose the belief that corresponds to the maximum a posteriori probability.

3.2.2 Luminance-based detection of motion

Such a Bayesian scheme can be applied to motion detection using a generative model of the luminance profile in the image [WSA02]. This is first based on the luminance conservation equation. Knowing the velocity \( \vec{V} = (u, v) \), we can assume that luminance is approximately conserved along this direction, that is, that after a small lapse \( dt \):

\[
I_{t+dt}(x + u \cdot dt, y + v \cdot dt) = I_t(x, y) + \nu_l
\]

where we define luminance at time \( t \) by \( I_t(x, y) \) as a function of position \( x, y \) and \( \nu_l \) is the observation noise. This noise is assumed to be Gaussian with zero mean and variance \( \sigma_I^2/dt \), that is, \( \nu_l \sim N(I; 0, \sigma_I^2/dt) \). Note that for convenience we scaled variance by \( dt \) such that the variance \( \sigma_I^2 \) can be represented per unit of time, independently of the time step \( dt \). Using the Laplacian approximation, one can derive the likelihood probability distribution \( p(I_t(x, y)|\vec{V}) \) as a Gaussian distribution. In such a representation, precision is finer for a lower variance. Indeed, it is easy to show that the logarithm of \( p(I_t(x, y)|\vec{V}) \) is proportional to the output of a correlation-based elementary motion sensors or equivalently to a motion-energy detector [AB85]. Interestingly, lower contrast motion results in wider distributions of likelihood and thus posterior \( p(\vec{V}|I_t(x, y)) \). Therefore, contrast dynamics for a wide variety of simple motion stimuli is determined by the shape of the probability distribution (i.e. Gaussian-like distributions) and the ratio between variances of likelihood and prior distributions as was validated experimentally on behavioral data [Bar+08].

The generative model explicitly assumes a translational motion \( \vec{V} \) over the observation aperture, such as the receptive field of a motion-sensitive cell. Usually, a distributed set \( \vec{V}_t(x, y) \) of motion estimations at time \( t \)
over fixed positions \( x, y \) in the visual field gives a fair approximation of a generic, complex motion that can be represented in a retinotopic map such as areas V1 and MT. This provides a field of probabilistic motion measures \( p(I_t(x,y)|\vec{V}_t(x,y)) \). To generate a global read-out from this local information, we may integrate these local probabilities over the whole visual field. Assuming independence of the local information as in [WSA02], spatio-temporal integration is modeled at time \( T \) by Equation (3.1) and

\[
p(\vec{V}|I_{0:T}) \propto \prod_{x,y,0 \leq t \leq T} p(I_t(x,y)|\vec{V}(x,y))p(\vec{V})
\]

where we write as \( I_{0:t} \) the information on luminance from time 0 to \( t \). Such models of spatio-temporal integration can account for several nonlinear properties of motion integration such as monotonic spatial summation and contrast gain control and are successful in explaining a wide range of neuro-physiological and behavioral data.

### 3.2.3 Motion-based predictive coding

The independence hypothesis set above formally states that the local measurement of global motion is the same everywhere, independently of the position of different motion parts. In fact, the independence hypothesis assumes that if local motion signals would be randomly shuffled in position, they would still yield the same global motion output (e.g. [Mov+85]). As shown by Watamaniuk, McKee, and Grzywacz [WMG95], this hypothesis is particularly at stake for motions along coherent trajectories: motion as a whole is more than the sum of its parts. A first assumption is that the retinotopic position of motion is an essential piece of information to be represented. In particular, in order to achieve fine-grained predictions, it is essential to consider that the spatial position of motion \( x, y \), instead of being a given parameter (classically, a value on a grid), is an additional random variable for representing motion along with \( \vec{V} \). Compared to the representation \( p(\vec{V}(x,y)|I) \) used in previous studies [BYG00; WSA02], the probability distribution \( p(x,y,\vec{V}|I) \) more completely describes motion by explicitly representing its spatial position jointly with its velocity. Indeed, it is more generic as it is possible to represent any distribution \( p(\vec{V}(x,y)|I) \) with a distribution \( p(x,y,\vec{V}|I) \), while the reverse is not true without knowing the spatial distribution of the position of motion \( p(x,y|I) \). By doing so, we introduce an explicit representation of the segmentation of motion in visual space as an essential ingredient in motion-based predictive coding.

Here, we explore the hypothesis that we may take into account most dependence of local motion signals between neighboring times and positions by implementing a predictive dependence of successive measurements of motion along a smooth trajectory. In fact, we know \textit{a priori} that natural scenes are predictable due to both rigidity and inertia of physical objects.
Due to the projection of their motion in visual space, visual objects preferentially follow smooth trajectories. We may implement this constraint into a generative model by using the transport equation on the motion itself. Assuming for simplicity that the sensory representation is updated at discrete, regularly spaced times, then, at time $t$, during the small lapse $dt$, motion is approximately translated with respect to its velocity:

$$x_{t+dt} = x_t + u_t \cdot dt + \nu_x$$
$$y_{t+dt} = y_t + v_t \cdot dt + \nu_y$$
$$u_{t+dt} = \gamma \cdot u_t + \nu_u$$
$$v_{t+dt} = \gamma \cdot v_t + \nu_v$$

with

$$\nu_x, \nu_y \propto \mathcal{N}(x, y; 0, D_X \cdot dt)$$
$$\nu_u, \nu_v \propto \mathcal{N}(u, v; 0, (\sigma_p^{-2} + D_V^{-1})^{-1} \cdot dt)$$

Where $\nu_x, \nu_y, \nu_u$ and $\nu_v$ are random variables that blur position and velocity at each time step. These are centered Gaussians defined by their variances in position space by $D_X \cdot dt$ and in velocity space defined as in [WSA02] by $(\sigma_p^{-2} + D_V^{-1})^{-1} \cdot dt$ (where the standard deviation of the prior is defined as $\sigma_p$). Note that for convenience we scaled variance by $dt$ such that the diffusion coefficients can be represented per unit of time, independently of the time step $dt$. Here, $\gamma = (1 + \frac{D_X^2}{\sigma_p^2})^{-1}$ is the damping factor introduced by the prior and $\gamma \approx 1$ for a high value of $\sigma_p$. The update rule (see [Per07] for a derivation) assumes independence of the prior on slow speeds with respect to predictive prior on smooth trajectories. Equations (3.3)-(3.4) describe the internal model of motion, while equations (3.5)-(3.6) give a model for the noise in this model.

As a consequence, the estimated positions and velocities are slightly randomized at each frame. We controlled that this procedure led to similar results as [WSA02] (see also [Per07] for an analytical solution). Note that we return to the usual formulation described in [PM12b] when setting $\sigma_p$ to a high value (100 was sufficient in our numerical simulations). We define $D_X$ and $D_V$ as the diffusion coefficients of position and velocity, respectively in a generative model of diffusion which are scaled to be represented in units of time, independently of $dt$. In fact, it is important to properly tune $D_X$ and $D_V$ since they explicitly quantify the precision of the prediction.

We may now use this generative model to integrate motion information using a Markov random chain on joint random variables $z_t = x_t, y_t, u_t, v_t$:

$$p(z_{t+dt} | I_{0:t}) = \int dz z_t p(z_{t+dt} | z_t) \cdot p(z_t | I_{0:t})$$

$$p(z_{t+dt} | I_{0:t+dt}) = p(I_{t+dt} | z_{t+dt}) \cdot p(z_{t+dt} | I_{0:t}) / p(I_{t+dt} | I_{0:t})$$
To implement this recursion, we first compute $p(I|z_t)$ from the observation model (Equation (3.1)). The predictive prior probability $p(z_t|z_{t-1})$ is defined by the generative model in Equation (3.3) and (3.4). Note that prediction (Equation (3.7)) always increases the variance by “diffusing” information. On the other hand, during estimation (Equation (3.8)), coherent data increases precision of the estimation while incoherent data increases the variance. This balance between diffusion and reaction will be the most important factor for the convergence of the dynamical system. Overall, these master equations, along with the definition of the prior transition $p(z_t|z_{t-1})$, define our model as a dynamical system with a simple global architecture but yet with complex recurrent loops.

Unfortunately, the dimensionality of the probabilistic representation makes it impossible to implement a realistic simulation of the full dynamical system on classical computer hardware. In fact, even with a moderate quantization of the relevant representation spaces, computing integrals over hidden variables in the filtering and prediction equations (respectively Equations (3.7) and (3.8)) leads to a combinatorial explosion of parameters that is intractable with the limited memory of current sequential computers. Alternatively, if we assume that all probability distributions are Gaussian, this formulation is equivalent to Kalman filtering on joint variables. Such an implementation may be achieved using for instance a neuromorphic approximation of the equations mentioned above [BYG00]. Indeed, one may assume that master equations are implemented by a finely tuned network of lateral and feed-back interactions. One advantage of this recursive definition in the master equations is that it gives a simple framework for the implementation of association fields. However, this implementation has the consequence of blurring predictions. To describe some nonlinear aspects of motion integration we have used particle filtering in our previous work [PM12b] and here we will use it to tackle the motion extrapolation problem.

### 3.2.4 Numerical simulations

Master equations can be approximated using Sequential Monte Carlo (SMC) [PM12b]. The SMC algorithm itself is controlled by only two parameters. The first one is the number of particles $N$ which tunes the algorithmic complexity of the representation. In general, $N$ should be large enough and an order of magnitude of $N \approx 2^{10}$ was always sufficient in our simulations. In the experimental settings that defined here (moving dots), the complexity of the scene is well controlled and remains low. Control experiments have tested the behavior for several numbers of particles (from $2^5$ to $2^{16}$). We found that, except with $N$ smaller than 100, results were always very similar. However, we kept $N$ to this quite high value to maintain the generality of the results in the perspective of further extensions of the model. The other parameter is the threshold above which particles are resampled. We found
that this parameter had little qualitative influence providing that its value is large enough to avoid staying in a local minima. Typically, a resampling threshold of 20% was sufficient.

Once the parameters of the SMC were fixed, the only free parameters of the system were the variances used to define the likelihood and the noise model. The likelihood of sensory motion was computed using Equation (3.1) using the same method as Weiss, Simoncelli, and Adelson [WSA02]. We defined space and time as the regular grid on the toroidal space to avoid border effects. Next, visual inputs were 128 × 128 grayscale images on 150 frames. All dimensions were set in arbitrary units and we defined speed such that $V = 1$ corresponds in toroidal space to the velocity of one spatial period within one temporal period that we defined arbitrarily to 100 ms biological time. Raw images were preprocessed (whitening, normalization) and we computed at each processing step the likelihood locally at each point of the particle set. This computation was dependent only upon image contrast and the width of the receptive field over which likelihood was integrated. We tested different parameters values that resulted in different motion direction or spatio-temporal resolution selectivities. For instance, a larger receptive field size gave a better estimate of velocity but a poorer precision for position, and reciprocally. Therefore, we set the receptive fields size to a value yielding to a good trade-off between precision and localization (that is 5% of the image’s width in our simulations). Similarly, the contrast of the likelihood was tuned to match the average noise value in the set of images. Once fixed, these two values were kept constant across all simulations. All simulations were performed using python with modules numpy [Oli07] and scipy (respectively version 2.6, 1.5.1 and 0.8.0) on a cluster of linux nodes. Visualization was performed using matplotlib [Hun07]. All scripts are available upon request from the corresponding author.

3.2.5 Experimental procedure

All of our experimental conditions include a single dot moving horizontally at constant speed, albeit in different contexts. The dot could be blanked transiently during its displacement (blank condition), presented with a high level of noise (noise condition) or blanked with a high level of background noise (blank + noise condition). In the two blanking conditions, the target disappeared for a duration of 28% of its whole spatial period. It then reappeared at the location it should be with a continuous motion and followed the same trajectory some the remaining lapse of time. Note that as $V = 1$, the blank duration is also 28% of a temporal period. We controlled that during the blank, the local motion energy gave an uniform likelihood for all velocities, as expected. To investigate the effects of background noise, we added to the signal one of 20 linearly increasing values of standard deviation of independent, identically distributed gaussian noise (from 0.01 to 0.20
relative to a peak signal value of 1). This noise could be added to a normal trajectory of a partially blanked one.

To quantify the efficiency of motion detection, we computed different statistics. A first representation looked at the temporal dynamics of motion distribution. For this we estimated the histograms of estimated position and velocity signals along both horizontal and vertical axis. When adding noise to the motion stimulus, we also measured contrast gain in order to quantify the tracking performance by plotting the estimated values of eye velocity as a function of contrast. All simulations were repeated for 20 trials and below we plot average values across trials. Error bars are their standard deviation across trials.

One main objective of the study was to compare our motion-based prediction (MBP) model with different controls. We defined two models as limiting cases of the MBP model. A first control model is obtained when motion estimation is made dependent upon the velocity prediction but not upon the position prediction. We called this model PV, and it corresponds to the model of Weiss, Simoncelli, and Adelson [WSA02] as defined by Equation 3.2. Then, the motion-based prediction was similar but we switched off the prediction in position by setting a high value for $D_X$ with respect to the spatial period (see equations (3.3)-(3.5)). Typically, a value of 100 was sufficient. Equations of motion’s position in the generative model then were simplified to $x_t, y_t \propto \mathcal{N}(x, y; 0, D_X \cdot t)$.

A second control model was by making motion estimation dependent upon the position prediction but not the velocity prediction. We called it PX and set an high value for $D_V$ with respect to the typical physical speed of stimuli (see equations (3.4) and (3.6)). Typically a value of 100 was sufficient in our simulations. The equations for the velocity in the generative model can then be simplified to $u_t, v_t \propto \mathcal{N}(u, v; 0, \sigma^2_p)$. This PX model will perform an isotropic diffusion of information since it lacks any prediction on the velocity component of motion. This is similar to the isotropic diffusion model of Tlapale et al [TMK10] when removing the contour-based information.

### 3.3 Results

Our goal was to investigate the role of prediction in motion extrapolation in the presence of different sources of uncertainty such as a transient disappearance of the target and high background noise. We tested the dependency of our model upon prediction versus current sensory input. Motion extrapolation for a single dot moving at constant velocity was tested in three different conditions: a transient blanking trajectory in absence of background noise; a complete trajectory with different background noise levels and a transient blanking trajectory with background noise.
3.3.1 Extrapolation of motion information in a blanked trajectory

We first used as an input the movie of single dot translating along a straight trajectory and that is transiently blanked after a short period of visible displacement. This situation is similar to those used in physiological [AM95], behavioral [BMM13] and theoretical previous studies [BYG00]. By doing so, we can challenge the dynamics of information being accumulated along the occluded trajectory that is, in absence of sensory input. We measured the estimated positions and velocities of dot motion at time windows located just before, during and after the blank.

In Figure 3.2, we plot the histogram of the estimated positions and velocities obtained with the three different models: PX (motion estimation is only predictive in position, not velocity), PV (motion estimation is only predictive in velocity, not position) and MBP that is the full motion-based model where there is a predictive motion estimation both in position and velocity. Remind that PX and PV were obtained simply by choosing high values of $D_V$ and $D_X$, respectively.

In Figure 3.2, for the earliest frames, velocity histograms first spread over a larger area but progressively fit into a narrow band centered on the physical velocity ($u = 1$ and $v = 0$, see rightmost columns). This strongly suggests a convergence of the estimated motion direction towards the veridical movement of the stimulus. During the blanking period marked by vertical white dashed lines, the histograms illustrate different states. In the PX control model (upper row), velocity estimations ($u$ and $v$) are largely scattered around zero, favoring the occurrence of slow speeds. Because of the measurements, the estimations still became narrower and centered on the physical velocity of stimulus both before and after blanking. During blanking, estimated positions diffused in an isotropic manner (two leftmost columns). With both PV and MBP model configurations, the dynamics of velocity estimations paused during blanking and distributions were maintained around the last estimated values computed right before target disappearance. At stimulus reappearance, the distributions immediately resume their convergence. The estimated positions ($x$ and $y$) computed with the full MBP model exhibited a dynamics similar to velocity estimations suggesting the existence of an internal model that updates the estimations with a slow diffusion. By contrast, in the PV control model, there is no prediction to update the next stimulus position and therefore estimation histograms spread across all possible positions (see second row, left columns). This difference between PV and MBP model performance is summarized in Figure 3.3. We plot the estimated velocity obtained with each model (mean and standard deviation) together with the control condition where the dot was continuously visible. Clearly, when the stimulus reappeared after blanking, motion-based predictive estimation tend to converge immediately back to the control speed with a quick catch up.
Figure 3.2: Histograms of estimated motion for a horizontally moving dot with a blanked trajectory under three predictive configurations of the model. Blanking period is indicated with dashed white lines and each row represents full motion estimation under the configuration denoted by the inner title. Each plot illustrates the probability distribution function of a relevant variable (vertical axis) with respect to time (horizontal axis) as in Figure 3.1. The color bar on top indicates the value of probability as it is estimated for each frame (one column in each image). In each configuration, the two left columns correspond to estimated positions \((x, y)\) while the right columns represent estimated velocities \((u, v)\). At the earliest frames, for all configurations, estimated variables are scattered in a rather wide area but then gradually converge to the veridical solution \((x, y, u, v) = (1, 0, 1, 0)\). (First row) PX configuration: motion estimation is only predictive in position of motion and not in velocity. (Second row) PV configuration: Motion estimation is only predictive in velocity of motion and not in position. (Third row) MBP configuration: in this configuration, motion estimation is predictive in both position and velocity of motion and predictive information is transported anisotropically using the velocity information (compare variable \(x\) with configurations PX and PV).
Figure 3.3: Estimated velocity in PV and MBP configurations of Figure (3.2). Estimation is measured as the response of the model to a horizontally moving dot with \( u = 1 \) and a short blank in the trajectory. Blanking period has been shown with shaded area. The trace in black dashed lines represent the control condition in which trajectory of stimulus includes no blank. Error bars show standard deviation of error over 20 trials. Note the quick catch up after reappearance of stimulus in MBP configuration.

Such dynamics was more sluggish with the PV model (blue curve): motion integration did resume but at roughly the same slope as observed at the onset of the blank. Note that we did not plot the performance of the PX model in Figure 3.3, because of the very large variability of estimated velocity observed across trials (see Figure 3.2). Moreover, the rather small difference observed between PV and MBP models is due to the rather simplistic dot motions used in the present study. As explained above, the sensory layer of both models is made of a bank of motion energy filters which are highly efficient in locally detecting straight dot translations. Choosing a high value for \( D_x \) and \( D_v \) in PV and PX models is then equivalent to switching to basic sensory level without prediction in position and velocity respectively. That is to say, our working hypothesis imposes a large scale coherency constraint on stimulus trajectory as reflected in the range of values for \( D_x \) and for \( D_v \). Still, we can observe the temporal dynamics of motion estimation as already shown with more ambiguous inputs [PM12b].

3.3.2 Predictive tracking in the presence of varying noise levels

In the next step, we tested the robustness of our model when using more realistic conditions such as low contrast (or low signal-to-noise ratio) inputs. This approach is similar to the previous psychophysical work on temporal coherency and predictability of motion [WMG95]. Below, we report the performance of two model configurations (PV and MBP) when gradually
increasing the level of background noise to an horizontally moving dot. We first did it for a fully visible trajectory in order to estimate the contrast (or SNR) thresholds at which the tracking states of the model change [PM12b].

We measured the estimated velocity averaged over 20 trials when the input image was corrupted by an independent and identically distributed Gaussian noise (see Figure 3.4). In order to first explore the role of prediction for overcoming the distracting effect of noise, we set our motion estimation routine to the PV case to minimally rely on position predictions. To do so, we chose a $D_X$ value high enough so that the model did favor any estimation in particular. We then repeated the same experiment but with the full MBP model. We found that including motion-based prediction led to a more precise tracking than in the PV case, at both low and high levels of noise. We found a range of contrast (or SNR) in which the MBP model was still maintaining perfect tracking while the PV model was in the no tracking state. Two particular aspects shall be noticed. First, with the PV model, increasing the noise level gradually decreased the convergence rate of the motion detection process. Second, with the full MBP model, we observed a binary response mode (i.e. the dot is either tracked or not tracked). In the tracking state, the convergence rate was found to be dependent upon the level of noise, as in the PV configuration. Increasing noise up to a certain level results in a shift of the onset of the tracking state, until the model reached the no tracking state. Our results in Figure 3.4 are similar to the outlier detection experiment observed in psychophysics by Watamaniuk, McKee, and Grzywacz [WMG95] where a horizontally moving dot was surrounded by many other distractor dots with random movements. This psychophysical study showed that the temporal coherency of the target dot motion rendered it detectable with a high confidence as measured by a tenfold increase of detection threshold. Our modeling results are consistent with this behavioral observation. As a consequence, we similarly found a binary tracking response in the sense that tracking is rather good up to some noise level. Therefore we have either tracking or not tracking states. Furthermore, increasing the noise level imposes a delay on emergence of tracking state which is reflected in smooth slowing of initial raising in velocity traces.

To summarize the effects of noise, we plotted the efficiency of model with and without prediction in position (i.e. the MBP and PV models) for a range of contrasts and fitted them with the Naka-Rushton function [NR66] that can adequately describe the different aspects of motion integration (e.g. [Per07]) (see figure 3.5). The gain was defined as the average estimated speed and contrast as the signal-to-noise ratio. The contrast response functions were plotted for both early and late phases (as defined in figure 3.2) of tracking. Remind that here the dot trajectory was fully visible. In the early phase (red curve), both models have very different best-fit contrast saturation values ($C_{50}$ of 10.35 and 27.37, respectively) and exponents ($n$ of 7.7 and 2.19, respectively). Interestingly, one can see in the late phase.
(blue curve) a global increase in contrast gain for both models, as illustrated by a leftward shift of the curves. The PV model led to a change in only the contrast saturation parameter (from $C_{50} = 7.71$ to $n = 3.42$), while the full MBP model exhibited a significant change in both in half-saturation ($C_{50} = 7.15$) and slope ($n = 20.89$) parameters of the contrast response function. This is characteristic of the emergence of the tracking behavior in the motion-based prediction model and complements the analysis done in [PM12b]. As a consequence, we have demonstrated here that this model is sufficient to explain some well-known static non-linear computations such as the gain control mechanism implemented by divisive normalization [Rus+06; SH98]. These are essential components of neural computations and we show here that they may emerge from a predictive coding formulation instead of an explicit descriptive mechanism. In a final experiment, we will see the potential function of this tracking response.

### 3.3.3 Motion extrapolation in noisy blanked trajectory

In the last series on experiments, we combined the different sources of uncertainty studied above by simulating a noisy dot moving along a partially blanked trajectory. Above, we have shown that motion extrapolation requires enough accumulation of information from the observed trajectory parts for allowing the emergence of the tracking state. Moreover, we found that there is contrast threshold for reaching this tracking state. Since our goal is to investigate the temporal evolution of the information that is accumulated from the observed trajectory, by imposing two independent sources of uncertainty (i.e. noise and blanking) we can highlight the differences between predictive and non-predictive motion estimation.

As in the previous sections, we quantified the efficiency of motion estimation by the estimated velocity of the tracking responses (see figure 3.6). We extend the results shown in figure 3.3 by now using blanked trajectory with low noise to higher levels of noise. As we mentioned before, a quick velocity catch up as illustrated in figure 3.3 indicates the emergence of a tracking state after stimulus reappearance. Such catch up was still visible in the presence of strong noise levels, at least up to a certain threshold. We expected a general degradation of motion extrapolation by increasing noise level and consequently a lower tracking performance, down to the no tracking state. For noises higher than contrast thresholds, no such velocity catch up was observed and the models in fact remained in the no tracking state (see figure 3.6). At all noise levels, incorporating position prediction as in the full MBP model revealed several differences in performance, when compared to the PV model. In particular, the MBP model was less sensitive to noise and its dynamics at intermediate signal-to-noise ratio was brisker than the PV case. Indeed, the MBP model remained able to match the stimulus trajectory after target reappearance in the presence of relatively high noise.
Figure 3.4: Estimated velocity of PV and MBP configurations averaged over 20 trials. Stimulus is a horizontally moving dot with \( u = 1 \) which includes different noise values at background and no blank in trajectory. Colors from dark to light correspond to the response to the stimulus with noise levels between 0.01 to 0.2. (Top) Estimated velocity of (PV) configuration while motion estimation only benefits from predictions in velocity of stimulus. (Bottom) Estimated velocity of motion-based prediction (MBP) configuration, where estimation is predictive in both position and velocity of motion. This configuration tracks well up to approximate noise value of 0.13 and after that enters into the “no tracking” state. For PV configuration this state transfer happens at noise value 0.06. As noise increases, in both configurations we observe a slower convergence in estimated velocity and more importantly a temporal shift of the emergence of tracking.
Figure 3.5: Response gain functions are plotted with best-fit Naka-Rushton functions [Per07] for both PV and MBP models. Increasing contrast produces a S-shape increase in response gain whose shape changes with both time and model configurations. Similar to the psychophysical reports by Watamaniuk, McKee, and Grzywacz [WMG95], gain and half-saturation values increase from the early to late tracking phases. There is an increase in the slope of the contrast response curve in the late response of the motion-based prediction configuration indicating a transition from no-tracking to tracking states.

level (up to 0.11). In comparison, the PV model remained in the no tracking state for noise levels higher than 0.05.

In summary, we found that making the motion extrapolation task more difficult by mixing two uncertainty sources deteriorates the tracking response. This can be explained by an insufficiently accumulated information about dot trajectory in the noisy and blanking conditions. This is evidenced by the comparison of responses at corresponding contrasts between figure 3.4 and figure 3.6. The full MBP model takes advantage from predictions in position and velocity domains, in comparison to the PV case and can accommodate higher noise levels before losing its tracking ability. In addition, a stronger internal representation of motion is maintained during blanking in this case (see MBP estimations in Figure 3.2). It also more quickly converges to the true, physical motion after reappearance. These results call for similar experiments to be done psychophysically by combining these different sources of uncertainty.

### 3.4 Discussion

In the present study, we investigated the role of motion-based prediction [PM12b] in motion extrapolation during target blanking, a condition frequently used in psychophysical, behavioral and neuronal studies to measure how the brain maintains an accurate representation of target motion despite large fluctuations in the input (e.g. [AM95; BF85; BMM13]). Our
Figure 3.6: Motion extrapolation with sensory noise: Stimulus is a horizontally moving dot with $u = 1$ which includes a blank as shown with shaded area. In addition there is a sensory noise and colors from dark to light correspond to noise levels increasing linearly between 0.01 to 0.2. (Top) Estimated velocity of model under PV configuration while motion estimation only benefits from predictions in velocity space. (Bottom) Estimated velocity of motion-based prediction (MBP) configuration, where estimation is predictive in both position and velocity of motion. In both configurations, increasing of noise corrupts tracking performance and after blank response converge only for noise values under a threshold and then enters to no tracking state. This threshold for PV and MBP configurations are 0.05 and 0.11 respectively. Note that the quick catch up after reappearance of stimulus never appears in PV but only in MBP in the cases in which a tracking state stabilized before blank.
goal was to test how the motion prediction framework described in our previous work [PM12b] can be extended to these conditions.

First, we probed the dynamics of motion extrapolation by measuring the impact of a transient absence of the stimulus, as imposed by a short blank in trajectory of the stimulus. We found a prototypical temporal pattern characterized by a pause in the motion integration process during the blank and a quick recovery of the actual position of the dot. This model behavior was largely different when turning off the anisotropic component of motion-based prediction. In this PV incomplete model, at the end of the blank, the integration dynamics resumed at a convergence rate similar to the one observed at the initial target motion onset. This difference can be explained by the fact that the full model can maintain a nearly accurate representation of the target trajectory in both position and velocity domain. In this regard, the full MBP model is more consistent with both physiological (e.g. Assad and Maunsell [AM95] and Newsome and Paré [NP88]) and behavioral (e.g. [BB03; BF85; BMM13]) observations. Interestingly, the comparison between PV, PX and the full MBP model further highlights the need of both position and velocity informations for correctly maintaining and predicting an accurate representation of target motion, an aspect that has been already introduced at the theoretical level [BYG00; PM12b; XL07].

An important issue was to answer to the question raised by the experimental study of Assad and Maunsell [AM95]. In monkeys, while MT does not represent motion during the blank, it seems that such information can be preserved in upstream cortical areas such as MST [NP88]. This later result is compatible with our approach, where neurons remain active during the transient disappearance of the stimulus, but it is still largely not known how and why such a dichotomy would emerge in the visual system. We demonstrate that a two layer model where motion information is primarily extracted locally before being diffused along a particular path can provide a solution. Such architecture presents the advantage of mixing different spatial and temporal scales and can be implemented in many biological systems, from retina to cortex. Future works will be conducted to propose a biological plausible implementation of our diffusion mechanism.

To further explore the model dynamics, we tested its robustness by adding background noise in different trajectory conditions. Increasing the background noise induced at some threshold value a sharp change in the dynamics, the model shifting from tracking to no tracking states. Such sharp transition as a function of signal-to-noise ratio is consistent with behavioral studies (e.g. [Spe+05]) showing a strong nonlinear relationship between pursuit gain and contrast (see [MP12] for a review). Interestingly, this sharp nonlinearity of the transition between tracking and non-tracking states —and which is classically implemented by some well-known static non-linear computations such as divisive normalization [Rus+06; SH98]— emerges here as a property of the dynamical system. The theoretical link
between Bayesian inference and divisive normalization has been already suggested by several authors (e.g. [HKC02]) including us ( [Bar+08]). The current study emphasizes that dynamical inference as implemented here can also reproduce the temporal dynamics of normalization mechanisms through lateral interactions [RMC12]. Further work remains to be done to validate this analogy in particular with respect to the adaptation of this non-linearity to the dynamical statistics of the input.

Our model investigates at an abstract level, the computational advantages of anisotropic diffusion of information within a probabilistic representation of motion. Previous work from Burgi, Yuille, and Grzywacz [BYG00] has suggested that there are multiple analogies of this computing architecture with the structure of neural computation in cortical areas. They originally proposed a constructive approach to implement such motion-based prediction with neural fields. However, their implementation was limited by severe constraints on the simulation of such neural-networks on classical computers. Indeed, this parallel structure is rather not optimal for a sequential computer and necessitate a large amount of memory to achieve a sufficient precision. Hopefully, the advent of novel computational architectures (clusters, neuro-morphic hardware) will foster the precision of the implementation of such models in a more biologically realistic fashion.

A last advance of our model is its ability to reproduce the dynamics of different brain responses to transiently occluded target, from neuronal activity up to highly accurate behaviors such a voluntary pursuit eye movements. Thus, our model has the potential to unify different approaches that were previously proposed to understand motion extrapolation. For instance, recent behavioral experiments imposing a blank during the straight trajectory of a tilted line shows complementing results in the light of our own results [BMM13]. Indeed, they show that if the object is tracked long enough and the blank is short enough, the bias that is characteristic of the aperture problem (the eye following first the direction perpendicular to the segment) disappears. This data is well fitted by a two-layer Bayesian network stacking a sensory and a motor levels. They explain motion extrapolation as a feed-back loop from the representation of the position of the eye to the sensory stage. Our model proposes that a complementary mechanism could be motion-based prediction and that the sensory representation of motion is sufficient to explain motion extrapolation. As we were careful to study the early stage of the tracking response (such that there can be no feedback from a motor stage), we predict that such systems should work in synergy and allow a more complete modeling of motion extrapolation. The main novelty of such scheme is that a simple generic framework —motion-based prediction— may explain a large range of mechanisms that are often explained by the explicit modeling of specialized computations. In our earlier work [PM12b], we have shown that motion-based prediction is sufficient in solving the aperture problem and that the specialized mechanisms that were
long supposed to be the source of this solution (texture-independent motion trackers, line-ending detectors, ...) instead emerged from the response to the model.

Many low-level classical problems such as motion extrapolation, the aperture problem or anticipation poses fundamental questions about the computational properties of large-scale networks of neurons. Moreover, their signatures can be found in many different species or neuronal architectures. They are shared by different sensory systems and can therefore be used as a way of unifying the search for generic computations using population codes. Complementing our previous work on the aperture problem [PM12b], we have shown here that the same architecture can solve another instance of low-level uncertainties, extrapolating current trajectories in the absence of sensory evidence. This study demonstrates the need to elaborate generic computational solutions that can eventually be implemented through realistic mechanisms such as divisive normalization mediated by lateral interactions.
Chapter 4

Statistics of edge co-occurrences are sufficient to categorize natural images

Abstract

Making a judgment about the category of a visual scene, such as whether it contains an animal, is typically assumed to involve high-level associative brain areas. Previous proposals use a hierarchy of processing steps that progressively analyze the scene at increasing levels of abstraction, from contour extraction to low-level object recognition and finally to object categorization [SOP07]. We explore an alternative hypothesis that the statistics of edge co-occurrences are sufficient to perform a rough yet robust (translation, scale, and rotation invariant) scene categorization. Using a scale-space analysis coupled with a sparse coding algorithm, we first achieved detailed and robust extraction of edges in images from different categories (natural, man-made, or containing an animal). We then computed the "association field" for each category by computing the statistics of edge co-occurrences. These statistics differed strongly between categories, with images of man-made objects having more straight and parallel configurations, and images of animals having more co-circular (curved) configurations.

4.1 Introduction

The capacity of the human visual system to robustly extract the category of an image is largely unchallenged by artificial systems, highlighting our ignorance of the neural mechanisms underlying categorization. Human performance in tasks like determining whether an image contains an animal is surprisingly accurate even at very rapid time scales [TFM96]. For instance, in a categorization based on eye saccades, subjects begin the eye movement
as early as $\approx 150$ ms after an image is briefly flashed for 20 ms [KT06]. This speed is surprising, because it is believed that categorization is performed at the end of a sequential, feed-forward chain of processing along the different layers of the hierarchy of the visual system: Starting from a rough sketch in the retina and the primary visual cortex [SP10], information flows to progressively more semantic concepts in the associative areas [Fre+01], and then to the motor areas where the decision is ultimately taken [RP00]. If this process is to complete by $\approx 150$ ms, there is time for at most a few spikes in each neural layer (the “one spike per neuron hypothesis” [Van+98]), which puts severe constraints on how much processing can occur in each neural region. A hierarchical computational model of these ideas by Serre, Oliva, and Poggio [SOP07] successfully replicated the behavioral data, matching human performance of about 82% correct categorization for a backward masking protocol [Bac+05], but it is not yet known whether the computational processing involved at each level could be performed in so short a time in humans.

Interestingly, ultra-rapid categorization in humans is highly robust to eccentricity of image presentation [Tho+01] or to a rotation of the image [CS11], which could suggest that this process is based on low-level cues instead. However, no model of low-level mechanisms has so far been able to match the human performance, and Serre, Oliva, and Poggio [SOP07] concluded that "The poor performance of simple classification strategies indicate that it is very unlikely that human observers could rely on low-level cues." (see their SI, Table 2). Other approaches have also been proposed, such as categorizing images using their spectral amplitude envelopes [OT01; TO03]. However, those models are not compatible with recent psychological observations that show that categorization in humans is robust to changes in the spectral envelope [GR09].

Here, we test the hypothesis that scene category can be determined from the statistics of edge co-occurrences, that is, by low-level properties of the rough sketch. Oriented edges in images of natural scenes tend to be aligned in co-linear or co-circular arrangements, with lines and smooth curves more common than other possible arrangements of edges (the “good continuation law” of Gestalt psychology). The visual system appears to take advantage of this prior knowledge about natural images, with human contour detection and grouping performance well predicted by such an “association field” [FHH93] between edge elements. Geisler et al. [Gei+01] have estimated this prior information computationally by extracting contours from a database of natural images and deriving statistics of edge co-occurrences. They showed that these statistics could predict behavioral data from humans in a line completion task. One possible candidate substrate for implementing an association field in mammals is the set of long-range lateral connections between neurons in the primary visual cortex (V1), which could act to facilitate detection of contours matching the association field [CM04] and
to either depress or highlight breaks in this continuity [HBG11]. To fill this role, lateral connections would need to be orientation specific and aligned along contours [HHF03], and indeed such an arrangement has been found in the primary visual cortex of the tree shrew [Bos+97; HBG11] and the monkey [SB01]. Such an elementary circuit could serve as the basis of synchronous activation of neurons along a contour [Sam+06]. This circuitry is compatible with the implementation of the good continuation of co-circular contours in natural images [Sig+01], and of curvature detection [PZ89].

However, it is not yet known whether the pattern of edge co-occurrences in any given image would be sufficient to extract the category of this image, which has been considered a much higher-level task involving semantic analysis. To investigate this possibility, we computed the statistics of edge co-occurrences in natural images extracted from different databases (natural, man-made, or containing an animal), and then measured the categorization performance of classifiers trained on the observed statistics. The paper is organized as follows: First, we replicate qualitatively the results from Geisler et al. [Gei+01] with the model outlined in Figure 4.1, confirming that information about continuations (co-linearity and co-circularity) appears consistently in natural images. In the Methods section at the end of the paper, we will detail the modified version of the algorithms from Geisler et al. [Gei+01] used to extract edges before computing the statistics of edge co-occurrences. Then, we present the results of a simple categorization scheme using the pattern of statistics computed for different categories. Compared to other natural images, we show that a man-made environment has a significantly higher probability of co-linear and parallel edge elements, as one might expect, whereas images with animals have a significantly higher probability of curved configurations. We then demonstrate that these patterns are sufficient to provide a rough, yet robust categorization with a similar performance level as that achieved by Serre, Oliva, and Poggio [SOP07], and therefore comparable to behavioral data on humans. We also show that the pattern of errors made by the model is more similar to the human data than for the hierarchical model by Serre, Oliva, and Poggio [SOP07]. Finally, we discuss the implications of these results for improving the efficiency of image analysis algorithms, and more generally for our understanding of neural computations in the visual system.

4.2 Results

4.2.1 Statistics of edges

Our goal is to study how the statistics of edge co-occurrence vary across three image categories, so we defined three testing databases. The first two consist of the image databases used by Serre, Oliva, and Poggio [SOP07], which contain either animals at different close-up views in a natural setting, or
natural images without animals (600 images each)\footnote{Available at \url{http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07.}}. A third database for comparison consists of self-acquired images from a biology laboratory setting, containing 600 indoor views of furniture, windows, and doors and cages in which animals are reared.

**First-order statistics**

One obvious candidate representation for categorization is the first-order statistics of edges. In natural images, edges are more frequently aligned to the cardinal axes, especially for man-made scenes, as has been reported and modeled previously by others [GLS11]. As the spectrum of edges is localized in the Fourier domain [Fis+07a], the representation of first-order statistics of edges is equivalent to using the amplitude spectrum obtained by Fourier analysis of the raw image. The spectral signature of scenes has previously been used by computational models to infer scene categories [OT01; TO03], and the human visual system could take advantage of these low-level natural image statistics. To compare with these previous results, we computed first-order statistics on the sparse representation described in the methods section. The histograms yielded similar results to those found on the amplitude spectrum of the raw image [OT01; TO03]. However, these first-order statistics, while tending to be different on average for different scene categories, are also highly variable within each category. The first-order histogram is highly dependent on geometrical constraints that are independent of the scene category, like the field of view (close-up or full-field view) or the orientation relative to the horizon, and we show below that they are not particularly reliable for classifying individual images into these different categories (see Table 4.1). Most importantly, these results fall to chance level with a rotation of the image or to changes in the spectral envelope, in contradiction with behavioral results [CS11; GR09]. First-order statistics are therefore a relatively poor indicator of scene category.

**Statistics of edge co-occurences**

Statistics of edge co-occurrences could represent a better alternative. Indeed, image semantics seem to depend not on spatial-frequency amplitude, but rather on phase information [OL81], which is also essential for discriminating textures [MK10]. Like Geisler et al. [Gei+01], we have chosen to compute the histogram of edge co-occurrences, that is, the frequentist probability of an edge knowing a reference edge (yielding \(N \cdot (N - 1)/2 = 523776\) samples per image when using \(N = 1024\) edges as we do here). This histogram is a 4-dimensional function of (1) the distance \(d\) between two edges, (2) the difference of azimuth \(\phi\) of the center of one edge with respect to the position and orientation of the reference edge, (3) the difference of orientation \(\theta\).
between the two edges, and (4) the ratio of edge scales $\sigma$ (see diagram in Figure 4.1-C). By definition of our representation, this set of statistics is independent of translations, rotations in the image plane, and scalings.

First, we replicated the results of Geisler et al. [Gei+01] on a set of natural images to validate our procedure, from the edge representation to the extraction. We first computed similar projections of the histograms as in Geisler et al. [Gei+01] and show qualitatively similar results despite the different datasets and methods used. As in Geisler et al. [Gei+01], the finding is that in natural images, edges are more likely to be organized in co-linear or parallel textures (see Figure 4.1-D) and along co-circular paths with a prior for low curvatures (see Figure 4.1-E). What is more interesting is that when using images from different environments such as a man-made environment (brownish edges), one finds a different pattern, where co-linearity dominates. This qualitative difference clearly indicates that the statistics of edge co-occurrences differ between databases. However, the precise way in which these sets differ is not necessarily clear, which will be analyzed in the next section.

Independence of variables in second-order statistics

The full set of second-order statistics is a function of four variables, which is difficult to plot and analyze, and so we considered whether it was possible to factorize this function into components that can be analyzed separately. We computed the mutual information of the joint probability with the 12 possible combinations of $d$, $\phi$, $\theta$, and $\sigma$. This calculation gives different Kullback-Leibler distances [CS02] in bits between the factorizations and the original function, in order to measure the independence of each factor. For all sets of images, four good candidate factorizations emerge (see Table 4.2): $p(\theta, \sigma, d) \cdot p(\phi)$, $p(\sigma, d, \phi) \cdot p(\theta)$, $p(\phi) \cdot p(\theta) \cdot p(\sigma, d)$ and $p(\phi, \theta) \cdot p(\sigma, d)$. An emergent pattern is that we may separate the characteristic angles ($\phi$ and $\theta$, individually or together) from distance-related statistics ($d$ and $\sigma$).

At this point, it is convenient to define $\psi = \phi - \theta/2$ as it is symmetric with respect to the choice of the reference edge (see methods for the use of geometrical symmetries). The angle $\psi$ is the angle between (1) the mediator of the segment joining the edges’ centers and (2) the line joining the center of this segment to the intersection of the normal of the segments (see Figure 4.1-C). Note that for a pair of edges on a common circle, we have $\phi = \theta/2$, that is, $\psi = 0$. This convention gives a simpler representation of circularities (for similar approaches see [HBG11; MK10; Sig+01]), and $\psi$ will denote the difference of azimuth in the rest of the paper. Doing similar independence measures on $\psi$ and $\theta$ further showed that these variables are relatively independent. Thus, we can characterize images by two independent indices, one for collinearity, one for cocircularity. To keep the generality of the results, we will in a first step study the whole $p(\psi, \theta)$ map.

The distribution $p(d, \sigma)$ proved to be quite similar across the different
classes of images, as it is more characteristic of the overall configuration of
the scene than of the objects within it (see Figure 4.2-(A,B)), while a clear
difference is seen for different angle configurations $p(\phi, \theta)$ (see Figure 4.2-C
and Figure 4.3). As its most representative projection can be reduced to
2 dimensions, we can plot the full probability as a “chevron map” $p(\psi, \theta)$,
where each chevron corresponds to a possible configuration of the angles
$\psi = \phi - \theta/2$ and $\theta$. Such a map is shown in Figure 4.2-C, with the saturation
of the colored circle indicating the frequency of occurrence of each possible
chevron configuration, i.e., for all possible difference of azimuth values $\psi$
on the horizontal axis and difference of orientation $\theta$ on the vertical axis.
Red denotes more frequent than a uniform reference, while blue denotes less
frequent.

Then, Figure 4.3 shows how the chevron map differs for the other two
datasets, now relative to the map computed for the natural dataset. Using
the variability on an image-by-image basis, we have plotted the 30% most sig-
ificantly different configurations in each map, with respect to the estimated
$z$ scores for each chevron. A first observation is that Figures 4.2-C and 4.3
show the configuration in a more compact fashion than Figures 4.1-D and 4.1-
E. In man-made versus natural environments, there is a significant excess of
parallel and co-linear edges, with a maximum for the co-linear co-occurrence
being about 2 times more likely than in natural images. Interestingly, in
animals versus natural scenes, there is a relative excess of co-circular and
converging configurations, with a maximum being about 1.2 times more
likely than in natural images. Note also a significant decrease for some
configurations for man-made images than for other natural images (with a
minimum being about 0.6 times less likely). This last point is consistent with
the observation from [HBG11] that significant relationships may be either
facilitating (for instance to group co-linear edges), but also depressing, to
rule out some configurations as a priori less probable.

An important issue is to examine if the qualitative differences that we
observe in the chevron maps can be assessed quantitatively. In order to
do that, we have built a simple classifier to measure if this representation
is sufficient to categorize different image categories. Such a finding would
suggest that information contained in the statistics of edge co-occurrence in
natural scenes may be used instead of or alongside a hierarchical analysis of
the visual scene, when making a quick judgment as in rapid-categorization
tasks.

4.2.2 Edge co-occurrences are sufficient to categorize natural
images

Here, we consider whether alternative low-level representations could be
more successful than those tested by Serre, Oliva, and Poggio [SOP07]. For
each individual image, we constructed a vector of features as either (FO)
the histogram of first-order statistics, (SO) the full histogram of edge co-occurrences, or (CM) the histogram $p(\psi, \theta)$ corresponding to the chevron map. To compare the representational power of each type of feature vector, we gathered these vectors for each different class of images and tested a standard linear Support Vector Machine (SVM) classification algorithm, as described below. Our results can be compared directly to those of Serre, Oliva, and Poggio [SOP07], who used the same classifier on both the last level of their hierarchical representation (successfully), and directly on the raw images (unsuccessfully). They can also be compared with the other unsuccessful low-level representations tested by Serre, Oliva, and Poggio [SOP07], such as the mean luminance, a single-template SVM classifier, texton features, global (context) features, or the output of their model V1 complex cell layer.

We first randomly divided each database into a training and a testing sub-set to validate the categorization performance. Then, we used the standard SVM library as implemented in Pedregosa et al. [Ped+11]. In order to evaluate a distance between histograms, we used the Jensen–Shannon divergence distance as the metric for SVM [CS02], directly supplying a precomputed Gram matrix of the distance between each pair of histograms to the classifier. We used the default parameters of the method. Other choices of parameters or of kernels (that is between linear, radial basis functions, or precomputed) gave qualitatively similar results. Fitting the classifier to the training set was done using an automatic line search algorithm from the same library [Ped+11]. The results of the SVM classifier are usually given as the precision, recall, or f1-score; here we used the latter to directly compare our method to that of Serre, Oliva, and Poggio [SOP07].

Table 4.1 gives the performance of different categorizations for the three types of representations, where several patterns can be seen. First, databases that are qualitatively different (such as natural versus artificial images) are very well categorized, with accuracy over 98% when using the full statistics of edge co-occurrences. For images of man-made objects this result may be obvious, given their prevalence of highly regular co-linear and parallel edges. It is perhaps more surprising, particularly given the claims from Serre, Oliva, and Poggio [SOP07] that low-level cues were unlikely to work, that we also achieved quite high, robust performance for classifying images containing animals versus other natural images. Second, it is interesting that results for the chevron map are almost as high as when using the full probabilities, confirming that the performance of the classifier comes primarily from a geometrical feature rather than a viewpoint-dependent feature (such as the scale of edges). This confirms our claim that configuration and geometrical variables are independent (see Figure 4.2).

Our hypothesis is that classifying images containing animals versus other natural images was successful because of the higher prevalence of co-circular and converging contours in images containing animals (see Figure 4.3). The match of the performance to human data extends to individual images, with
Table 4.1: Classification results for different sets of natural images (50% is chance level) and different sets of features: the first-order statistics (SVM 1), the chevron-map subset of the second-order statistics (SVM C), or the full statistics of edge co-occurrences (SVM 2).

<table>
<thead>
<tr>
<th>Database 1</th>
<th>Database 2</th>
<th>SVM 1</th>
<th>SVM C</th>
<th>SVM 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>Artificial</td>
<td>89%</td>
<td>96%</td>
<td>98%</td>
</tr>
<tr>
<td>Natural</td>
<td>Animal</td>
<td>70%</td>
<td>79%</td>
<td>82%</td>
</tr>
<tr>
<td>Natural (noise)</td>
<td>Animal (noise)</td>
<td>70%</td>
<td>76%</td>
<td>79%</td>
</tr>
</tbody>
</table>

performance on each image much more closely matching human behavioral results than for the hierarchical model of Serre, Oliva, and Poggio [SOP07] (see Figure 4.4).

4.2.3 Robustness to noise, translation and rotations

Note that by definition, our measure of the statistics of edge co-occurrence is invariant to translations, scalings, and rotations in the plane of the image (unlike the first-order statistics). Thus, under any of these transformations, one can efficiently differentiate between images from different categories. This property makes it possible to explain the rather unintuitive result that ultra-rapid categorization in humans is relatively independent to rotations [CS11] (see also the supplementary information of Serre, Oliva, and Poggio [SOP07]). We also performed the same classification where images from both databases were perturbed by adding independent Gaussian noise to each pixel such that signal-to-noise ratio was halved. As can be seen in Table 4.1, results are degraded but still quite good. Edge extraction in the presence of noise may detect false edges, but the underlying statistics of the chevron maps are still robustly captured, thanks to the high number of co-occurrences that are measured.

4.3 Discussion

Using standard tools from different disciplines (image processing, machine learning, neuroscience), we have provided a novel categorization algorithm that has potential impact on all three fields. Our main results are that statistics of edge co-occurrence can be computed reliably for natural images, and that they contain important information about local geometry that is sufficient to categorize individual images accurately. These results call into question previous claims that a hierarchical, semantic-like, analysis of the visual scene is necessary for classification into high-level categories [SOP07].

We speculate that the observed differences in second-order statistics for images with animals have an underlying basis in the physical constraints...
governing the shapes of animals, compared to other natural stimuli. Specifically, animals generally have compact shapes, constrained by their capacity to move, unlike plants, for instance, which are rooted in one location and often become elongated as they stretch to reach towards sunlight. The bodies of animals are also often enveloped by a flexible membrane, which by the constraint of minimal surface area will tend to assume a circular shape, just as a bubble will. These overall principles for shape patterning are echoed in subparts of the body, such as the eyes and the hair, markings, and textures that pattern the skin of the animal, further biasing the statistics. Conversely, man-made objects tend to have long, straight lines, with sharp orthogonal edges due to their methods of manufacture. We would expect that other categories of objects could similarly be distinguished by their second-order statistics, assuming that the form of those objects follows their function in ways analogous to the categories tested here.

Several improvements to the method could be considered for future applications. First, the image processing methods that are used to represent the data that is used to reach a classification result can be extended. For instance, edge extraction algorithms are mostly based on measuring a match between the raw image and a model of an edge (a log-Gabor patch, in this case). Using the average prior on edge co-occurrences (knowing their distance as in Figure 4.2-B or edge configuration as in Figure 4.2-C), corresponding to average knowledge across all image classes (the so-called “association field”), one could improve the measure of this match by accumulating knowledge from neighboring edges, thus improving the overall signal-to-noise ratio. Moreover, using the prior knowledge that this image belongs to a certain class (for instance due to some contextual cue), this match can be improved by using the expected second-order statistics for that class as a prior probability. Results from monkey primary visual cortex suggest that animals may similarly be able to modulate their association field dynamically. For instance, McManus, Li, and Gilbert [MLG11] reported a gating of horizontal connections dependent on the task (detecting collinear or cocircular edges), which is similar to what we show in figure 4.3 if we assume that these tasks correspond to the categorization of natural images versus man-made or animal images. Finally, these algorithms can also be complemented with a segmentation scheme using a local classification rule similar to the one we proposed here. Images containing animals usually consist of a figure (the animal), and a background consisting of a natural image without an animal. Classifying the category of each edge using such a method would itself allow refining the histogram for each class and the corresponding chevron maps, along with providing a segmentation of each image. Ultimately, such a method should help to improve the measurement of edge co-occurrences and the different patterns associated with different classes of objects.

We predict that this result can be used to improve existing machine learning algorithms for visual scene comprehension. Despite the intensive focus
in the machine-learning community on improving algorithms for classifying images, to our knowledge the second-order statistics have not previously been tested as a representation. In contrast, hierarchical models such as deep-learning networks have recently attracted a great deal of attention, in part due to their presumed similarity with the architecture of the cortex. The dependence between variables within one level of the architecture is often overlooked, assuming that it would lead to intractable algorithms. Here, we focused on the statistics of edge co-occurrences and used a classical classification method (SVM) as a generic illustration for its performance. The objective was to prove the usefulness of such a simple, one-layer representation. Ultimately, this approach could be integrated with more complex models such as the one proposed by Serre, Oliva, and Poggio [SOP07]. The probabilistic setting common to such models would allow a seamless integration that has already been proven efficient for predicting biological results [LM03; RB99].

Finally, these results are highly relevant for neuroscience. First, they challenge the premise that ten “processing layers” are necessary to categorize image classes, a premise that is the root of the hypothesis that ultra-rapid categorization necessarily implies that only a few spikes per neuron are used [Tho+01]. Instead, they suggest a complementary source of information based on the computation of the probability of co-occurrences using intra-cortical connectivity (lateral connections). As such, we offer a simple explanation of the role of lateral interactions in lower visual areas [Bos+97; HBG11]. This source of information would be integrated at the same time as the feed-forward pass and would share some of the same mechanisms, such that they would lead to a progressive refinement of the representation of information. We propose that the statistics of edge co-occurrences play an important role in the performance measured in experiments measuring ultra-rapid categorization. A further prediction would be that that categorization could be manipulated by modifying edge co-occurrence in a set of images (e.g., producing an image containing animals but with the statistics of the ‘natural’ set, and vice versa); rapid human responses are predicted to primarily follow the statistics of edge co-occurrences, rather than the actual presence of an animal. I.e., this approach predicts that an image with significantly more curves than a typical image would be falsely detected to contain an animal, in a rapid categorization task.

Second, these results have consequences for our interpretation of the observed pattern of lateral interactions in V1, such as the results of Bosking et al. [Bos+97], and therefore on our understanding through models [Bed12; GO08]. If the lateral connection patterns are learned by experience-driven plasticity [CK90], we predict that the patterns in wild-raised tree shrews would be very different from those measured by Bosking et al. [Bos+97] and analyzed by Hunt, Bosking, and Goodhill [HBG11], with shorter-range correlations and less emphasis on co-linear continuations. This prediction
can be tested in future experiments on matching groups of animals reared in different environments. From our results, we predict that $p(\theta, \psi)$ should be adaptive, since it varies strongly across environments, while $p(d, \sigma)$ could be hard-wired. The biggest challenge is to understand how such probabilities may be represented in the neural activity and how the probabilistic calculations could be implemented in animal brains.

4.4 Methods

A method for measuring the statistics of edge co-occurrences in natural images was demonstrated by Geisler et al. [Gei+01]. Here we extend their method in two important ways (see Figure 4.1-A). First, we use an over-complete, multi-scale representation of edges, which is more similar to the output of the primary visual cortex. Second, we use a synthesis model for the edge representation, so that the edges we detect are guaranteed to be sufficient to regenerate the image with a low error. Here we describe each of these procedures, along with the construction of the statistics of edge co-occurrences and the implementation of the classifier.

4.4.1 Linear representation of edges

The first step of our method involves defining the dictionary of templates or filters for detecting edges. We use a log-Gabor representation, which is well suited to represent a wide range of natural images [Fis+07b]. This representation gives a generic model of edges as defined by their shape, orientation, and scale. One advantage of this choice of dictionary is that it allows parameterization of the family of edges; here we set the parameters of the edges to match what has been reported for simple cells’ responses in macaque area V1. In particular, we set the bandwidth of the Fourier representation of the filters to 1 and $\pi/8$ respectively in the log-frequency and polar coordinates to get a family of elongated and thus selective filters (see Fischer et al. [Fis+07a] and Figure 4.1-A for examples of such edges). These values are similar to the values used by Geisler et al. [Gei+01]. Prior to the analysis of each image, we used the spectral whitening filter described by Olshausen and Field [OF97] to provide a good balance of the energy of output coefficients [Fis+07b; PST04a].

A linear convolution model provides a translation-invariant representation. Such invariance can be extended to rotations and scalings by choosing to multiplex these sets of filters at different orientations and spatial scales. Although orthogonal representations are popular for computer vision due to their computational tractability, it is desirable in our context that we have a high over-completeness in the representation to have a detailed measure of the association field. Ideally, the parameters of edges would vary in a continuous fashion, to provide relative translation, rotation, and scale
invariance. We chose to have 8 dyadic levels (that is, by doubling the scale at each level) for the set of 256 × 256 images, with 24 different orientations. Orientations are measured as an angular difference, in the range from 0 to $\pi$ (but not including $\pi$). Tests with a range of different numbers of orientations yielded similar results. Finally, each image is transformed into a pyramid of coefficients. This pyramid consists of approximately $\frac{4}{3} \times 256^2 \approx 8.7 \times 10^4$ pixels multiplexed on 8 scales and 24 orientations, that is, approximately $16.7 \times 10^6$ coefficients, an over-completeness factor of about 256.

This transform is linear and can be performed by a simple convolution repeated for every edge type. Following Fischer et al. [Fis+07a], convolutions were performed in the Fourier (frequency) domain for computational efficiency. The Fourier transform allows for a convenient definition of the edge filter characteristics, and convolution in the spatial domain is equivalent to a simple multiplication in the frequency domain. By multiplying the envelope of the filter and the Fourier transform of the image, one may obtain a filtered spectral image that may be converted to a filtered spatial image using the inverse Fourier transform. We exploited the fact that by omitting the symmetrical lobe of the envelope of the filter in the frequency domain, the output of this procedure gives a complex number whose real part corresponds to the response to the symmetrical part of the edge, while the imaginary part corresponds to the asymmetrical part of the edge (see Fischer et al. [Fis+07a] for more details). More generally, the modulus of this complex number gives the energy response to the edge—as can be compared to the response of complex cells in area V1, while its argument gives the exact phase. This property further expands the richness of the representation.

4.4.2 Sparse coding and validation of the edge extraction method

Obviously, this dictionary of edge filters is over-complete. The linear representation would thus give a inefficient representation of the distribution of edges (and thus of the statistics of edge co-occurrences) as there would exist a priori correlations between coefficients. Therefore, starting from this linear representation, we searched for the most sparse representation. Minimizing the $\ell_0$ pseudo-norm (the number of non-zero coefficients) leads to an expensive combinatorial search with regard to the dimension of the dictionary (it is NP-hard). As proposed first by Perrinet, Samuelides, and Thorpe [PST04b], we may approximate a solution to this problem using a greedy approach.

In general, a greedy approach is applied when finding the best combination is difficult to solve globally, but can be solved progressively, one element at a time. Applied to our problem, the greedy approach corresponds to first choosing the single filter $\Phi_i$ that best fits the image along with a suitable coefficient $a_i$, such that the single source $a_i\Phi_i$ is a good match to the image.
Examining every filter $\Phi_j$, we find the filter $\Phi_i$ with the maximal correlation coefficient, where:

$$i = \arg\max_j \left( \frac{\|I\|}{\|\Phi_j\|} \cdot \frac{\Phi_j}{\|\Phi_j\|} \right),$$

(4.1)

$(\cdot, \cdot)$ represents the inner product, and $\| \cdot \|$ represents the $\ell_2$ (Euclidean) norm. Since filters at a given scale and orientation are generated by a translation, this operation can be efficiently computed using a convolution, but we keep this notation for its generality. The associated coefficient is the scalar projection:

$$a_i = \left\langle I, \frac{\Phi_i}{\|\Phi_i\|^2} \right\rangle$$

(4.2)

Second, knowing this choice, the image can be decomposed as

$$I = a_i \Phi_i + R$$

(4.3)

where $R$ is the residual image. We then repeat this 2-step process on the residual (that is, with $I \leftarrow R$) until some stopping criterion is met. Note also that the norm of the filters has no influence in this algorithm on the choice function nor on the reconstruction error. For simplicity and without loss of generality, we will thereafter set the norm of the filters to 1: $\forall j, \|\Phi_j\| = 1$.

Globally, this procedure gives us a sequential algorithm for reconstructing the signal using the list of sources (filters with coefficients), which greedily optimizes the $\ell_0$ pseudo-norm (i.e., achieves a relatively sparse representation given the stopping criterion). The procedure is known as the Matching Pursuit (MP) algorithm [MZ93], which has been shown to generate good approximations for natural images [Per10b].

We have included two minor improvements over this method: First, we took advantage of the response of the filters as complex numbers. As stated above, the modulus gives a response independent of the phase of the filter, and this value was used to estimate the best match of the residual image with the possible dictionary of filters (Matching step). Then, the phase was extracted as the argument of the corresponding coefficient and used to feed back onto the image in the Pursuit step. This modification allows for a phase-independent detection of edges, and therefore for a richer set of configurations, while preserving the precision of the representation.

Second, we used a smooth Pursuit step. In the original form of the Matching Pursuit algorithm, the projection of the Matching coefficient is fully removed from the image, which allows for the optimal decrease of the energy of the residual and allows for the quickest convergence of the algorithm with respect to the $\ell_0$ pseudo-norm (i.e., it rapidly achieves a sparse reconstruction with low error). However, this efficiency comes at a cost, because the algorithm may result in non-optimal representations due to choosing edges sequentially and not globally. This is often a problem when
edges are aligned (e.g. on a smooth contour), as the different parts will be removed independently, potentially leading to a residual with gaps in the line. Our goal here is not to get the fastest decrease of energy, but rather to provide a good representation of edges along contours. We therefore used a more conservative approach, removing only a fraction (denoted by $\alpha$) of the energy at each pursuit step (for MP, $\alpha = 1$). We found that $\alpha = 0.5$ was a good compromise between rapidity and smoothness. One consequence of using $\alpha < 1$ is that, when removing energy along contours, edges can overlap; even so, the correlation is invariably reduced. Higher and smaller values of $\alpha$ were also tested, and gave classification results similar to those presented here.

In summary, the whole learning algorithm is given by the following nested loops in pseudo-code:

1. draw a signal $I$ from the database; its energy is $E = \|I\|^2$,
2. initialize sparse vector $s$ to zero and linear coefficients $\forall j, a_j = <I, \Phi_j>$,
3. while the residual energy $E = \|I\|^2$ is above a given threshold do:
   (a) select the best match: $i = \text{ArgMax}_j |a_j|$, where $|\cdot|$ denotes the modulus,
   (b) increment the sparse coefficient: $s_i = s_i + \alpha \cdot a_i$,
   (c) update residual image: $I \leftarrow I - \alpha \cdot a_i \cdot \Phi_i$,
   (d) update residual coefficients: $\forall j, a_j \leftarrow a_j - \alpha \cdot a_i <\Phi_i, \Phi_j>$,
4. the final non-zero values of the sparse representation vector $s$, give the list of edges representing the image as the list of couples $(i, s_i)$, where $i$ represents an edge occurrence as represented by its position, orientation and scale.

This class of algorithms gives a generic and efficient representation of edges, as illustrated by the example in Figure 4.1-B. We also verified that the dictionary used here is better adapted to the extraction of edges than Gabors [Fis+07b]. The performance of the algorithm can be measured quantitatively by reconstructing the image from the list of extracted edges. Measuring the ratio of extracted energy in the images, $N = 1024$ edges were on average enough to extract on average 95% of the energy of $256 \times 256$ images on all sets of images. All simulations were performed using Python (version 2.6) with packages NumPy (version 1.6.2) and SciPy (version 0.7.2) [Oli07] on a cluster of Linux computing nodes. Visualization was performed using Matplotlib (version 1.1.0) [Hun07]. All scripts are available upon request to the corresponding author.
4.4.3 Histogram of edge co-occurrences and geometrical symmetries

Note that since we are considering only relative orientations, co-occurrences have several geometrical symmetries: if an occurrence exists for a configuration \((\phi, \theta)\), then it exists also for \((\phi + \pi, \theta + \pi)\) (considering other orientations of the first edge by a rotation of \(\pi\) radian), \((\phi + \pi - \theta, \pi - \theta)\) (swapping both edges) and \((\phi - \theta, -\theta)\) (rotation of \(\pi\) radians). For that reason, it is convenient to define \(\psi = \phi - \theta/2\) as it is symmetric: for a configuration \((\psi, \theta)\), we have also the following symmetries \((\psi + \pi, \theta + \pi)\), \((\psi + \pi, \pi - \theta)\) and \((\psi, -\theta)\). Note that for a pair of edges on a common circle, we have \(\phi = \theta/2\), that is, \(\psi = 0\); this convention also represents circularities more conveniently. Since \(\phi\), \(\psi\) and \(\theta\) are angles defined for instance between \(-\pi\) and \(\pi\), this allows us to consider only a single quadrant (by convention the upper right, that is \(-\pi/2 < \phi \leq \pi/2\), \(-\pi/2 < \psi \leq \pi/2\) and \(0 \leq \theta \leq \pi/2\)), the rest being inferred by the above relations. Furthermore, we observed that there is typically an axial symmetry with respect to the mediator (that is, in any given image set, a configuration \((\psi, \theta)\) is as likely as \((-\psi, \theta)\)), corresponding to mirror versions of images. We used this additional looser type of symmetry only for simplifying visualizations (see Figure 4.2 and 4.3).

To validate the categorization performance, we used the standard SVM library as implemented by Pedregosa et al. [Ped+11]. We used the Jensen–Shannon divergence distance as a metric between histograms [CS02]. The results of the SVM classifier are given as the f1-score to directly compare our method to that of Serre, Oliva, and Poggio [SOP07].
Figure 4.1 (following page): Model for edge extraction and statistics of edge co-occurrences in natural scenes. (A) Architecture of the model. An input image (“Photoreceptors”) is first linearly convolved with a bank of filters at different orientations and scales (“Edge & scale map”), similar to the properties of cells in layer 4 of the primary visual cortex of primates. A second layer (“Sparse edge image”) removes redundant information, so that only the information about the position, orientation and scale of edges remains (see example output in (B)). A third layer (“Edge co-occurrences”) pools the different possible co-occurrences of edge configurations into a map that serves as the input of a classifier (“Decision layer”). (B) An example image with the list of extracted edges overlaid. As in Geisler et al. [Gei+01], edges outside the dashed red circular mask are discarded to avoid artifacts. We controlled the quality of the reconstruction from the edge information such that the residual energy is less than 5% over the whole set of images, a criterion met when identifying 1024 edges per image. (C) The relationship between a pair of edges can be quantified in terms of the difference between their orientations \( \theta \), the ratio of scale \( \sigma \) relative to the reference edge, the distance \( d \) between their centers, and the difference of azimuth (angular location) \( \phi \) of the second edge relative to the reference edge. Additionally, we define \( \psi = \phi - \theta/2 \) so that for co-circular edges, \( \psi = 0 \) (see text). (D) From this representation, it is possible to compute the statistics of edge co-occurrences for each set of images and we show here a replication of the results from Geisler et al. [Gei+01] for natural images (in greenish color) but also for man-made images (in brownish color, slightly shifted vertically for clarity) and animals (in blueish color, slightly shifted vertically). As in Geisler et al. [Gei+01], we can project this function to see the most probable orientation difference knowing any possible position (determined by the distance \( d \) and difference of azimuth \( \phi \)) relative to the reference edge (i.e., \( \arg \max_\theta p(\theta|d, \phi) \)). (E) Additionally, as in Geisler et al. [Gei+01], we can project this function onto similar axes to show the most likely azimuth for each orientation difference and each given distance (i.e., \( \arg \max_\phi p(\phi|d, \theta) \)). The results show that when difference of orientation \( \theta \) is nonzero, it tends to be for co-circular contours (so-called “good completions”) in natural images [Sig+01] and in natural images containing an animal, while sharp edges dominate in man-made images. Here, we have taken advantage of geometrical symmetries of edge co-occurrences to show only the top or the bottom of these configurations, respectively. In both plots, the value of the maximum probability relative to the central reference edge is represented by the transparency of the edge shown and relative to the reference edge in the center of the plot. These results replicate Geisler et al. [Gei+01] and suggest that image categories contain important information about the statistics of edge co-occurrences that could be used as a prior. In addition, while parallel textures dominate for all categories in (D), the pattern in (E) clearly differs qualitatively between databases, though further analysis in subsequent figures will be required to demonstrate these differences quantitatively.
Independent configuration and geometrical variables. An independence analysis shows that in natural images, the statistics of edge co-occurrences can be factorized into independent components \( p(\psi, \theta) \cdot p(\sigma, d) \) (see text). First, to show the shape of \( p(\sigma, d) \), we show in (A) the distribution of scale ratios \( p(\sigma) \) and in (B) the distribution of distances to a reference edge \( p(d) \). Counts are plotted for each dataset in the colors indicated (blue for natural, green for animals and red for laboratory), along with the statistics obtained after shuffling each edge variable (in black). The bar heights allow comparison across categories, while the error bars indicate variation within each category. In (A) there is an overall decrease in probability with increasing difference in scale similar to that in the shuffled case, which is due to the finite number of scales. The distribution is consistent across all databases, with variability comparable within and between databases, and thus the scale differences are not useful for categorizing the class of an image. Similarly, (B) shows that edges are clustered, with other edges significantly more likely to be closer to a given reference edge (with a maximum of about 25% more probable than in the shuffled case at the shortest range). The results for shuffled images show that there is a bias due to the finite size of images. For non-shuffled images, the change in probability with distance is mainly due to a prior preference for a clustering of edges. This distribution is consistent with scenes consisting mostly of small objects, as is well described by the dead-leaves model [Pit10]. Again, the variation within each database in (B) is high relative to the variation between them, and so the distances are also not significant for categorizing the image class. (C) A second component \( p(\theta, \psi) \) represents the distribution of the different geometrical arrangements of edges’ angles, which we call a “chevron map”. We show here the histogram for natural images, illustrating the preference for co-linear edge configurations. For each chevron configuration, deeper and deeper red circles indicate configurations that are more and more likely with respect to a flat prior (all configurations equally likely), with an average maximum of about 4 times more likely, and deeper and deeper blue circles indicate configurations less likely than a flat prior (with a minimum of about 1.5 times less likely). Conveniently, this “chevron map” shows in one graph that natural images have on average a preference for co-linear and parallel angles (the horizontal middle axis), as seen in Figure 4.1-D, along with a slight preference for orthogonal configurations (top and bottom rows). Figure 4.3 will show that the chevron patterns differ reliably between image classes, which will allow them to be used for classification.
Figure 4.3: Chevron maps in man-made and animal images compared to the ‘natural’ set. As for figure 4.2, we show the probability of edge configurations as chevron maps for two databases (man-made, animal). Here we show the probability relative to that of the natural image dataset, as a log ratio (blue: less likely, red: more likely; the opacity is proportional to the absolute log ratio, with deeper colors indicating stronger differences). To further highlight the most reliable differences between these databases and natural images, we selected only the top 20% most relevant features (32 chevrons) for plotting. Selection is based on the p-value corresponding to the discriminability of each class, on an image-per-image basis. The man-made images reliably have more co-linear and parallel elements (red chevrons in (A)), and fewer converging patterns (blue chevrons in (A)). The animal images reliably have more circular continuations (red chevrons in middle of (B)), and more converging patterns (red chevrons at left and right of (B)).
Figure 4.4: **Comparison of the model’s performance with human data on an image by image basis.** Following Serre, Oliva, and Poggio [SOP07], we defined an animalness index (AI) as the fraction of times a human or model classified a given image as having an animal. For each image, we plot the AI for \( n = 24 \) human observers (from [SOP07]) against the AI for the model of Serre, Oliva, and Poggio [SOP07] (red symbols), and then against the AI for the chevron-based SVM-C model (in blue). A model with a perfect match would have datapoints along the dashed diagonal line from bottom left to top right. The symbol shape for each image depends on its ground truth category; target images (actually containing animals) are indicated by a + marker, and distractors (not containing animals) are indicated by a × marker. The model of Serre, Oliva, and Poggio [SOP07] often saturates at confidence 0 or 100\%, while our chevron-based SVM-C model shows a closer match to the human results. These results suggest that the chevron model makes a pattern of errors more consistent with those made by human observers, compared to the hierarchical model.
Table 4.2: **Testing the relative dependence of all histogram variables.** Given the histogram of second order statistics \( p(d, \phi, \theta, \sigma) \) (which we will subsequently denote simply \( p \)), we have computed its entropy and the Kullback-Leibler (KL) divergence with all combinations of the factorization of the variables \( d, \phi, \theta \) and \( \sigma \) (in bits, averaged over all images from all categories). Entropy is the KL divergence of \( p \) with the uniform distribution and serves as an upper bound. The KL measure is positive and only the full statistics can achieve \( KL = 0 \) (Gibbs’ inequality), but the next four factorizations cause very little error, and are thus good candidates. For these factorizations, we may separate the characteristic angles (\( \phi \) and \( \theta \), individually or together) from distance related statistics (\( d \) and \( \sigma \)) which are characteristic of the configuration of the view. The remaining factorizations are much more different from the full distribution, and thus will not be considered further.

<table>
<thead>
<tr>
<th>Factorization</th>
<th>KL divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
<tr>
<td>( KL(p</td>
<td></td>
</tr>
</tbody>
</table>

\( Entropy(p) \approx 0.724 \)
Chapter 5

Anisotropic connectivity implements motion-based prediction in a spiking neural network

Abstract

The hypothesis of predictive coding, that is, that the brain explicitly infers upcoming sensory input to establish a coherent representation of the world, is becoming generally accepted. However, it is not clear on which level spiking neural networks may implement such predictive coding and which function the connectivity may have. We present a network model of conductance-based integrate-and-fire neurons inspired by the architecture of retinotopic cortical areas and which assumes that the basis for predictive coding is implemented through network connectivity, namely in the connection delays and in the connection selectivity, based on the tuning properties of source and target cells. We show that the applied connection pattern leads to motion-based prediction in an experiment tracking a moving dot. In contrast to our proposed model, a network with random or isotropic connectivity fails to predict the path when the moving dot disappears. Furthermore, we show that a simple linear decoding approach is sufficient to transform neuronal spiking activity into a probabilistic estimate to read out the target trajectory.
5.1 Introduction

5.1.1 Problem statement: Codage prédictif dans les transformations visuo-motrices

In a dynamical world, prediction is a highly relevant evolutionary advantage. This is crucial in sensory systems, as the raw data that is processed is most often noisy, and possibly ambiguous or distorted. Take for example the task performed by the primate visual system of tracking the trajectory of a moving object and accurately moving the eyes in order to stabilize the image on the retina. The image of the object may be blurred, or the measure of its velocity may depend on its geometry instead of its trajectory. Another problem occurs when the object is occluded, or simply when the observer blinks. It is an advantage to be able to predict the position and speed of the object at the end of this blanking period. This problem is classically referred to as motion extrapolation (see figure 5.1). While predictive coding mechanisms may have different aspects and occur at different levels ranging from the retina to higher level areas [GM10], we will focus on this particular phenomenon as prototypical example.

Particularly in primates, object motion information is extracted along a cascade of feed-forward cortical areas, where primary visual area (V1) extracts local motion information that is integrated in extra-striate middle temporal (MT) and medial superior temporal (MST) areas [NWK88]. MT and MST process visual motion and oculomotor signals driving pursuit (see [Ilg97] for a review) and are therefore key elements in motion extrapolation. Specifically, we will focus on the dynamics of neural activity during the period without informative sensory input (to which we will refer as the blank) and just after its reappearance. Indeed, the capacity of the dynamics to transform such fragmented input into a correct, continuous representation is a major pressure on the evolution of the visual system [GM10]. It was shown in the monkey visual system that neural activity was mostly absent during the blank in lower areas of the visual hierarchy while it was maintained in some higher level areas [AM95]. More precisely, neural activity in MT is driven by the motion of the dot and quickly devolves to spontaneous activity during a blank, while activity in its efferent area MST is maintained to the level of neural activity expected if the dot was not blanked when there is no retinal image motion. This can happen during a transient image occlusion [NWK88] or while tracking an imaginary target covering the visual field outside of the receptive field currently being recorded [IT03]. Similar sustained activity during target occlusion has been found in primate posterior parietal cortex, and it is linked to image motion prior to target disappearance [AM95], that is, by a predictive signal.

Motion extrapolation is also seen in lower level neuronal structures, such as the retina [Ber+99], and calls for a more generic computational
framework. However, direct evidence for such neural mechanisms is still lacking. Before proposing a solution using a connectivity pattern based on motion-based prediction, we will first review some existing experimental and theoretical evidence. Along this study, our aim is to provide a basis for future applications of neuromorphic hardware [Bru+11; Sch+10].

5.1.2 Neuro-physiological correlates of prediction for motion extrapolation

At the neural level, it seems that the topography of neural representation is an essential constraint to prediction. Indeed, it is more efficient that populations of neurons that represent similar parameters should be close. This is due to the cost of wiring neurons (length and volume of axon and dendrites) [CSS02] but also due to the limited speed of information propagation in neural wires. Such aspect is particularly acute on the surface of the cerebral cortex and this hypothesis has been an efficient construct to understand the organization of visual areas [Mii+05]. This is also implemented in other cerebral structures and species such as the convergence of inputs from place cells in the hippocampus of rats that code for path integration of body position in an environment [McN+06]. Physiological evidence shows that similar mechanisms are present in the deep superior colliculus of primates allowing for the integration of the belief on the position of a visual target in visual space for the guidance of saccadic or smooth eye movements [Kra04]. Here, we will focus on low-level visual areas based on the neurophysiology of the macaque brain (V1, MT and MST), but we will keep a rather generic formulation to explore the functional role of some key parameters.

Neurons in such areas receive connections from neighboring neurons in the same cortical area (local connectivity) but also respectively by feed-forward or feed-back connections from lower or higher areas. Focusing on area MT, early physiological studies in macaque monkey identified this area as a specialized module for visual motion processing [AKL73; DZ71]. This involves extracting speed and direction of the moving object. MT neurons respond selectively to visual motion and are tuned for local speed and direction of luminance features moving in their receptive fields [MV83]. Concerning motion integration, Pack and colleagues [PB01] have shown that the temporal dynamics of behavior can correspond with the firing rates of MT neurons. They showed that neuronal responses quickly progress from local to global motion direction in about 100 ms, suggesting that such integrative mechanisms are dynamical and progressive. These results pinpoint the key role of MT neurons in local motion analysis and global motion integration. Area MT and MST receive feed-back connections that may modulate the activity of their neurons [SB95]. However, these connections (mostly myelinated) introduce constant delays and are mostly related to higher level contextual modulations. Provided that motion extrapolation is implemented in one single cortical area, a finely structured
set of diffusive mechanisms would be required. A potential candidate is naturally the dense network of lateral interactions found in sub-cortical and cortical structures involved in sensory processing and sensorimotor control. Of particular relevance is the role of the connectivity pattern in the emergence of a solution to this problem. In this paper we will focus on a smaller spatio-temporal scale and study the role of lateral, intra areal (mostly unmyelinated) connections.

A possible correlate of prediction may lay in the traveling waves of neural activity that may be observed on the cortical surface. Covey was the first to show a precisely tuned synaptic integration field [Bri+99] (see [SNC12] for a review). Theoretical studies suggest that for such waves to exist, there should exist some specific anisotropy connectivity pattern [BC98]. It is established that the speed of propagation of activity along these mostly unmyelinated connections is of the order $0.1 \text{ mm ms}^{-1}$ to $0.4 \text{ mm ms}^{-1}$ but there is an ongoing debate on their selectivity. In the primary visual cortex, a set of patchy connections in the long-range horizontal connections found in superficial layers of cortex [Bos+97] that preferentially connect columns with similar orientation preference has been observed in ferrets. This is consistent with the fact that columnar interactions determine horizontal propagation of recurrent network activity in neocortex [WC12]. It has also been observed that activity spreads in macaque anisotropically, but for all orientation columns [Cha+11]. Anisotropies in the connection pattern necessarily leads to a wide range of traveling wave parameters (speed, direction) and introducing inhomogeneities can in addition lead to more complex wave profiles and possibly even wave propagation failure [Bre01]. However, the function of these traveling waves, and therefore the underlying structure of the intracortical connectivity, is mostly unknown.

### 5.1.3 Existing neuromorphic models for prediction

There have been numerous attempts at modeling generic predictive neural mechanisms. Here, we review some prototypical examples at different modeling levels, from a more abstract level to a neuromorphic implementation.

Following the idea of the Kalman filter as an adaptive predictive filter and extending the work of Montagnini et al. [Mon+07b], Bogadhi et al. [BMM13] proposed a hierarchical recurrent Bayesian framework to understand the behavioral response to motion extrapolation as observed in smooth pursuit eye movements. Indeed, probabilistic inference has been successful in explaining motion perception to a variety of stimuli [WSA02] under the hypothesis that sensory areas use predictive coding as a generic neural computation [RB99]. They are somewhat similar to engineering models proposed earlier [NS95] but allow for a more explicit formulation of the underlying hypothesis. Such a framework accommodates uncertainty in the motion information of the measurement likelihoods [HSS11; SS06; WSA02]. Representing uncertainty
in the measurements and prior expectation gives a simple, yet powerful framework to investigate the predictive behavior of the system, and offers the possibility to optimally adapt to changes in the measurements. The approach from Bogadhi and colleagues [BMM13] allows for a mix of prediction and measurement based on their reliability measured from their respective variances. The combined estimate is used to drive the pursuit response. The hierarchical framework allows investigation of the adaptive behavior of pursuit as well as the role of prediction on motion integration as observed in pursuit responses. Such Bayesian models give a generic account of the motion extrapolation mechanism but do not provide a neural implementation.

A direct translation could in theory be performed by a probabilistic population code approach [Bec+08]. This requires that neural responses represent probability distributions and assume “Poisson-like” spike response variability. Under that hypothesis, one could derive from a Bayesian model the architecture of a network of spiking neurons. Another approach is to use a global and generic functional cost for the problem (such as the free-energy of a system designed to track a dot) and derive the optimal system. Such endeavors allow one to propose a hierarchical neural architecture [Fri09], which predicts behavioral results under visual occlusion for control and schizophrenic patients [APF12]. Such models are in essence similar to other modeling approaches where neural activity is represented by average firing rate on a cortical map (forming a so-called neural field). Such models were successful in accounting for a large range of classical and non-classical receptive field properties of V1 including orientation tuning, spatial and temporal frequency tuning, cross-orientation suppression, surround suppression, and facilitation and inhibition by flankers and textured surrounds [Spr10]. Similar models were applied to problems specific to motion detection and a link can be drawn between such solutions and classical solutions drawn in computer vision [TMK10]. However, these models do not take advantage of the specificity of computing with spiking neurons, that is the dual property of being able to integrate information and detect synchrony in the input.

Some models that propose solutions related to motion extrapolation using neuronal networks (spiking and non-spiking). A recent model of spiking units [LC08] explains the phenomenon of the flash-lag effect [Nij08] by a motion-extrapolation mechanism provided by facilitating synapses, but acts on the single cell level only. Baldo et al. [BC05] present a feed-forward network of leaky integrate-and-fire (LIF) neurons performing prediction, but that does not account for the role of recurrent connectivity abundant in cortical networks. In that aspect, [LW08] present a more realistic recurrent network but focuses on a binary decision task, whereas we focus on a generic solution for the problem by studying prediction performance for a spectrum of possible directions. Recently, Jancke [JE10] used a recurrent neural field model to explain visual illusions like the Fröhlich, the flash-lag, and the representational momentum effect. Our approach is similar to theirs in the
sense that the mechanism for motion-extrapolation can be seen in spreading activation to surrounding neuronal populations, but differs fundamentally in the way that connections are set up, as connection selectivity for directional tuning is not considered in their model.

As an intermediate observation, we see that though there exist a wide spectrum of models, a common feature is that these models use diffusive mechanisms implemented by the connectivity to propagate predictive information (probabilities, population activity, spikes) from a local to a global scale. The richness of behaviors is then mostly obtained by using different types of neurons (for instance by varying their polarity — excitatory or inhibitory, or the time constant of the synapses), which implements complex non-linear mechanisms such as gain control. This may be sufficient to account for motion extrapolation. However, it should also be highlighted that all these models assume a prediction in all directions and therefore that the connectivity is \textit{a priori} isotropic. We challenge this assumption by introducing anisotropy in the connectivity as a key mechanism transporting motion information in a coherent manner.

5.1.4 Our approach: an anisotropic connectivity pattern implementing motion-based prediction

At the behavioral level, Yuille et al. \cite{YG89} have shown that motion integration in humans is highly dependent on the smoothness of the trajectory of the stimulus. Humans can detect a target dot moving in a smooth trajectory embedded in randomly moving dots, while the target dot is not distinguishable from noise in each frame separately. Introducing a preference for smooth trajectories, the activity from local motion detectors are made more coherent in space and time and this globally lowers the threshold for detecting stimuli moving in smooth versus segmented trajectories. In particular, during a transient blanking, it is most likely that such processes (along with the knowledge that the sensory input was indeed blanked and not definitively removed) underlie motion extrapolation. For instance, when a moving target disappears, smooth pursuit eye movements continue at the same velocity during the initial period of occlusion \cite{BB03}. Therefore, it seems that neural computations take advantage of the information about motion, but it is yet not clear how this can be done efficiently in a network of spiking neurons.

At an abstract level, a preference for temporal coherency of motion can be defined in a probabilistic framework. This was formulated theoretically \cite{BYG00} proposed as a neural field implementation including local to short-range connectivity but lacked the precision needed to efficiently represent realistic input sequences. In our earlier work \cite{PM12b}, we implemented an efficient prior for smooth trajectories to investigate different aspects of spatio-temporal motion integration. Particularly, this model focused on the aperture problem and proposed that local, diffusive predictive coding
is sufficient to infer global motion from local, ambiguous signals. The aperture problem is a challenging problem to study integration of local motion information [PB01]. The model proposed that instead of specific mechanisms such as line-ending detectors, the gradual spatio-temporal integration of motion relies on prediction based on the current knowledge of motion in terms of its velocity and position (motion-based prediction). Compared to previous models the main difference of this implementation is that, it is possible to predict that information about motion velocity at a known position will be transported in the direction given by the velocity.

Indeed in motion-based prediction, the retinotopic position of the velocity of motion is an essential piece of information that allows routing information and allow implementation of predictive coding on smooth trajectories. By including explicitly the dependence of local motion signals between neighboring times and positions knowing the current speed along a smooth trajectory, incoherent features should get canceled out, while coherent information should get progressively enhanced. As such, this context-dependent, anisotropic diffusion in the probabilistic representation of motion also results in the formation of a tracking behavior favoring temporally coherent features. Such model was recently extended to account for motion extrapolation [KMP13] and has been able to replicate some behavioral results from [BMM13]. Our goal here is to show that the idea of motion-based prediction (as described in [PM12b]) can be implemented in a generic network of spiking neurons through anisotropic connectivity and that this is sufficient to solve a motion extrapolation task. The novelty compared to previous studies is the transition from an abstract, probabilistic framework to a spiking neural network and the link between anisotropic connectivity to motion-extrapolation, a task of functional relevance. Of course, we will not exclude that other complementary solutions may exist, but we will argue that it constitutes one of the simplest solutions for a network of spiking neurons. For that purpose, we will use a classical implementation of recurrent networks using conductance-based integrate-and-fire neurons with three prototypical connectivities: random, isotropic or anisotropic. While the consequence of non-homogeneous connectivities has been somewhat explored [VP12], it is —up to our knowledge— the first study of the functional consequence of anisotropic connections in a large-scale neural network.

5.1.5 Objectives and outline

This paper has been prepared in following order: First, we develop a network of spiking neurons with the connectivity directly drawn from the probabilistic modeling framework proposed for the solution to the aperture problem [PM12b] and that was extended to the motion extrapolation problem [KMP13] (see chapter 3). We will include in Section 5.2.1 details
on structure and implementation of the model but also details from the experimental and numerical aspects.

Then, we report results in Section 5.3 from simulations where we studied the network response to a disappearing moving dot under three different connectivities: random, isotropic or anisotropic.

Finally in the discussion (Section 5.4), we interpret these results in the light of current knowledge on probabilistic inference and dynamical systems and we will discuss the limitations of the current study along with suggestions for future work.

5.2 Methods

5.2.1 Neuron parameters

Simulations were performed with PyNN [Dav+08] as interface to the NEST simulator [GD07] on a Cray XE6 system using 96 cores. For analysis we used python modules numpy [Oli07], scipy [J+01] and visualization was performed using matplotlib [Hun07]. Neurons were simulated as LIF neurons with conductance based synapses. The membrane potential \( V \) of a neuron with index \( k \) obeys the following equation:

\[
C_m \frac{dV_k}{dt} = g_l (E_l - V_k(t)) + \sum_j [g_{j,k,E}(t)(E_E - V_k(t)) + g_{j,k,I}(t)(E_I - V_k(t))]
\]

where \( j \) is the index of the sources, \( g_{j,k}(t) = w_{j,k} \cdot \exp(-\frac{t-t_{\text{spike}}}{\tau_p}) \) is the synaptic conductance time course with \( p \in \{E, I\} \): \( \tau_p \) are the synaptic time constants, and \( E_p \) is the reversal potential for excitatory (\( p = E \)) and inhibitory (\( p = I \)) synapses respectively, \( g_l \) is the constant leakage conductance, and \( E_l \) the leakage or resting potential. When the membrane potential \( V \) is above the threshold voltage \( V_{\text{thresh}} = -50 \text{mV} \) a spike is emitted and \( V \) is set to \( V_{\text{reset}} = -70 \text{mV} \) for a refractory period of \( \tau_{\text{refrac}} = 1 \text{ ms} \). Table 5.1 lists the parameters used for both excitatory and inhibitory neurons. The cell and synapse parameters have been chosen to be in a similar range as seen in experimental studies (see table 3 in the study by [Rau+03]) to allow for comparison between future modeling and experimental studies. The model is in principle not dependent on the cell parameters and different parameters would not change the fundamental outcome, but retuning of other parameters like connection strengths would be necessary. The initial values of the membrane potentials are drawn from a normal distribution around \( V_{\text{init}} = -65 \text{ mV} \) with a standard deviation of 10 mV.
### Table 5.1: Neuron parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>$C_m$</th>
<th>$g_l$</th>
<th>$\tau_m$</th>
<th>$E_I$</th>
<th>$E_E$</th>
<th>$\tau_E$</th>
<th>$\tau_I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>1</td>
<td>0.1</td>
<td>10</td>
<td>-70</td>
<td>0</td>
<td>-70</td>
<td>5</td>
</tr>
<tr>
<td>Unit</td>
<td>nF</td>
<td>$\mu$S</td>
<td>ms</td>
<td>mV</td>
<td>mV</td>
<td>ms</td>
<td>ms</td>
</tr>
</tbody>
</table>

#### 5.2.2 Tuning properties

The model is inspired by retinotopic cortical areas like areas V1 or V5/MT of primates. In our model, each neuron has four tuning properties: $(x_i, y_i, u_i, v_i)$ parameterizing the center of the spatial receptive field of neuron $i$ at position $\vec{x}_i = (x_i, y_i)$ and its preferred direction $\vec{v}_i = (u_i, v_i)$. The width of this receptive field defines the tuning selectivity of neurons and is parameterized by $\beta_X$ and $\beta_V$, respectively for space ($x$ and $y$) and velocity ($u$ and $v$). The spatial receptive fields are arranged in a hexagonal grid to optimally cover the input space which is set to span a $1 \times 1$ area in arbitrary units. As we will model a network size of approximately $10^4$ neurons, this will in practice correspond to a spatial scale of the order of millimeters. Velocities should therefore be in the range of ms$^{-1}$.

In order to have receptive fields for all possible directions (up to a certain maximum velocity of approximately $|\vec{v}_{\text{max}}| = 4.0$ ms$^{-1}$) at all positions, the midpoint of each of the 100 hexagonal grid cells contains neurons with ten different preferred velocities for ten different angles, hence 100 different preferred directions per hexagonal grid cell. The lengths of preferred directions are distributed according to a distribution favoring low velocities [WSA02] with a logarithmic scale for the speed according to Weber’s law [SS06]. In order to avoid boundary effects, both spatial dimensions are closed and continuous. This leads to a horn torus as input space, i.e. if a stimulus leaves the $1 \times 1$ space it reappears on the opposite side (so-called “pac-man topology”). This topology holds also for the network connectivity, e.g. connections reaching beyond the virtual border at $x_{\text{target}} = 1$ will be wrapped around. After all tuning properties are set, they get dispersed to account for natural variability [PR11].

#### 5.2.3 Input Stimulus

A classical way of studying motion extrapolation is by presenting a moving target that travels behind an occluder for a short period of time. A seminal study used timing estimation by asking participants to make a button press response at the time they judge the occluded target to have reached a particular point [Ros75]. Motion extrapolation can be carried out for lateral motion with the target moving across the fronto-parallel plane, or for approach motion, when the object moves towards the observer [DeL04]. Herein, we
investigate visual, lateral motion extrapolation only. For simplicity, we study
the network’s response to a moving dot stimulus and the network’s ability
to predict the trajectory of the dot when it disappears behind an obstacle
which leads to a blank gap in the input signal.

From the definition of the tuning properties of a neuron \( i \), we may model
the response to a moving dot as an inhomogeneous Poisson process with
a parametrically defined envelope. Indeed, we will use the following input
stimulus \( L_i(t) \) as the envelope for a Poisson process with a maximum
of 5 kHz (when \( L_i(t) \) reaches 1) and a time step of 0.1 ms:

\[
L_i(t) = \exp(-\frac{\|\vec{x}_{stim}(t) - \vec{x}_i\|^2}{2\beta_X^2} - \frac{\|\vec{v}_{stim} - \vec{v}_i\|^2}{2\beta_V^2})
\] (5.2)

where \( \vec{x}_i \) is the neuron’s receptive field central position, \( \vec{v}_i \) the neuron’s
preferred direction, \( \vec{x}_{stim}(t), (\vec{v}_{stim}) \) is the position (direction) of the moving
dot (see figure 5.1). As the trajectory of the dot is rectilinear and constant,
we have

\[
\vec{x}_{stim}(t) = \vec{x}_{stim}(0) + \vec{v}_{stim} \cdot t
\] (5.3)

The resulting inhomogeneous Poisson spike train is connected to the re-
spective neurons via one excitatory synapse of strength \( w_{input} = 5 \text{nS} \). This
formalization allows to study for the different roles of theses parameters.
In particular, the tuning width may play an important role as it is known
that in low level visual areas (such as the retina), receptive fields are small
(position is accurate, motion is imprecise) while in higher level areas motion
is more finely represented, while position is less precise (as the receptive
fields’ size increase). In the rest, all neurons have the same tuning width
defined by \( \beta_X = 0.15 \) and \( \beta_V = 0.15 \text{s}^{-1} \). The \( \beta_X, \beta_V \) values have been set
so that a reasonable part of the network receives sufficient input from the
moving dot stimulus. Increasing \( \beta_X, \beta_V \) would make the dot appear broader,
whereas smaller \( \beta_X, \beta_V \) would make a smaller fraction of the network respond
to the stimulus. Changes in the \( \beta_X, \beta_V \) parameters would not change the
working concept of the model, but would require a retuning of connectivity
parameters like number of connections and connection strengths.

For simplicity, we studied only networks in which excitatory neurons
receive input because inhibitory neurons primarily provide a normalization
mechanism in our model, even though this might not reflect real cortical
circuits [Fre+03]. All neurons receive additional noise in form of Poisson spike
trains with a rate of \( f_{noise} = 2 \text{ kHz} \) injected via excitatory and inhibitory
synapses with a weight of \( w_{noise} = 4 \text{nS} \) to simulate the input from external
networks. For all simulations of this paper, the network was stimulated with
a dot moving at a speed of \( \vec{v}_{stim} = 0.5 \text{s}^{-1} \) from left to right (and an initial
position defined by \( \vec{x}_{stim}(t = 0) = (.1,.5) \)). Crucially, during the blank phase,
the stimulus vector in the network was permuted randomly at each time
step among the whole excitatory population, so that the selectivity of the
input was completely lost while keeping a similar input average frequency as compared to phases when stimulus is active. As the input vector is shuffled the network does not receive a coherent continuous input signal. During the blank phase, cells that are well-tuned to the stimulus receive input spike trains with a larger inter-spike-interval leading to a decrease in the effective input. This is due to the integration of post-synaptic potentials on the membrane in the context of LiF neurons. Compared to an empty input vector, this input vector shuffling during the blank phase elevates the mean membrane potential of the population slightly which can help the network to fill the blank phases with meaningful input.

5.2.4 Network connectivity

The network study consists of an excitatory population with \( N_{\text{exc}} = 13000 \) neurons and an inhibitory population with \( N_{\text{inh}} = 2520 \) neurons (that is, with a ratio of 16.2% inhibitory cells over the whole population). Both populations are mutually and recurrently connected in one of the following ways which will be explained in the next sections: randomly, isotropically or anisotropically.

Random and isotropic connectivities

Connections within and between populations can be set up in an isotropic manner that does not depend on the source or target neuron’s tuning properties. When neurons are connected in this way, connection probabilities are computed according to:

\[
p_{ij} = p_{\text{max}} \cdot \exp\left(-\frac{d_{ij}^2}{2 \cdot \sigma_X^2}\right)
\]  

(5.4)

where \( p_{\text{max}} \) is a normalizing factor and \( d_{ij} = \|\vec{x}_i - \vec{x}_j\| \) represents the distance (in visual space) between both neurons. \( p_{\text{max}} \) is set so that the total number of connections between two populations drawn isotropically is equal to an overall probability of \( p_{k,l} \), \((k,l) \in \{E,I\}\). The connection probabilities utilized are: \( p_{EE} = 0.5\% \), \( p_{EI} = 2\% \), \( p_{IE} = 2\% \), \( p_{II} = 1\% \).

Weights are drawn from a normal distribution with mean \( \mu_{\text{iso}}^{w} \) and standard deviation \( \sigma_{\text{iso}}^{w} = 0.2 \cdot \mu_{\text{iso}}^{w} \). The value of \( \mu_{\text{iso}}^{w} \) is set so that the expected sum of incoming weights equals a certain target value \( w_{kl} \) specific to the type of the source and target population \((k,l \in \{E,I\})\): \( w_{EE} = 0.3 \) \( \mu \text{S} \), \( w_{EI} = 1.8 \) \( \mu \text{S} \), \( w_{IE} = 0.8 \) \( \mu \text{S} \), \( w_{II} = 0.15 \) \( \mu \text{S} \) (if not stated differently).

Delays are drawn from a normal distribution with a mean value \( \mu_{\text{iso}}^{\delta} = 3 \) ms and a standard deviation of \( \sigma_{\text{iso}}^{\delta} = 1 \) ms. Self-connections have been discarded. A completely random connectivity may be then achieved by setting \( \sigma_X \) to a sufficiently big value (relative to the scale of the spatial period). This results in a flat, uniform probability of connection over the whole population.
Anisotropic, motion-based prediction connectivity

Inspired by motion-based prediction [PM12b], we may define a connectivity by wiring neurons that are linked by a smooth trajectory with a higher probability. Connectivity will then be specifically anisotropic as it provides a mechanism for motion-based prediction by diffusing motion information across the network in a forward, asymmetric manner. Specifically, we will take advantage of the latency that exists between neurons in the same cortical area and use that parameter to connect cells matching a smooth trajectory. The motivation underlying this formula is based on the idea that smooth trajectories are more likely seen in natural scenes and are promoted by the network connectivity. If the target position is situated at the position where the source neuron predicts the stimulus to be in a certain time $\tau_{ij}$ and if the target neuron predicts the stimulus to move in a similar direction $\vec{v}_j$ as the preferred direction of the source neuron $\vec{v}_i$, the source neuron connects with a high probability to the target neuron.

As a consequence, the connection probability is computed from the tuning properties of the source neuron $i$ and target neuron $j$ according to the sampling of the prior defined in [PM12b]:

$$p_{i,j} = p_{\text{max}} \cdot \exp\left(-\frac{\|\vec{x}_{i,j}^* - \vec{x}_j\|^2}{2 \cdot \sigma_X^2}\right) \cdot \exp\left(-\frac{\|\vec{v}_i - \vec{v}_j\|^2}{2 \cdot \sigma_V^2}\right) \quad (5.5)$$

$$\vec{x}_{i,j}^* = \vec{x}_i + \vec{v}_i \cdot \tau_{ij} \quad (5.6)$$

$$\tau_{ij} = \frac{\|\vec{x}_i - \vec{x}_j\|}{\|\vec{v}_i\|} \quad (5.7)$$

In this formulation, $\vec{x}_{i,j}^*$ is the position predicted for a motion that would leave the source neuron’s receptive field (therefore from position $\vec{x}_i$ and with velocity $\vec{v}_i$) after a latency $\tau_{ij}$. Then, parameter $\tau_{ij}$ corresponds to the expected latency knowing the respective position and velocity of source and target neurons. In equation (5.5), the parameters $\sigma_X$ and $\sigma_V$ determine the strength of the tuning properties of motion-based prediction. Unless stated otherwise, we will use $\sigma_X = 0.1$ and $\sigma_V = 0.1 \text{s}^{-1}$ (see figure 5.2). Note that the precision of prediction in the velocity domain (that is $\sigma_V$) determines a scaling factor for the degree of anisotropy: The lower $\sigma_V$ is, the more the outgoing connections of a neuron are aligned with the preferred direction of the source neuron. Note also that only $\sigma_V$ includes the predictive prior on velocity and that we may retrieve an isotropic connectivity by setting $\sigma_V$ to a sufficiently high value.

The probabilities are then sorted and each target neuron receives input from 0.5% of the source neurons that have the highest connection probability. Those 0.5% highest probabilities are converted to connection weights so that the sum of incoming weights per neuron equals a certain target value $w_{kl}$ specific to the type of the source and target population ($k, l \in \{E, I\}$):
$w_{EE} = 0.20 \, \mu S$ (for motion-based connectivity and $w_{EE} = 0.25 \, \mu S$ for direction-based connectivity), $w_{EI} = 1.8 \, \mu S$, $w_{IE} = 0.8 \, \mu S$, $w_{II} = 0.15 \, \mu S$ (these values are only for the example networks and might differ depending on the exact implementation and require an adjustment for different network sizes). This means the connectivity becomes deterministic (based on the tuning properties of the source and target cell) and the term probability refers only to the overall selection of source cells in the network.

**Anisotropic, direction-based prediction connectivity**

However, if we use the previous equation to connect cells (equation (5.5)), and scale our network realistically, it appears that latencies depend on the velocity coded by the cells, and in turn, this leads to unrealistically high delay values with the range of velocities we used. As a consequence, we defined another way of setting up the connectivity which only take into account the angle between source and target cells and the angle between the directions coded for by the source and target cells. It is therefore independent of the preferred speed (modulus of velocity) of the neurons and on the latency used to connect the cells.

We use von Mises probability distribution functions to define the tuning in the range of all directions:

$$p_{i,j} = p_{\text{max}} \cdot \exp\left(\frac{\cos(\vec{x}_j - \vec{x}_i, \vec{v}_i)}{\sigma_X^2}\right) \cdot \exp\left(\frac{\cos(\vec{v}_i, \vec{v}_j)}{\sigma_V^2}\right) \quad (5.8)$$

Where the first term guarantees that information spreads in the direction that is preferred by the source cell (and where $\sigma_X$ gives approximately the width of tuning in radians). The second term ensures that information is passed only to cells that code for motion that moves in a similar direction as preferred by the source cell (and where similarly $\sigma_V$ gives approximately the tuning width in radians). Note that in position-velocity space, the probability of connection is maximal in the direction given by the preferred velocity of the source cell and centered on the position of that cell’s receptive field. The density therefore defines a cone around this half-line, defined by widths $\sigma_X$ and $\sigma_V$ (see middle panel in figure 5.2). Note that this formulation may be derived from the formulation of motion-based prediction by lowering the strength of prediction on the radial component of velocity. As such, this connection probability gives a similar mechanism for promoting smooth trajectory, and provides the diffusion of motion information in the direction detected by the network. A comparison of these two network connectivities is visualized in figure 5.2.

Whereas encoding and decoding of direction information is now largely understood in various neuronal systems, how the human brain accurately represents speed information remains largely unknown. Speed tuned neurons
have been identified in several early cortical visual areas in monkeys. However, how such speed tuning emerges is not yet understood. A working hypothesis is that speed tuned neurons nonlinearly combine motion information extracted at different spatial and temporal scales, taking advantage of the statistical spatiotemporal properties of natural scenes. Furthermore, the population code underlying perceived speed is not yet elucidated and therefore we are still far from understanding how speed information is decoded to drive and control motor responses or perceptual judgments. As a consequence, such a connectivity profile will serve as a further control to test if restraining the prediction to direction is sufficient to solve the motion extrapolation problem.

5.2.5 Choice of parameters

The number of receptive fields has been set so that the four-dimensional space of tuning properties is covered with a reasonable density of cells. Decreasing the number of receptive fields would decrease the number of cells in the network and would impede the diffusion of information between cells. This is because the weight of connections is sensitive to the distribution of cells in the tuning property space, and a over-sparsely populated tuning property space can lead to unwanted effects for activity spread in the network. The parameters describing receptive field sizes, $\beta_X$ and $\beta_V$, determine the distribution of the input signal in the network. They have been chosen so that a small part of the network receives a sufficient amount of excitation that brings this small part above the spike threshold and initiates the spread of activity, and by that, the diffusion of motion-information in the network. Increasing $\beta_{X,V}$ would make the stimulus appear fuzzier making the extrapolation task more difficult. A decrease of $\beta_{X,V}$ would make the stimulus appear sharper. But it would not necessarily make the task easier since the source of activity would be smaller and the seed for the diffusion of information could possibly be too small to propagate through the network, depending on the network connectivity parameters. The parameters determining the network connectivity $p_{k,l}$ and $w_{k,l}$, $(k,l) \in \{E,I\}$ were chosen to be in a range comparable to physiological values for large networks. Especially the connection weights needed to be fine tuned to solve the motion-extrapolation task. Redistributing the tuning properties could easily lead to instabilities, i.e. that the trajectory could not be extrapolated, and too high weights could lead to an explosion of activity in the network.

5.2.6 Prediction readout

A crucial issue when trying to map a Bayesian inference algorithm to a network of spiking neurons is to understand how probability can be expressed in terms of neural activity. Herein, we applied a simple vector averaging
method to infer the prediction about stimulus position and direction from the activity of the excitatory population. Indeed such decoding scheme may be justified as a simple implementation of probabilistic codes as done by [Bec+08]. Their approach requires several assumptions which are not guaranteed in our model: First of all, neurons are assumed to have Poisson-like spiking statistics, which is obviously not true in our model since activity is strongly driven by the stimulus and hence neuronal activity is not Poisson-like (see 5.3). Secondly, they assume that network activity is uncorrelated on timescales of 50 ms, which is likewise not realistic for our model. Furthermore, their approach works on probability distributions gained over several trials, which could principally be done with our model, but it is computationally more expensive than the single-trial vector-averaging method described above. However, this provided a decoding approach which seemed to robustly represent the activity in the network.

In particular, we used a similar formulation as the decoding framework proposed for neurons in area MT [JM06]. Indeed, the definition of our model fits well to their implementation. In both models, the activity of sensory neurons is pooled in a simple additive feed-forward architecture. In contrast to their model, we extend the application beyond the angle of motion and apply the readout framework to position and direction. More precisely, the tuning properties are in the exponential family and tile uniformly the position-velocity space. Thanks to the definition of the tuning selectivity of the neurons in the network, the position and velocity corresponding to the Maximum Likelihood estimation corresponds to the average over all neurons of each central tuning parameter weighted by the activity of the neuron (independently of $\beta_X$ and $\beta_V$ as they are uniform for all neurons).

To define a continuous activity at each time bin a weight $p_i(t)$ is defined for each excitatory neuron $i$ based on the number of output spikes fired during a time bin $t$:

$$p_i = \frac{n_i(t)}{\sum_i n_i(t)}$$  \hspace{1cm} (5.9)

where $n_i(t)$ is the number of spikes fired by neuron $i$. The time bin size was set to 50 ms, but it could be chosen differently without qualitative changes.

Such decoding schemes are classically implemented on unbounded variables. However, we defined space on a torus in order to avoid edge effects. Hence, the network average must be computed for circular quantities [MJ09]: The idea behind equation 5.10 is that in order to compute the mean of a circular quantity, the position and direction first need to be transformed into an angle, which is then projected to the 2D unit circle where the arithmetic mean is computed. After that, the angle that the mean position forms is transformed back from an angle to space. For positions, this takes the form:
where \( x_i \) is the center of the spatial receptive field of neuron \( i \) (the same formula is applied to compute \( y_{\text{pred}}(t) \)). The subtraction of \( \pi t \) in the \( \sin \) and \( \cos \) functions is necessary to map the interval of position which is between 0 and 1 to the interval of \((-\pi, \pi)\) required for the projection of position on the unit circle.

Similarly, for reading our the direction of the stimulus predicted by the network:

\[
x_{\text{pred}}^n(t) = \frac{1}{2\pi} \arctan2\left( \sum_i p_i(t) \sin(2\pi x_i - \pi), \sum_i p_i(t) \cos(2\pi x_i - \pi) \right)
\]

(5.10)

The difference between the equations 5.10 and 5.11 lies in the fact the positions are bound to be between 0 and 1, whereas directions can be negative and larger than 1 or \(-1\), which changes the transformation to and from angles for position and direction. Taken together, this gives an easy decoding scheme from the neural activity to a probabilistic read-out.

5.3 Results

5.3.1 Anisotropic diffusion transports information during the blank

Recurrent excitatory connectivity is a candidate for providing mechanisms for motion-based prediction. In order to prove the functionality of our approach, we show results for single example networks of three different connectivity types applied to the recurrent excitatory population: isotropic, anisotropic motion-based (speed tuning dependent) and anisotropic direction-based (speed tuning independent). For simplicity, the other connectivities are set up according to the isotropic scheme. Networks using the anisotropic scheme for one or more of the other pathways \((E-I, I-E, I-I)\) could be tuned to perform similarly (not shown).

The connectivity in our model is mainly controlled by two parameters: the sum of incoming weights and the number of connections received by a cell. The sum of incoming weights for excitatory-excitatory connections have been tuned so that the activity initiated by the stimulus is strong enough to propagate through the network when the stimulus is turned off after 400 ms of input driven activation. Weights from the excitatory to the inhibitory population have been chosen so that inhibitory neurons exhibit a reasonable
level of average activity of approximately 5 Hz when the stimulus is driving the excitatory population. The role of inhibitory to excitatory connections is to balance the network activity when the stimulus is active after the blank. In contrast to balanced random networks, the inhibitory to excitatory weights could not be set to high values that compensate for the higher number of excitatory neurons as in previous models (e.g., [Bru00; MAD07]) because strong isotropic inhibitory feedback would silence excitatory neurons in the vicinity and impede the propagation of motion information during the blanking period. The interplay between the excitatory and inhibitory populations is crucial for balancing network activity, but more importantly, for suppressing activity that creates false predictions about the target trajectory. All connectivity parameters were tuned so that the spread of activity within the excitatory population is strong enough to fill a realistically long blanking period, where the average duration of a single blink is between 100 ms and 400 ms [Sch01].

Before stimulus onset, the network input consists of background noise that persistently drives the network at low firing rates. The stimulus spike trains are dispersed over the whole excitatory population (see figure 5.3). As this input is not coherent, the type of connectivity has no effect on the activity before onset. The stimulus activates the network between 200 ms and 600 ms of the simulation before another blank phase of 200 ms in which the input is equal to the phase before stimulus onset. Neurons that are well tuned to the stimulus fire at very high rates (temporarily up to 250 Hz) when the stimulus is present. This is due to the strong input stimulus and the amplification by the recurrent excitation. The average firing rate of neurons being active at least once during the simulation grows to up to 20 Hz when the stimulus is persistently present.

During the blank phase the global network activity drops rapidly to a low average rate of approximately 2 Hz and those neurons that convey the remaining motion information fire approximately 5 to 15 spikes during the blanking period, as individual output rates remain elevated to levels of 25 Hz to 75 Hz. Due to the anisotropic connectivity the activity triggered by the stimulus propagates through the network in the direction that was initiated by the target (see figures 5.3 and 5.4).

We observed that there needs to be a balance between stimulus induced excitation and the recurrent excitation: When recurrent excitation is too strong, the internal neural dynamics dominate over the activity triggered by the stimulus and it is likely that false tracking behavior occurs, i.e., network activity spreads too fast in the direction of the stimulus and the predicted trajectory gets ahead of the target. When recurrent excitation is not strong enough, the network activity fails to fill the blank by its own dynamics like in the network with isotropic connectivity.

The connectivity parameters $\sigma_X$ and $\sigma_Y$ need to be chosen differently for the two anisotropic networks, because their role in determining the connection
probability between cells is slightly different according to equations 5.5 and 5.8. For motion-based (MB) connectivity we used $\sigma_{X,V}^{MB} = 1$, and for direction-based (DB) $\sigma_{X,V}^{DB} = 0.5$. When $\sigma_X < \sigma_V$ two main effects could be observed. As the network connectivity is more specific in the spatial domain, the prediction performance in the target direction tends to be lower and in the target position gets more precise. But the prediction can get ahead of the stimulus because excitation spreads to fast along the predicted trajectory. In the opposite case, when $\sigma_V < \sigma_X$, the prediction of target direction gets more precise. Since the connectivity is spatially more distributed, the network is less likely to fill the blank because excitation is diluted across space.

5.3.2 Reading-out the population response

Jazayeri and Movshon [JM06] presented a framework for a generic representation of likelihoods of sensory stimuli by neural activity. Here, we used a similar approach (see 5.2.6) which allows us to transform the binary spiking activity into a continuous valued representation of probability about the target motion. By these means, the activity of individual neurons can be interpreted as time-continuous confidence measure.

The phase before stimulus onset is like a prior probability. It is dominated by noisy activity that seems uniform and does not converge into a coherent probability distribution (see figure 5.4). At stimulus onset, the network activity increases instantaneously and the probability distribution changes into a meaningful representation of motion information. During the blank period, the network activity drops rapidly (sometimes more gradually) and the probability distribution becomes more noisy, but changes less dramatically. Hence, despite the overall decrease in activity, information is not lost when the stimulus disappears. Instead, activity continues to propagate through the network, driven by the anisotropic connectivity. When the stimulus reappears the network activity grows again and continues to grow up to an average rate of 20 Hz until it is counterbalanced by the inhibitory feedback.

5.3.3 Motion-based predictive anisotropic diffusion solves motion extrapolation

In order to get a global estimation of the motion information we combine the probability estimates of individual neurons as described earlier (see 5.2.6) by a linear weighting of their time-varying activity. This provides a single valued, time-continuous prediction, i.e. readout signal of target position and direction. We will now compare the maximum confidence response for the three different connectivities to the exact same input in order to investigate the effect of the network connectivity on the readout signal.

Before stimulus onset the readout signal of all three networks follows the same noisy time course (see figure 5.5). After stimulus onset and after the
blank, all three estimations coincide with the actual target position and are very close to the target direction. This shows that our simple linear decoding approach is sufficient to translate the network activity into a meaningful readout signal.

The difference between the three networks can be seen during the blank phase. During this period, the readout signal of the network with the isotropic connectivity returns to the noisy time course, just like before the stimulus onset. In contrast, readout from the networks with anisotropic connectivity continues to give a precise estimation of the target position and direction of motion, as can be seen from the low prediction error (see the right most columns of figure 5.5). The readout from the direction-based connectivity is less accurate than the motion-based connectivity, but it still shows that the direction-based diffusion mechanism allows for inference of the target position and direction during the blank phase. Thus, it can solve the motion extrapolation problem.

In some experiments, we observed that direction prediction appears to be biased towards higher velocities - especially during the blank. Improvements to the connectivity rule might be necessary to gain a “perfect” prediction performance (zero root-mean-square error). The reason for the drift towards higher velocities can be seen as an unbalanced distribution of incoming weights. Neurons with higher velocities are more seldom and hence have less cells with similar tuning properties in their vicinity. Due to the fact that all cells receive the same sum of incoming weights, the comparatively few cells that project to cells with high preferred velocities do this via few, strong connections, possibly leading to the observed drift and instabilities in the network dynamics. This could possibly be solved by improving how probabilities are mapped to connection weights, e.g., by introducing a non-linearity that prohibits weights above a certain value. Nevertheless, it was not our objective to present an optimal ad-hoc connection algorithm that gives perfect prediction performance, but to prove the fact that anisotropic, tuning property-based connectivity could be an important mechanism achieving motion-based prediction.

### 5.3.4 Conclusion

The comparison of the prediction performance of the three different networks shows two main points. Anisotropic connectivity provides a mechanism for the diffusion of motion information, which is relevant to predict future trajectories in noisy environments where the flow of information is interrupted frequently. Also, our simple approach to read out network activity linearly is sufficient to solve the given task, and does not require knowledge about probability distributions gained over many trials.
5.4 Discussion

5.4.1 Summary and comments

Following our previous study [PM12b], we have confirmed that anisotropic diffusion of information is a sufficient mechanism to realize motion-based prediction as defined by the moving-dot blanking experiment. We have studied the role of different anisotropic and isotropic connectivity patterns and have shown that network connectivities that take into account the tuning properties of neurons and prefer smooth trajectories are more efficient in predicting the trajectory of a disappearing moving stimulus than isotropic networks. The main contribution of this study is to show that anisotropic diffusion of motion information can be implemented in networks of spiking neurons and thus could be a generic mechanism for motion-based prediction. Furthermore we have presented, to the best of our knowledge, the first model for motion prediction using spiking neurons and selective anisotropic connectivity that is inspired by a probabilistic framework.

The presented model is certainly limited and unrealistic in many ways. We have intentionally chosen a simple model (in terms of neuronal and synaptic features) to focus on the effects of connectivity patterns and to explore possible future applications for neuromorphic hardware systems [Bru+11; Sch+10]. One of the main limitations consists of using long interconnection delays that are necessary to achieve its main function in the current state of the model. According to the spatial scale and the conduction delays of cortical networks, the connection delays resulting from our model are not on the same order of magnitude. Assuming that the concept of anisotropic diffusion of information operates in motion processing networks, neuronal mechanisms other than such long delays would be required to achieve the desired predictive function. One reason for long delays is likely the size of our network and that we are sub-sampling neurons in comparison to realistic cortical network sizes. In larger networks it would be sufficient to have recurrent excitation work on a more local scale and long-range connectivity would implement the transport of an expectation signal, possibly in the subthreshold domain. We have successfully explored one possibility to relieve the need for long delays by constraining the connectivity to more local scales, which reduced the required delays by more than one order of magnitude. Instead of the axonal delays as employed by our model, dendritic delays, long synaptic time constants (provided e.g. NMDA or GABA-B currents), or a combination of those three mechanisms could be used to implement the same principle. In summary, larger network sizes and longer synaptic time constants would likely help to realize our approach with shorter, more realistic connection delays. The fact that both models with longer and shorter delays show very similar performance could be a hint for the generality of the idea of motion information being transported by anisotropic connectivity.
5.4.2 Context of existing models

Based on early ideas by Hubel and Wiesel [HW62], models employing anisotropic connectivity have been used to describe orientation selectivity [FHC78] and its general use for visual information processing of static images with non-spiking units [Ryb+91]. Models that are more similar to ours in motion coherence and the representation of motion trajectories (e.g. [BYG00; JE10]) do not use anisotropy in setting up the network connectivity. Other continuous recurrent network models have been used for various tasks like spatial working memory [Com+00] and categorical discrimination with veridical judgment of motion [LW08]. Our model works on a different level, but combines the functional features of previous models in the sense that motion trajectories are represented in a spiking and probabilistic way. Prediction signals are transported through recurrent connectivity, but none of the earlier models has shown that anisotropy could be a key element for this.

The dynamics of our network show some similarity to synfire chains [Pru+98] and we believe a model with similar dynamics and functionality could be implemented by attractor networks making use of a columnar organization [Lun+06] that is prominent in motion processing areas like are V5/MT [ADG84]. Work by Bressloff [Bre01] has shown that weak heterogeneities in excitable neural media can lead to wave propagation failure. We have shown that in principle, it is favorable to have heterogeneity (i.e. the anisotropy in connectivity) to promote the spread of activity in a meaningful way. Still, there is much experimental evidence showing that cortical networks can spread activity in the form of traveling waves [SNC12], and it is believed that long-range horizontal connections might be one of the underlying mechanism. It is arguable how well the conclusions of the Wilson-Cowan formalism used by [Bre01; JE10; Ryb+91] can be translated into the context of spiking networks, especially if the neuronal and synaptic machinery get more complex. We leave it for future analysis to determine if our network model could show behaviors similar to ones observed in experiments (for a review on traveling waves see [SNC12]).

We applied a simple vector averaging strategy to decode the position and motion direction from the network’s response [GSK], but the optimal way of decoding is up to debate. Several studies [PHB05; PL04] suggest a vector averaging approach with a bias term to estimate speed, but more recent experiments suggest that "perceived speed is not based on a labeled-line interpretation of MT cells" [KWA06]. Alternative decoding approaches involve a winner-take-all mechanism [LW08] or probabilistic codes [Bec+08]. We may use an existing method to decode the optimal estimate from the population of neurons as is done in real neural data [JM06], though our goal here was to show in a simple way that anisotropic and selective connectivity could be of great importance for motion prediction.
Based on the results of our model, we predict that the connectivity in higher cortical motion-processing areas like MT or MST is not isotropic, but that effective connectivity between cells depends on their tuning properties. A sign of this anisotropic, tuning-property based connectivity could possibly be seen in future experiments similar to those in [Guo+07] in the form of an anticipatory signal in cells that “expect” to receive stimulus input via the recurrent network connections.

5.4.3 Outlook and future work

After having shown a proof-of-concept for the idea that motion-based prediction can be achieved through anisotropic connectivity, many problems could be explored by the presented framework. One of the most urgent challenges in our view is the question how the recurrent connectivity can develop in a self-organized and robust manner. In order to integrate our model into the visual hierarchy, we need to understand how the tuning properties introduced in our model could be constructed through either feed-forward connections from lower cortical areas, through recurrent mechanisms that shape the desired properties, or both. Similarly important is the question of how a selective connectivity involving the inhibitory population influences the effective receptive field sizes along with the performance and stability of the presented framework. Another future challenge future is to use our probabilistic framework of spiking neurons for more realistic input and towards real-world applications.
Figure 5.1: The motion extrapolation problem. Sensory input, such as the smooth motion of a dot in visual space, may be perturbed by disruption of sensory drive, like when the eye blinks during a visual stimulation. It is essential that some mechanisms may fill this blank: this defines the motion extrapolation problem. We first define the problem by parameterizing a generic input and its perturbation. Left: The input is a Gaussian hill of activity in a topographically organized space, moving on a straight trajectory. We show here a snapshot in time of the input (blue) and the resulting input activity to the network (gray) for a period of 400 ms. This corresponds for instance to the activation of a low-level visual area to a single dot represented by a bell-shaped hill of activity (blue blurred circle). In addition, this input carries information about the motion of the object (blue arrow) and drives neurons which have a close selectivity in position and velocity (gray arrows). Right: We show the time course of the input for one representative neuron (denoted by the yellow star in the left panel). Top: The blue trace shows the envelope of the inhomogeneous Poisson process that creates the input spike train. For $0 \text{ ms} < t \leq 200 \text{ ms}$ and $600 \text{ ms} < t \leq 800 \text{ ms}$ the stimulus is blanked, that is, that all neurons in the sensory layer receive input from a Poisson process with the same rate. We permuted the input vector fed into the network among all the cells in the network for each time step during the blank. Black vertical lines indicate input spikes. Bottom: Histogram of the input spike train with a bin size of 50 ms. This shows clearly the missing information during the blank. We define the goal of solving the motion extrapolation problem as representing the prediction of information on motion (speed and position) during the blank.
Figure 5.2: From aniso- to iso-tropic connectivities. We propose that the local connectivity pattern of lateral connections play a crucial role in solving the motion-extrapolation problem. In order to show that, we compared different connectivity patterns in response to the blanked input (see figure 5.1). In the three panels, we show the incoming and outgoing connections for the same single neuron (as marked by the yellow diamond) for different connection rules. The preferred direction of that neuron is shown by the yellow arrow. Cells targeted by the yellow cell are marked with black circles and cells projecting to the yellow cell are marked with red triangles. The relative weights of incoming (outgoing) connections are indicated by the size of the source (target) neuron, respectively. The preferred direction of source and target neurons is shown by solid arrows. Connection delays are color coded. Left: Motion-based prediction anisotropic connectivity. Inspired by previous work on motion-based prediction [PM12b], we propose a first pattern of connectivity based on connecting a source neuron to a target neuron if and only if the position and velocity of the target is compatible with a smooth trajectory that would originate from the source neuron. The strength of this prediction is parameterized by the width of the lateral tuning selectivity and here we show the prototypical pattern for $\sigma_X = 0.3$ and $\sigma_V = 0.3$ ms$^{-1}$ (as used for the simulations). Middle: Direction-dependent connectivity. To create a more realistic connectivity pattern, we used the same rule but independently of speed (i.e., the modulus of velocity), but only as a function of the direction of motion. The target neuron is connected if and only if its direction is close to the source’s direction and if its position is predicted to be in the direction given by the source neuron. Additionally, to account for physiological constraints on lateral interaction, only connections within a radius of $r_{Conn} = 0.10$ or latencies shorter than 100 ms are allowed. This leads to a more local connectivity and smaller connection delays compared to the previous connectivity. We show here the resulting connectivity pattern for $\sigma_X = 0.3$ and $\sigma_V = 0.3$ ms$^{-1}$ with the motion-dependent connectivity and $\sigma_X = 1.0$ and $\sigma_V = 1.0$ ms$^{-1}$ with the direction-dependent connectivity. Right: An isotropic connectivity pattern was chosen as a control. There is no prediction in velocity space, but we still predict that activity should diffuse locally, as the connection probability drops with the distance between cells.
Figure 5.3: **Rasterplot of input and output spikes.** The raster plot from excitatory neurons is ordered according to their position. Each input spike is a blue dot and each output spike is a black dot. While input is scattered during blanking periods (figure 5.1), the network output shows some tuned activity during the blank (compare with the activity before visual stimulation). To decode such patterns of activity we used a maximum-likelihood estimation technique based on the tuning curve of the neurons.
Figure 5.4: Probabilistic population decoding and the resolution of the motion extrapolation problem using anisotropic connectivity.

The computed prediction confidence resulting from our simulations is shown using the motion-based anisotropic connectivity pattern with respect to time. The three vertical dashed lines correspond to: the onset of the stimulus, the onset of the blank and finally the reappearance of the stimulus, respectively. The cells’ prediction confidence is defined in equation 5.9 and have been sorted and binned according to their tuning properties. The accumulated confidence within each time bin is color coded. 

Top: left (right): We show the prediction confidence of movement direction $u = v_x$ ($v = v_y$). Bottom: left (right): Prediction confidence about $x$ and $y$ position, respectively. While information is distributed before stimulation and quickly converges during stimulation, it is predicted during the blank: the motion extrapolation is solved and information is very quickly recovered at the reappearance of the stimulus.
Figure 5.5: **Comparison of prediction performance for the different connectivities.** The performance of direction (top panels) and position (bottom panels) prediction as decoded from the network activity is shown (see equations 5.10, 5.11). First and second columns show the horizontal and vertical components, respectively, while the last column shows the mean squared error of the predicted position with respect to the known position of the target. The color of the lines correspond to the different connectivities presented in figure 5.2: motion-based prediction (solid blue), direction-dependent prediction (dashed green), isotropic (dash-dotted red). While an isotropic connectivity clearly fails to predict the fate of the input during the blank, we show here that the anisotropic connectivities may efficiently solve the motion extrapolation problem, even with an approximate solution such as the direction-based prediction.
Chapter 6

Active inference, eye movements and oculomotor delays

Abstract

This chapter considers the problem of sensorimotor delays in the optimal control of (smooth) eye movements under uncertainty. Specifically, we consider delays in the visuo-oculomotor loop and their implications for active inference. Active inference uses a generalization of Kalman filtering to provide Bayes optimal estimates of hidden states and action in generalized coordinates of motion. Representing hidden states in generalized coordinates provides a simple way of compensating for both sensory and oculomotor delays. The efficacy of this scheme is illustrated using neuronal simulations of pursuit initiation responses, with and without compensation. We then consider an extension of the generative model to simulate smooth pursuit eye movements—in which the system believes both the target and its centre of gaze are attracted to a (fictive) point moving in the visual field. Finally, the generative model is equipped with a hierarchical structure, so that it can recognize and remember unseen (occluded) trajectories and emit anticipatory responses. These simulations speak to a straightforward and neurobiologically plausible solution to the generic problem of integrating information from different sources with different temporal delays and the particular difficulties encountered when a system—like the oculomotor system—tries to control its environment with delayed signals.

6.1 Introduction

The oculomotor system produces eye movements to deploy sensory (retinal) epithelia at very fast timescales. In particular, changes in the position of a
visual object are compensated for with robust and rapid eye movements, such that the object is perceived as invariant, despite its motion [Ilg97; LMT87]. This near-optimal control is remarkable, given the absence of any external clock to coordinate dynamics in different parts of the visual-oculomotor system. An important constraint, in this setting, is axonal conduction, which produces delays in sensory and motor signalling within the oculomotor system. For example, in humans, retinal signals arriving at motion processing areas report the state of affairs at least about 50 ms ago, while the action that follows is executed at least 40 ms in the future [IK06]; for a review, see [MI10] (see Figure 6.1). Different sources of delays exist – such as the biomechanical delay between neuromuscular excitation and eye movement. Due to these delays, the human smooth pursuit system responds to unpredictable stimuli with a minimum latency of around 100 ms [WP87]. In addition, these delays may produce oscillations about a constant velocity stimulus [RGG86; Rob65], whose amplitude and frequency can be altered by artificially manipulating the feedback [GKL92].

Eye movements can anticipate predictable stimuli, such as the sinusoidal movement of a pendulum [BA91; DTF30; Wes54]; for a review, see [Bar08]. Interestingly, ocular tracking can compensate for sensorimotor delays after around one or two periods of sinusoidal motion – producing a tracking movement with little discernible delay [BA91]. This suggests that the oculomotor system can use sensory information from the past to predict its future sensory states (including its actions), despite the fact that these sensory changes can be due to both movement of the stimulus and movement of the eyes. The time taken to compensate for delays increases with the unpredictability of the stimulus [MJ66], though the system can adapt quickly to complex waveforms, with changes of velocity [BS02], single cycles [BBC00] or perturbed periodic waves – where subjects appear to estimate their frequency using an average over recent cycles [CB09]. Further studies suggest that different sources of information, such as auditory or verbal cues [Kow89] or prior knowledge about the nature of sensory inputs [MSM06] can evoke anticipatory eye movements.

The aim of this work was to establish a principled model of optimal visual motion processing and oculomotor control in the context of sensorimotor delays. Delays are often ignored in treatments of the visual-oculomotor system; however, they are crucial to understanding the system’s dynamics. For instance, delays may be important for understanding the pathophysiology of impaired oculomotor control: schizophrenic smooth pursuit abnormalities are due to impairments of the predictive (extra-retinal) motion signals that are required to compensate for sensorimotor delays [Nka+10; Tha+99]. Surprisingly, delays may also explain a whole body of visual illusions [Cha+08; Cha01; CW02; VE13], even for visual illusions that involve a static display. Delays are also an important consideration in control theory and engineering. Finally, neuronal solutions to the delay problem speak to the representation
of time in the brain, which is essential for the proper fusion of information in the central nervous system.

6.1.1 Existing solutions and the proposed hypothesis

A principled approach to optimal oculomotor control is provided by Bayesian filtering schemes that use probabilistic representations to estimate visual and oculomotor states. These states are hidden; i.e., they cannot be measured directly. A popular scheme for linear control problems is the Kalman filter [Kal60]. The Kalman scheme can be extended to accommodate biomechanical constraints, such as transmission delays (e.g., fixed-lag smoothers). However, their solutions can become computationally complex when delays are large in relation to discretisation time and are not biologically plausible.

We have previously considered generalized Bayesian filtering in continuous time as a metaphor for action and perception. This approach has been applied to eye movements [Fri+10a] and saccades in particular [Fri+12]. However, these applications ignored sensorimotor delays and their potentially confounding effects on optimal control.

Crucially, the active inference schemes we have considered previously are formulated using representations in generalized coordinates of motion; that is, states (such as position) are represented along with their higher order temporal derivatives (such as speed, acceleration, jerk, etc). This means that one has an implicit representation of hidden states in the recent past and future that can be used to finesse the problems of delays. For example, it has been shown that acceleration is a necessary component of the predictive drive to eye movements [Ben+07]. In brief, generalized representations can be projected to the past and to the future using simple (linear) mixtures of generalized motion. Note that a representation of generalized motion can be explicit or implicit by using a population coding scheme – as has been demonstrated for acceleration [LM99]. Representations of generalized motion may be important for modelling delays when integrating information in the brain from distal sources – such as other cortical columns in the same cortical area or other areas that are connected with fixed but different delays [Roe+97]. The integration of information over time becomes particularly acute in motor control, where the products of sensory processing couple back to the sampling of sensory information through action.

In the context of action, acted inference finesse the problems with delayed control signals in classical formulations of motor control by replacing command signals with descending cortico-spinal predictions. For instance, the location of receptive fields in the parietal cortex in monkeys is shown to shift transiently before an eye movement [DCG92]. These predictions are fulfilled at the peripheral level, using fast closed loop mechanisms (peripheral reflex arcs). In principle, “these predictions can anticipate delays if they are part of the generative model,” [Fri11]: However, this anticipation has
never been demonstrated formally. Here, we show how generalized Bayesian filtering – as used in active inference – can compensate for both sensory and motor delays in the visual-oculomotor loop.

It is important to mention what this work does not address. First, we focus on tracking eye movements (pursuit of a single dot stimulus for a monocular observer with fixed head position): we do not consider other types of eye movements (vergence, saccades, or the vestibulo-ocular reflex). Second, we take an approach that complements existing models, such as those of [RGG86] and [KL89]. Existing models account for neurophysiological and behavioural data by refining block-diagram models of oculomotor control to describe how the system might work. We take a more generic approach, in which we define the imperatives for any system sampling sensory data, derive an optimal oculomotor control solution and show why this solution explains the data. Although the two approaches should be consistent, ours offers a principled approach to identifying the necessary solutions (such as predictive coding) to a given problem (oculomotor delays). We hope to demonstrate the approach by modelling pursuit initiation and smooth pursuit – and then consider the outstanding issue of anticipatory responses: in previous treatments [RGG86], “[ anticipation] has not been adequately modelled and no such attempt is offered. . . only unpredictable movements are considered.”

6.1.2 Outline

The main contributions of our work are described in the subsequent five sections. First, section 6.2 summarises the basic theory behind active inference and attempts to link generalized filtering to conventional Bayesian filters used in optimal control theory. This section then considers neurobiological implementations of generalized filtering, in terms of predictive coding in generalized coordinates of motion. This formulation allows us to consider the problem of delayed sensory input and motor output in section 6.3 – and how this problem can be finessed in a relatively straightforward way using generalized representations. Having established the formal framework (and putative neuronal implementation), the final three sections deal with successively harder problems in oculomotor control. We start in Section 6.4 by considering pursuit initiation using a simple generative model of oculomotor trajectories. Using simulations, we consider the impact of motor delays, sensory delays and their interaction on responses to a single sweep of a visual target. The subsequent section turns to smooth pursuit eye movements – using a more sophisticated generative model of oculomotor trajectories, in which prior beliefs about eye movements enable the centre of gaze to predict target motion using a virtual or fictive target (see Section 6.5). In the final section, we turn to hierarchical models of target trajectories that have explicit memories of hidden dynamics, which enable anticipatory responses (see Section 6.6). These responses are illustrated using simulations of antici-
patory pursuit movements using (rectified) hemi-sinusoidal motion. In short, these theoretical considerations lead to a partition of stimulus-bound eye movements into pursuit initiation, smooth pursuit and anticipatory pursuit, where each mode of oculomotor control calls on formal additions to the underlying generative model; however, they all use exactly the same scheme and basic principles. Where possible, we try to simulate classic empirical results in this field – at least heuristically.

In short, these theoretical considerations lead to a partition of stimulus-bound eye movements into pursuit initiation, smooth pursuit and anticipatory pursuit, where each mode of oculomotor control calls on formal additions to the underlying generative model. However, these models all use exactly the same scheme and basic principles; in particular, they all use the same solution to the oculomotor delay problem. These simulations illustrate that the active inference scheme can reproduce classical empirical results in three distinct experimental contexts.

6.2 From predictive coding to active inference

This section sets out the basic theory, before applying it to the special problem of oculomotor delays in the following sections. We first introduce the general framework of active inference in terms of generalized Bayesian filtering and variational free energy minimisation. In brief, active inference can be regarded as equipping standard Bayesian filtering schemes with classical reflex arcs that enable action, such as an eye movement, to fulfil predictions about hidden states of the world. Second, we will briefly describe the formalism of active inference in terms of differential equations describing the dynamics of the world and internal states of the visual-oculomotor system. The neurobiological implementation of these differential equations is considered in terms of predictive coding, which uses prediction errors on the motion of hidden states – such as the location of a visual target. In the next section, we will turn to the special problem of oculomotor delays and how this problem can be finessed using active inference in generalized coordinates of motion. This solution will be illustrated in subsequent sections using simulations of pursuit initiation responses and smooth pursuit. Finally, we shall exploit the richness of hierarchical generative models – which underlie active inference – to illustrate anticipatory eye movements that cannot be explained by simply compensating for oculomotor delays.

6.2.1 From free energy to generalized filtering

The scheme used here to model oculomotor behaviour has been used to model several other processes and paradigms in neuroscience. This active inference scheme is based upon three assumptions:
Figure 6.1: Problem statement: optimal motor control under axonal delays. The central nervous system has to contend with axonal delays, both at the sensory and the motor levels. For instance, in the human visuo-oculomotor system, it takes approximately $\tau_s = 50$ ms for the retinal image to reach the visual areas implicated in motion detection, and a further $\tau_m = 40$ ms to reach the oculomotor muscles. As a consequence, for a tennis player trying to intercept a ball at a speed of $20 \, \text{m} \cdot \text{s}^{-1}$, the sensed physical position is 1 m behind the true position (as represented here by $\tau_s \cdot \vec{V}$), while the position at the moment of emitting the motor command will be .8 m ahead of its execution ($\tau_m \cdot \vec{V}$). Note that while the actual position of the ball when its image produced by the photoreceptors on the retina hits visual areas is approximately at 45 degrees of eccentricity (red dotted line), the player’s gaze is directed to the ball at its *present* position (red line), in anticipatory fashion. Optimal control directs action (future motion of the eye) to the expected position (red dashed line) of the ball in the future — and the racket (black dashed line) to the expected position of the ball when motor commands reach the periphery (muscles).
Exchange with the environment

\[
\psi = f(\psi, a) + \omega \\
\overline{\psi} = \arg\min_a F(\overline{s}, \overline{\mu})
\]

\[
\overline{\mu} = \arg\min_{\overline{\mu}} F(\overline{s}, \overline{\mu})
\]

\[
s = g(\psi, a) + \omega \\
a = \arg\min_a F(\overline{s}, \overline{\mu})
\]

Figure 6.2: This schematic shows the dependencies among various quantities modelling exchanges of an agent with the environment. It shows the states of the environment and the system in terms of a probabilistic dependency graph, where connections denote directed dependencies. The quantities are described within the nodes of this graph – with exemplar forms for their dependencies on other variables (see main text). Hidden (external) and internal states of the agent are separated by action and sensory states. Both action and internal states – encoding a conditional probability density function over hidden states – minimise free energy. Note that hidden states in the real world and the form of their dynamics can be different from that assumed by the generative model; this is why hidden states are in bold. See main text for further details.
The brain minimises the free energy of sensory inputs defined by a generative model.

The generative model used by the brain is hierarchical, nonlinear and dynamic.

Neuronal firing rates encode the expected state of the world, under this model.

The first assumption is the free energy principle, which leads to active inference in the context of an embodied interaction of the system with its environment; where the system can act to change its sensory inputs. The free energy here is a variational free energy that provides a computationally tractable upper bound on the negative logarithm of Bayesian model evidence (see Appendix 1). In Bayesian terms, this means that the brain maximises the evidence for its model of sensory inputs [BHS83; BNT01; Day+95; Gre80; Gro+97; KP04; OF96]. This is the Bayesian brain hypothesis [YK06]. If we also allow action to maximise model evidence, we get active inference [Fri+10a]. Crucially, unlike conventional optimal control schemes, there is no ad hoc value or loss function guiding action: Action minimises the free energy of the system's model. This permits the application of standard Bayesian solutions and simplifies the implicit neuronal architecture; for example, there is no need for an efference copy signal [Fri11]. In this setting, desired movements are specified in terms of prior beliefs about state transitions or the motion of hidden states in the generative model. Action then realises prior beliefs (policies) by sampling sensory data that provides evidence for those beliefs.

The second assumption above is motivated by noting that the world is both dynamic and nonlinear – and that hierarchical structure emerges inevitably from a separation of temporal scales [Gin55; Hak83]. The third assumption is the Laplace assumption that, in terms of neural codes, leads to the Laplace code, which is arguably the simplest and most flexible of all neural codes [Fri09].

In brief, the Laplace code means that probabilistic representations are encoded explicitly by synaptic activity in terms of their mean or expectation (while the second order statistics such as dispersion or precision are encoded implicitly in terms of synaptic activity and efficacy). This limits the representation of hidden states to continuous variables, as opposed to discrete states; however, this is appropriate for most aspects of sensorimotor processing. Furthermore, it fineses the combinatoric explosion associated with discrete state space models. Restricting probabilistic representations to a Gaussian form clearly precludes multimodal representations. Having said this, the hierarchical form of the generative models allows for fairly graceful modelling of nonlinear effects (such as shadows and occlusions). For example, a Gaussian variable at one level of the model may enter the lower levels in
highly non-linear way — we will see examples of this later. See Appendix 2 for a motivation of the Laplace assumption from basic principles.

Under these assumptions, action and perception can be regarded as the solutions to coupled differential equations describing the dynamics of the real world (the first pair of equations) and the behaviour of an agent (the second pair of equations); expressed in terms of action and internal states that encode conditional expectations about hidden states of the world \[\text{Fri+10a}\]:

\[
\begin{align*}
    s &= g(x, \nu, a) + \omega_x \\
    \dot{x} &= f(x, \nu, a) + \omega_x \\
    \dot{a} &= -\partial_a F(\tilde{s}, \tilde{\mu}) \\
    \dot{\tilde{\mu}} &= D\tilde{\mu} - \partial_{\tilde{\mu}} F(\tilde{s}, \tilde{\mu})
\end{align*}
\] (6.1)

For clarity, real-world states are written in boldface, while internal states of the agent are in italics: Here \((s, x, \nu, a)\) denote sensory input, hidden states, hidden causes and action in the real world, respectively. The variables in the second pair of equations \((\tilde{s}, \tilde{\mu}, a)\) correspond to generalized sensory input, conditional expectations and action. Generalized coordinates of motion, denoted by the \(\tilde{}\) notation, correspond to a vector representing the different orders of motion of a variable: position, velocity, acceleration, and so on \[\text{Fri+10b}\]. Using the Lagrangian notation for temporal derivatives, we get e. g. for \(s\): \(\tilde{s} = (s, s', s'', \ldots)\). In the absence of delays \(\tilde{s}(t) = \tilde{s}(t)\) the agent receives instantaneous sensations from the real world. The differential equations above are coupled because sensory states depend upon action through hidden states and causes \((x, \nu)\) while action \(a(t) = a(t)\) depends upon sensory states through internal states \(\tilde{\mu}\).

By explicitly separating real-world states — hidden from the agent — to its internal states, one can clearly separate the generative model from the updating scheme that allows to minimise the agent’s free-energy: The first pair of coupled stochastic differential equations describes the dynamics of hidden states and causes in the world and how they generate sensory states. These equations are stochastic because sensory states and the motion of hidden states are subject to random fluctuations \((\omega_x, \omega_\nu)\).

The second pair of differential equations corresponds to action and perception respectively — they constitute a (generalized) gradient descent on variational free energy. The differential equation describing changes in conditional expectations (perception) is known as generalized filtering or predictive coding and has the same form as standard Bayesian (Kalman-Bucy) filters — see also \[\text{Bea03; RB99}\]. The first term is a prediction based upon a differential operator \(D\) that returns the generalized motion of the conditional expectations; namely the vector of velocity, acceleration, jerk and so on.
such that $D\tilde{\mu} = (\mu', \mu'', \mu'''\ldots)$. However, the expected velocity is not the velocity of the expectation and comprises both prediction and update terms: The second term reflects this correction and ensures the changes in conditional expectations are Bayes-optimal predictions of hidden states of the world – in the sense that they maximise the free energy bound on Bayesian model evidence. See Figure 6.2 for a schematic summary of the implicit conditional dependencies implied by Equation 6.1.

6.2.2 Hierarchical form of the generative model

To perform simulations using this scheme, one simply integrates or solves Equation 6.1 to simulate (neuronal) dynamics that encode conditional expectations and ensuing action. Conditional expectations depend upon a generative model, which we assume has the following (hierarchical) form

\begin{equation}
\begin{aligned}
s &= g^{(1)}(x^{(1)}, v^{(1)}) + \omega^{(1)}_s \\
\dot{x}^{(1)} &= f^{(1)}(x^{(1)}, v^{(1)}) + \omega^{(1)}_x \\
\vdots \\
\nu^{(i-1)} &= g^{(i)}(x^{(i)}, v^{(i)}) + \omega^{(i)}_\nu \\
\dot{x}^{(i)} &= f^{(i)}(x^{(i)}, v^{(i)}) + \omega^{(i)}_x \\
\vdots
\end{aligned}
\end{equation}

Where $(i)$ indexes the level in the hierarchical model. Note that we denote the sensory layer as $i = 0$, but this indexing is somewhat arbitrary. This equation is just a way of writing down a generative model that specifies a probability density function over sensory inputs and hidden states and causes. This probability density is needed to define the free energy of sensory input: it is specified in terms of functions $(f^{(i)}, g^{(i)})$ and Gaussian assumptions about random fluctuations $(\omega^{(i)}_x, \omega^{(i)}_\nu)$ on the motion of hidden states and causes. It is these that make the model probabilistic – they play the role of sensory noise at the first level and induce uncertainty about states at higher levels. The precisions of these fluctuations are quantified by $\Pi^{(i)}_x, \Pi^{(i)}_\nu$ which are defined as the inverse of the respective covariance matrices.

The deterministic part of the model is specified by nonlinear functions of hidden states and causes $(f^{(i)}, g^{(i)})$ that generate dynamics and sensory consequences. Hidden causes link hierarchical levels, whereas hidden states link dynamics over time. Hidden states and causes are abstract quantities that the brain uses to explain or predict sensations – like the motion of an object in the field of view. In hierarchical models of this sort, the output of one level acts as an input to the next. This input can produce complicated
Figure 6.3: Schematic detailing a neuronal message passing scheme (generalized Bayesian filtering or predictive coding) that optimises conditional expectations about hidden states of the world, given sensory (visual) data and the active (oculomotor) sampling of those data. This diagram shows the speculative cells of origin of forward driving connections (in red) that convey prediction error from a lower area to a higher area and the backward connections (in black) that construct predictions [Mum92]. These predictions try to explain away prediction error in lower levels. In this scheme, the sources of forward and backward connections are superficial (red) and deep (black) pyramidal cells respectively. The equations on the right represent a generalized descent on free energy under the hierarchical model described in the main text – this can be regarded as a generalisation of predictive coding or Kalman filtering: see [Fri08]. State-units are in black and error-units are in red. Here, we have placed different levels of some hierarchical model within the visual-oculomotor system. Visual input arrives in an intrinsic (retinal) frame of reference that depends upon the angular position of a stimulus and the direction of gaze. Exteroceptive input is then passed to the lateral geniculate nuclei (LGN) and to higher visual and prefrontal (e.g., motion sensitive, such as the frontal eye field) areas in the form of prediction errors. Crucially, proprioceptive sensations are also predicted, creating prediction errors at the level of the cranial nerve nuclei (pons). The special aspect of these proprioceptive prediction errors is that they can be resolved through classical reflex arcs – in other words, they can elicit action to change the direction of gaze and close the visual–oculomotor loop.
convolutions with deep (hierarchical) structure. We will see examples of this later in particular in the context of anticipatory movements.

6.2.3 Perception and predictive coding

Given the form of the generative model (Equation 6.2) one can write down the differential equations (Equation 6.1) describing neuronal dynamics in terms of prediction errors on the hidden causes and states. These errors represent the difference between conditional expectations and predicted values, under the generative model (using $A \cdot B := A^T B$ for the scalar product and omitting higher-order terms):

$$\dot{\tilde{\mu}}^{(i)}_x = D\tilde{\mu}^{(i)}_x + \frac{\partial g^{(i)}}{\partial \tilde{\mu}^{(i)}_x} \cdot \Pi^{(i)}_x \tilde{\varepsilon}^{(i)}_x + \frac{\partial \tilde{f}^{(i)}}{\partial \tilde{\mu}^{(i)}_x} \cdot \Pi^{(i)}_x \tilde{\varepsilon}^{(i)}_x - D\Pi^{(i)}_x \tilde{\varepsilon}^{(i)}_x$$

$$\dot{\tilde{\mu}}^{(i)}_\nu = D\tilde{\mu}^{(i)}_\nu + \frac{\partial g^{(i)}}{\partial \tilde{\mu}^{(i)}_\nu} \cdot \Pi^{(i)}_\nu \tilde{\varepsilon}^{(i)}_\nu + \frac{\partial \tilde{f}^{(i)}}{\partial \tilde{\mu}^{(i)}_\nu} \cdot \Pi^{(i)}_\nu \tilde{\varepsilon}^{(i)}_\nu - \Pi^{(i)}_\nu \tilde{\varepsilon}^{(i)}_{\nu+1}$$

$$\tilde{\varepsilon}^{(i)}_x = D\tilde{\mu}^{(i)}_x - \tilde{f}^{(i)}(\tilde{\mu}^{(i)}_x, \tilde{\mu}^{(i)}_\nu)$$

$$\tilde{\varepsilon}^{(i)}_\nu = \tilde{\mu}^{(i-1)}_\nu - \tilde{g}^{(i)}(\tilde{\mu}^{(i)}_x, \tilde{\mu}^{(i)}_\nu)$$

(6.3)

The quantities $\tilde{\varepsilon}^{(i)}$ correspond to prediction errors (on hidden states $x$ or hidden causes $\nu$). These are weighted by their respective precision vectors $\Pi^{(i)}$ in the update scheme. Equation 6.3 can be derived fairly easily by computing the free energy for the hierarchical model in Equation 6.2 and inserting its gradients into Equation 6.1. This gives a relatively simple update scheme, in which conditional expectations are driven by a mixture of prediction errors, where prediction errors are defined by the equations of the generative model.

It is difficult to overstate the generality and importance of Equation 6.3 – its solutions grandfather nearly every known statistical estimation scheme, under parametric assumptions about additive noise [Fri08]. These range from ordinary least squares to advanced variational deconvolution schemes. In this form, one can see clearly the relationship between predictive coding and Kalman-Bucy filtering – changes in conditional expectations comprise a prediction (first term) plus a weighted mixture of prediction errors (remaining terms). The weights play the role of a Kalman gain matrix and are based on the gradients of the model functions and the precision of random fluctuations.

In neural network terms, Equation 6.3 says that error-units receive predictions from the same hierarchical level and the level above. Conversely, conditional expectations (encoded by the activity of state units) are driven by prediction errors from the same level and the level below. These constitute bottom-up and lateral messages that drive conditional expectations towards a
better prediction to reduce the prediction error in the level below. This is the essence of recurrent message passing between hierarchical levels to suppress free energy or prediction error: see [FK09] for a more detailed discussion. In neurobiological implementations of this scheme, the sources of bottom-up prediction errors, in the cortex, are thought to be superficial pyramidal cells that send forward connections to higher cortical areas. Conversely, predictions are conveyed from deep pyramidal cells by backward connections, to target (polysynaptically) the superficial pyramidal cells encoding prediction error [FK09; Mum92]. This defines an elementary circuit that may be the basis of the layered organisation of the cortex [Bas+12]. Figure 6.3 provides a schematic of the proposed message passing among hierarchically deployed cortical areas.

6.2.4 Action

In active inference, conditional expectations elicit behaviour by sending predictions down the hierarchy to be unpacked into proprioceptive predictions at the level of (pontine) cranial nerve nuclei and spinal-cord. These engage classical reflex arcs to suppress proprioceptive prediction errors and produce the predicted motor trajectory

$$\dot{a} = -\partial_a F = -\left(\partial_a \tilde{\varepsilon}^{(1)}_\nu \right) \cdot \Pi^{(1)}_\nu \tilde{\varepsilon}^{(1)}_\nu$$

The reduction of action to classical reflexes follows because the only way that action can minimise free energy is to change sensory (proprioceptive) prediction errors by changing sensory signals. This highlights the tight relationship between action and perception; cf., the equilibrium point formulation of motor control [FL95]. In short, active inference can be regarded as equipping a generalized predictive coding scheme with classical reflex arcs: see [FDK09; Fri+10a] for details. The actual movements produced clearly depend upon (changing) top-down predictions that can have a rich and complex structure. This scheme is consistent with the physiology and anatomy of the oculomotor system (for a review see [Ilg97; Kra04]; although our goal here is not to identify the role of each anatomical structure but rather to give a schematic proof-of-concept.

6.2.5 Summary

In summary, we have derived equations for the dynamics of perception and action using a free energy formulation of adaptive (Bayes-optimal) exchanges with the world and a generative model that is both generic and biologically plausible. A technical treatment of the material above will be found in [Fri+10b], which provides the details of the generalized filtering used to produce the simulations in the next section. Before looking at these simulations, we consider how delays can be incorporated into this scheme.
6.3 Active inference with sensorimotor delays

If action and sensations were not subject to delays, one could integrate (solve) Equation 6.1 directly; however, in the presence of sensory and motor delays ($\tau_s$ and $\tau_a$ respectively) Equation 6.1 becomes a (stochastic and non-linear) delay differential equation because $\tilde{s}(t) = \tilde{s}(t - \tau_s)$ and $a(t) = a(t + \tau_a)$. In other words, the agent receives sensations from (sees) the past, whilst emitting motor signals that will be enacted in the future (we will only consider delays from the sensory and motor sub-systems and neglect delays between neuronal systems in this chapter).

To finesse the integration of these delay differential equations one can exploit their formulation in generalized coordinates: By taking linear mixtures of generalized motion one can easily map from the present to the future, using the matrix operators:

$$T(\tau) = \exp(\tau D) = \begin{bmatrix}
1 & \frac{1}{1!} \tau & \frac{1}{2!} \tau^2 & \cdots \\
0 & 1 & \frac{1}{1!} \tau & 0 \\
0 & 0 & 1 & \cdots \\
0 & 0 & 0 & \cdots
\end{bmatrix}$$

(6.5)

with $D = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0
\end{bmatrix}$

The first differential operator simply returns the generalized motion $D\tilde{x}(t) = \tilde{x}'(t)$ while the second delay operator produces generalized states in the future $T(\tau)\tilde{x}(t) = \tilde{x}(t + \tau)$ (we define delays as positive by convention). Note that shifting forward and backwards by the same amount of time produces the identity operator $T(\tau)T(-\tau) = I$ and that, more generally, $T(\tau_1)T(\tau_2) = T(\tau_1 + \tau_2)$.

These delay operators are simple to implement computationally (and neurobiologically) and allow an agent to finesse the delayed coupling above by replacing (delayed) sensory signals with future input $\tilde{s}(t) = T(\tau_s)\tilde{s}(t - \tau_s) = \tilde{s}(t)$ for subsequent action and perception. Alternatively, one can regard this compensation for sensory delays as attempting to predict the past (see below). Generalized coordinates allow the representation of the trajectory of a given variable at any time (that is, its evolution in the near past and present) and thus allow its projection into the future or past. Generalized representations are more extensive than ‘snapshots’ at a particular time and enable the agent to anticipate the future (of delayed sensory trajectories) and represent hidden states in real time – that is, representations that are synchronised with the external events. In terms of motor delays, the agent can replace its
internal motor signals with action in the future $a(t) = T(\tau_a)a(t - \tau_a) = a(t)$, such that when action signals reach the periphery they correspond to the action encoded centrally. These substitutions allow us to express action and perception in Equation 6.1 as\(^1\):

\[
\dot{a}(t) = -\partial_a F(T(\tau_a)T(\tau_a)\tilde{s}(t - \tau_s - \tau_a), T(\tau_a)\tilde{\mu}(t - \tau_a)) = -\partial_a F(T(\tau_a - \tau_s + \tau_a - \tau_a)\tilde{s}(t), T(\tau_a - \tau_a)\tilde{\mu}(t))
\]

(6.6)

This equation distinguishes between true delays ($\tau$) and those assumed by the agent ($\tau$). When the two are the same, the delay operators $T(\tau - \tau) = I : \tau = \tau$ become identity matrices and Equation 6.6 reduces to Equation 6.1. When the two differ, Equation 6.6 permits the simulation of a system with uncompensated delays. Notice how the dynamics of action in the first differential equation are driven by a gradient descent on the free energy of sensations with composite sensory and motor delays. In other words, action in the real world depends upon sensory states generated $\tau_s + \tau_a$ in the past.

One can now solve Equation 6.6 to simulate active inference, with or without compensation for sensorimotor delays. We use a standard local linearisation scheme for this integration [Oza92], where delays enter at the point at which sensory prediction error is computed and when it drives action: from Equations 6.3 and 6.4:

\[
\begin{align*}
\dot{\varepsilon}^{(1)}_{\nu} &= T(\tau_s)\tilde{s}(t - \tau_s) - g^{(1)}(\tilde{\mu}_{\nu}^{(1)}, \tilde{\mu}_{\nu}^{(1)}) \varepsilon^{(1)}_{\nu} = T(\tau_s - \tau_s)\tilde{s}(t) - g^{(1)}(\tilde{\mu}_{\nu}^{(1)}, \tilde{\mu}_{\nu}^{(1)}) \\
\dot{a}(t) &= -(\partial_a \varepsilon^{(1)}_{\nu}) \cdot \Pi^{(1)}(\tau_a)\varepsilon^{(1)}_{\nu}(t - \tau_a) = -(\partial_a \varepsilon^{(1)}_{\nu}) \cdot \Pi^{(1)}(\tau_a - \tau_a)\varepsilon^{(1)}_{\nu}(t)
\end{align*}
\]

(6.7)

Equation 6.7 means that perfect (errorless) prediction requires $T(\tau_s)\tilde{s}(t - \tau_s) = g^{(1)}(\tilde{\mu}_{\nu}^{(1)}, \tilde{\mu}_{\nu}^{(1)})$. In other words, errorless prediction means that the agent is effectively predicting the future projection of the past. Note again the dependency of action, via prediction errors, on sensory states $\tau_s + \tau_a$ in the past.

\(^1\)We have made a slight approximation here because $T(\tau_a)\tilde{\mu}(t - \tau_a) = T(\tau_a - \tau_a)\tilde{\mu}(t)$ when, and only when, the free energy gradients are zero and $\tilde{\mu}(t) = D\tilde{\mu}(t)$. Under the assumption that the perceptual destruction of these gradients is fast, in relation to action, this can be regarded as an adiabatic approximation.
6.3.1 Summary
This section has considered how the differential equations describing changes in action and internal (representational) states can be finessed to accommodate sensorimotor delays. This is relatively straightforward — in the context of generalized schemes — using delay operators that take mixtures of generalized motion to project states into the future or past. Sensory delays can be (internally) simulated and corrected by applying delays to sensory input producing sensory prediction error, while motor delays can be simulated and corrected by applying delays to sensory prediction error producing action. Neurobiologically, the application of delay operators just means changing synaptic connection strengths to take different mixtures of generalized sensations and their prediction errors. We will now use these operators to look at the effects of sensorimotor delays with and without compensation.

6.4 Results: pursuit initiation
This section focuses on the consequences of sensory delays, motor delays and their combination — in the context of pursuit initiation — using perhaps the simplest generative model for active inference. Our purpose is to illustrate the difficulties in oculomotor control that are incurred by delays and how these difficulties dissolve when delays are accommodated during active inference. We start with a description of the generative model and demonstrate its behaviour when delays are compensated. We then use this normal behaviour as a reference to look at failures of pursuit initiation induced by delays. In this section, responses to a single sweep of rightward motion are used to illustrate basic responses. In the next section, we consider pursuit of sinusoidal motion (with abrupt onsets) and the implications for generative models that may be used by the brain.

6.4.1 Generative model of pursuit initiation
The generative model for pursuit initiation used here is very simple and is based upon the prior belief that the centre of gaze is attracted to the target location. The processes generating sensory inputs and the associated generative model can be expressed as follows:
\[ \begin{align*}
    s &= \begin{bmatrix}
    s_o \\
    s_t
\end{bmatrix} = \begin{bmatrix}
    x_o \\
    x_t - x_o
\end{bmatrix} + \omega_{\nu}^{(1)} \\
    \dot{x} &= \begin{bmatrix}
    \dot{x}_o \\
    \dot{x}_t
\end{bmatrix} = \begin{bmatrix}
    \frac{1}{t_o} a - \frac{1}{t_o} x_o \\
    \frac{1}{t_m}(\nu^{(1)} - x_t)
\end{bmatrix} + \omega_{x}^{(1)}
\end{align*} \tag{6.8}
\]

The first pair of equations corresponds to a noisy sensory mapping from hidden states and the equations of motion for states in the real world. These pertain to real-world variables representing the position of the target and of the eye (in boldface). The remaining equations constitute the generative model of how sensations are generated using the form of Equation 6.2. These define the free energy in Equation 6.1 – and specify behaviour under active inference. The variables constitute the first layer of the hierarchical model (see Equation 6.2, but for simplicity, we have written \( x \) instead of \( x^{(1)} \) and \( x \) instead of \( x^{(1)} \).

The real-world provides sensory input in two modalities: proprioceptive input from cranial nerve nuclei reports the angular displacement of the eye \( s_o \in \mathbb{R}^2 \) and corresponds to the centre of gaze. Note that, using the approximation of relatively small displacements, we use Cartesian coordinates to follow previous treatments e.g \cite{Fri+10b}. However visual space is better described by bounded polar coordinates, and treatments of large eye movements should account for this. Exteroceptive (retinal) input reports the angular position of a target in a retinal (intrinsic) frame of reference \( s_t \in \mathbb{R}^2 \). The indices \( o \) and \( t \) thus refer to states of the oculomotor system or of the target, respectively. Note that \( s_t \) is just the difference between the centre of gaze and target location in an extrinsic frame of reference \( x_t - x_o \). In this chapter, we are modelling the online inference of target position, and we are ignoring the problem of how the causal structure of the environment is learned. We simply assume that this structure has already been learned accurately, and therefore the dynamics of the real-world and the generative model are the same. Clearly, this model of visual processing is an enormous simplification: we are assuming that place coded spatial information can be summarised in terms of displacement vectors. However, more realistic simulations – using a set of retinotopic inputs with classical receptive fields covering visual space – produce virtually the same results. We will use more realistic models in future publications that deal with smooth pursuit and
visual occlusion. Here, we use the simpler formulation to focus on delays and the different sorts of generative models that can provide top-down or extra-retinal constraints on visual motion processing.

The hidden states of this model comprise the true, real-world oculomotor displacement \( \mathbf{x}_o \in \mathbb{R}^2 \) and target location \( \mathbf{x}_t \in \mathbb{R}^2 \). The units of angular displacement are arbitrary, but parameters are tuned to correspond to a small displacement of 4 degrees of visual angle for one arbitrary unit (that is approximately 4 times the width of a thumb’s nail at arm’s length). The oculomotor state is driven by action with a time constant of \( t_a = 64 \text{ ms} \) and decays (slowly) to zero through damping, with a time constant of \( t_o = 512 \text{ ms} \). The target location is perturbed by hidden causes \( \mathbf{x}_t \in \mathbb{R}^2 \) that describe the location to which the target is drawn, with a time constant of \( t_m = 16 \text{ ms} \).

In this chapter, the random fluctuations on sensory input and on the motion of hidden states are very small, with a log precision of 16. In other words, the random fluctuations have a variance of \( \exp(-16) \). This completes our description of the process generating sensory information; in which hidden causes force the motion of a target location and action forces oculomotor states. Target location and oculomotor states are combined to produce sensory information about the target in an intrinsic frame of reference.

The generative model has exactly the same form as the generative process but with one important exception: there is no action and the motion of the hidden oculomotor states is driven by the displacement between the target location and the central gaze (with a time constant of \( t_s = 32 \text{ ms} \)). In other words, the agent believes that its gaze will be attracted to the location of the target, which, itself, is being driven by some unknown exogenous force or hidden cause. The log-precisions on the random fluctuations in the generative model were four, unless stated otherwise. This means that uncertainty about sensory input, (motion of) hidden states and causes were roughly equivalent.

Having specified the generative process and model, we can now solve the active inference scheme in Equation 6.1 and examine its behaviour. Sensorimotor delays are implemented in the message passing from the generative process to the generative model. This generative model produces pursuit initiation because it embodies prior beliefs that the centre of gaze will follow the target location. This pursuit initiation rests on conditional expectations about the target location in extrinsic coordinates and the state of the oculomotor plant, where the location is driven by hidden causes that also have to be inferred.

The generative model described in this section provides the equations required to simulate active inference using the formalism of the previous section. In short, we now consider the generative model that defines the variational free energy and (Bayes) optimal active inference.
Figure 6.4: This figure reports the conditional estimates of hidden states and causes during the simulation of pursuit initiation, using a single rightward (positive) sweep of a visual target, while compensating for sensory motor delays. We will use the format of this figure in subsequent figures: the upper left panel shows the predicted sensory input (coloured lines) and sensory prediction errors (dotted red lines) along with the true values (broken black lines). Here, we see horizontal excursions of oculomotor angle (upper lines) and the angular position of the target in an intrinsic frame of reference (lower lines). This is effectively the distance of the target from the centre of gaze and reports the spatial lag of the target that is being followed (solid red line). One can see clearly the initial displacement of the target that is suppressed after a few hundred milliseconds. The sensory predictions are based upon the conditional expectations of hidden oculomotor (blue line) and target (red line) angular displacements shown on the upper right. The grey regions correspond to 90% Bayesian confidence intervals and the broken lines show the true values of these hidden states. One can see the motion that elicits following responses and the oculomotor excursion that follows with a short delay of about 64 ms. The hidden cause of these displacements is shown with its conditional expectation on the lower left. The true cause and action are shown on the lower right. The action (blue line) is responsible for oculomotor displacements and is driven by the proprioceptive prediction errors.
Figure 6.5: This figure illustrates the effects of sensorimotor delays on pursuit initiation (red lines) in relation to compensated (optimal) active inference – as shown in the previous figure (blue lines). The left panels show the true (solid lines) and estimated sensory input (dotted lines), while action is shown in the right panels. Under pure sensory delays (top row), one can see clearly the delay in sensory predictions, in relation to the true inputs. The thicker (solid and dotted) red lines correspond respectively to (true and predicted) proprioceptive input, reflecting oculomotor displacement. The middle row shows the equivalent results with pure motor delays and the lower row presents the results with combined sensorimotor delays. Of note here is the failure of optimal control with oscillatory fluctuations in oculomotor trajectories, which become unstable under combined sensorimotor delays.
6.4.2 Simulations

All simulations were performed with a time bin of 16ms and we report results in milliseconds. All results were replicated with different time bins (16ms, 8ms, 4ms, 2ms, 1ms) with minimal changes to the results. Figure 6.4 reports the conditional estimates of hidden states and causes during the simulation of pursuit initiation, using a simple rightward sweep of a visual target and compensating for sensorimotor delays: $\tau_s = \tau_{sa}$ and $\tau_a = \tau_{sa}$. This compensation is effectively the same as simulating responses in the absence of delays – because the delay operators reduce to the identity matrix. Target motion was induced using a hidden cause that was a ramp function of post-stimulus time. Note that ramp stimuli are often used in psychophysics, and this generative model – using velocity in place of position — produces the same results in velocity space. Indeed, most models, such as [RGG86] or [KL89], focus on modelling velocity responses. We choose to model the tracking of position for two reasons: First, it is easy to generalise position results to velocity using generalized coordinates of motion. Second, positional errors can induce slow eye movements [KS79; WP81] and we hoped to accommodate this in the model. If we assume that the units of angular displacement are 4 degrees of visual angle, then the resulting peak motion corresponds to about 20 degrees per second.

The upper left panel shows the predicted sensory input (coloured lines) and sensory prediction errors (dotted red lines) along with the true values (broken black lines). Here, we see horizontal excursions of oculomotor angle (upper lines) and the angular position of the target in an intrinsic frame of reference (lower lines). This is effectively the distance of the target from the centre of gaze and reports the spatial lag of the target that is being followed (solid red line). One can see clearly an initial retinal displacement of the target that is suppressed after approximately 20 ms. This effect confirms that the visual representation of target position is predictive and that the presentation of a smooth predictable versus an unpredictable target would induce a lag between their relative positional estimates, as is evidenced in the flash-lag effect [Nij94].

The sensory predictions are based upon the conditional expectations of hidden oculomotor (blue line) and target (red line) angular displacements shown on the upper right. The grey regions correspond to 90% Bayesian confidence intervals and the broken lines show the true values. One can see clearly the motion that elicits pursuit initiation responses, where the oculomotor excursion follows with a short delay of about 64 ms. The hidden cause of these displacements is shown with its conditional expectation on the lower left. The true cause and action are shown on the lower right. The action (blue line) is responsible for oculomotor displacements and is driven by proprioceptive prediction errors. Action does not return to zero because the sweep is maintained at an eccentric position during this simulation. This
eye position slightly undershoots the target position: it is held at around 95% of the target eccentricity in the upper right panel. Note that this corresponds roughly to the steady-state gain observed in behavioural data, which was modelled explicitly by [RGG86]. For our purposes, these simulations can be regarded as Bayes optimal solutions to the pursuit initiation problem, in which sensorimotor delays have been accommodated (discounted) via absorption into the generative model. We can now examine the performance in the absence of compensation and see how sensory and motor delays interact to confound pursuit initiation:

The above simulations were repeated with uncompensated sensory delays ($\tau_s = 0$ ms and $\tau_s = 32$ ms), uncompensated motor delays ($\tau_a = 0$ ms and $\tau_a = 32$ ms) and combined sensorimotor delays of 64 ms ($\tau_a = \tau_s = 0$ ms and $\tau_a = \tau_s = 32$ ms). To quantify behaviour, we focus on the sensory input and underlying action. The position of the target in intrinsic coordinates corresponds to spatial lag and usefully quantifies pursuit initiation performance. Figure 6.5 shows the results of these three simulations (red lines) in relation to the compensated (optimal) active inference shown in the previous figure (blue lines). True sensory input corresponds to solid lines and its conditional predictions to dotted lines. The left panels show the true and predicted sensory input, while action is shown in the right panels. Under pure sensory delays (top row) one can see the delay in sensory predictions, in relation to the true inputs. The thicker (solid and dotted) red lines correspond respectively to (true and predicted) proprioceptive input, reflecting oculomotor displacement. Crucially, in contrast to optimal control, there are oscillatory fluctuations in oculomotor displacement and the retinotopic location of the target that persists even after the target is stationary. These fluctuations are similar to the oscillations elicited by adding an artificial feed-back delay [GKL92]. Here, the fluctuations are caused by damped oscillations in action due to, and only to, sensory proprioceptive and exteroceptive delays. These become unstable (increasing in their amplitude) when the predicted value oscillates in counter-phase with the real value. Similar oscillations are observed with pure motor delays (middle row). However, here there is no temporal lag between the true and predicted sensations (solid versus dashed lines). Furthermore, there is no apparent delay in action – action appears to be emitted for longer, reaching higher amplitudes. In fact, action is delayed but the delay is obscured by the increase in the amplitude of action – that is induced by greater proprioceptive prediction errors. If we now combine both sensory and motor delays, we see a catastrophic failure of oculomotor tracking (lower row). With combined sensorimotor delays the pursuit initiation becomes unstable, with exponentially increasing oscillations as action over-compensates for delay-dependent errors.

In effect, the active inference scheme has undergone a phase transition from a stable to an unstable fixed point. We have illustrated this bifurcation
by increasing sensorimotor delays under a fixed motor precision or gain in Equation 6.7. The results in Figure 6.5 used a motor gain with a log precision of 2.5. We chose this value because it produced stable responses with sensory or motor delays alone and unstable dynamics with combined delays. These results illustrate the profound and deleterious effects of sensorimotor delays on simple pursuit initiation, using biologically plausible values – namely sensorimotor delays of 64 ms and a target velocity of about 16 degrees per second. This also illustrates the necessity of compensation for these delays so that the system can achieve a more robust and stable response. One would anticipate, in the face of such failures, real subjects would engage interceptive saccades to catch the target, of the sort seen in schizophrenic patients [Lev+93]. In the remainder of this chapter, we will concentrate on the nature of pursuit initiation and smooth pursuit with compensated sensorimotor delays, using a reasonably high motor gain with a log precision of four.

### 6.4.3 Pursuit initiation and visual contrast

Before turning to more realistic generative models of smooth pursuit, we consider the empirical phenomena in which following responses to the onset of target movement are suppressed by reducing the visual contrast of the target [Tho82]. In simulations of this sort, visual contrast is modelled in terms of the precision of sensory information in accord with Weber’s law – see [FF10] for details. Contrast-dependent effects are easy to demonstrate in the context of active inference. Figure 6.6 shows the spatial lag – the displacement in intrinsic coordinates of the target from the centre of gaze depicted by the solid red line in Figure 6.4 – as a function of contrast or log-precision of exteroceptive sensory input. The upper panel shows the true (solid lines) and predicted (dotted lines) spatial lag as a function of peristimulus time for different log precisions, ranging from two (low) to eight (high). The peak lags are plotted in the lower panel as a function of visual contrast or log precision. Since estimation error decreases as visual contrast increases, both curves converge, leading to a decrease to zero of the prediction error. These results show, in accord with empirical observations, how the spatial lag (position error) increases with contrast [AOR09], while the true lag decreases [Spe+05]. A similar difference between perception and action was recently reported [Sim+12]. The explanation for this contrast-dependent behaviour is straightforward – because pursuit initiation is based upon proprioceptive prediction errors, it depends upon precise sensory information. Reducing the precision of visual input – through reducing contrast – increases uncertainty about visual information (sensory estimation error) and places more weight on prior beliefs and proprioceptive sensations. This reduces the perceived motion of the target and reduces the amplitude of prediction errors driving action.
Figure 6.6: This figure reports the spatial lag (the displacement of the target from the centre of gaze) as a function of contrast (log precision of exteroceptive sensory input). The upper panel shows the true (solid lines) and predicted (dotted lines) spatial lag as a function of peristimulus time for different log precisions, ranging from two (black lines) to eight (red lines). The peak lags are plotted in the lower panel as a function of visual contrast or log precision. These results show how the perceived lag increases with contrast, while the true lag decreases in accord with empirical observations.
Figure 6.7: This figure uses the same format as Figure 6.4 – the only difference here is that the target motion is sinusoidal. The key thing to take from this simulation is that the peak spatial lag at the onset of the second cycle of target motion is greater than the peak lag at the onset of the first. This is contrary to empirical predictions.
6.4.4 Summary

In this section, we have seen that sensorimotor delays can have profound and deleterious effects on optimal oculomotor control. Here, optimal control means Bayes optimal active inference, in which pursuit initiation emerges spontaneously from prior beliefs about how a target attracts the centre of gaze. These simulations demonstrate that it is relatively easy to compensate for sensorimotor delays by exploiting representations in generalized coordinates of motion. Furthermore, the resulting scheme has some construct validity in relation to experimental manipulations of the precision or contrast of visual information. However, there are certain aspects of oculomotor tracking that suggest the pursuit initiation model above is incomplete: when presented with periodic target motion, the latency of motor gain (defined operationally in terms of the target and oculomotor velocities) characteristically reduces after the first cycle of target motion [BBC00]. This phenomenon cannot be reproduced by the pursuit initiation model above:

Figure 6.7 shows the responses of the pursuit initiation model to sinusoidal motion using the same format as Figure 6.4. Here, the hidden cause driving the target was a sine wave with a period of 512 ms that started after 256 ms. If we focus on the spatial lag (solid red line in the upper left panel), one can see that the lag is actually greater after one period of motion than at the onset of motion. This contrasts with empirical observations, which suggest that the spatial lag should be smaller after the first cycle [BBC00]. In the next section, we consider a more realistic generative model that resolves this discrepancy and takes us from simple pursuit initiation to smooth pursuit.

6.5 Results: smooth pursuit

In this section, we consider a slightly more realistic generative model that replaces the prior beliefs about the target attracting the centre of gaze with the belief that both the target and centre of gaze are attracted by the same (fictive) location in visual space. This allows pursuit initiation to anticipate the trajectory of the target and pursue the target more accurately – providing the trajectories are sufficiently smooth. The idea behind this generative model is to account for the improvements in tracking performance that are not possible at the onset of motion and that are due to inference on smooth target trajectories.

6.5.1 Smooth pursuit model

The smooth pursuit model considered in this chapter rests on a second-order generalisation of the pursuit initiation model of previous section. Previously, we have considered the motion of the oculomotor plant to be driven directly by action. This form of action can be considered as an (adiabatic) solution
to a proper second-order formulation, in which action exerts a force and thereby changes the angular acceleration of oculomotor displacement. This second-order formulation can be expressed in terms of the following generative process and model

\[
\mathbf{s} = \begin{bmatrix} \mathbf{s}_o \\ \mathbf{s}_t \end{bmatrix} = \begin{bmatrix} \mathbf{x}_o \\ \mathbf{x}_t - \mathbf{x}_o \end{bmatrix} + \omega^{(1)}
\]

\[
\mathbf{\dot{x}} = \begin{bmatrix} \mathbf{\dot{x}}_o \\ \mathbf{\dot{x}}'_o \\ \mathbf{\dot{x}}_t \end{bmatrix} = \begin{bmatrix} \frac{1}{t_o} a - \frac{1}{t_o} \mathbf{x}'_o \\ \frac{1}{t_m} (\nu^{(1)} - \mathbf{x}_t) \end{bmatrix} + \omega^{(1)}_x
\]

(6.9)

Here, the only thing that has changed is that we have introduced new hidden states corresponding to oculomotor velocity \( \mathbf{x}'_o \in \mathbb{R}^2 \). Action now changes the motion of the velocity (i.e., acceleration), as opposed to the velocity directly. This difference is reflected in the generative model but with one crucial addition – the hidden oculomotor state is not driven by the displacement between the target and the centre of gaze but by the displacement between the hidden cause and the centre of gaze. In other words, the hidden oculomotor states are attracted by the hidden cause of target motion – not the target motion \( \textit{per se} \). The idea here is that inference about the trajectory of the hidden cause should enable an anticipatory optimisation of pursuit initiation, provided these trajectories are smooth – hence a smooth pursuit model. Note that the equation of motion in the oculomotor model \( \mathbf{\dot{x}}_o = \frac{1}{t_o} (\mathbf{x}_t - \mathbf{x}_o) \) (see Equation 6.8) is the (adiabatic) solution to the equation used to model smooth pursuit: \( \frac{1}{t_v} (\nu^{(1)} - \mathbf{x}_o) - \frac{1}{t_v} \mathbf{x}'_o = 0 \) when \( \nu^{(1)} = \mathbf{x}_t \) (see Equation 6.9). As a result (and as confirmed by simulations) this model behaved similarly for the sweep stimulus used in Figures 6.4, 6.5, 6.6.

6.5.2 Simulations

We repeated the simulation reported in Figure 6.7 using the smooth pursuit generative model. The results of this simulation are shown in Figure 6.8 using the same format as Figure 6.7. The key difference – in terms of performance – is that the peak spatial lag after one cycle of motion is now less than the peak lag at the onset of motion. The response to the sinusoid trajectory
Figure 6.8: This figure uses the same format as the previous figure – the only difference here is that we have replaced the pursuit initiation model with a smooth pursuit model. In the smooth pursuit model the centre of gaze is attracted by a hidden cause of target motion, as opposed to the target \textit{per se}. Note that, in comparison to the previous figure, the peak lag at the onset of the second cycle of target motion is now smaller than at the onset to the first.
Figure 6.9: This figure uses the same format as the previous figure – the only difference is that the target motion has been rectified so that it is (approximately) hemi-sinusoidal. The thing to note here is that the improved accuracy of the pursuit previously apparent at the onset of the second cycle of motion has now disappeared – because active inference does not have access to the immediately preceding trajectory. This failure of an anticipatory improvement in tracking is contrary to empirical predictions.
contrasts with simple pursuit initiation and is more consistent with empirical observations. The true and expected hidden states show that the oculomotor trajectory now follows the target trajectory more accurately, particularly at the peaks of right and leftward displacement. Interestingly, the amplitude of action has not changed very much (compare Figures 6.7 and 6.8, upper right panels). However, action is initiated with a slightly shorter latency, which is sufficient to account for the improved pursuit when informed by the prior beliefs about the smooth trajectory of the target.

6.5.3 Summary

In summary, by simply replacing the target with the hidden cause of target motion – as the attractor of oculomotor trajectories – we can account for empirical observations of improved pursuit during periodic target motion. In the context of active inference, this smooth trajectory can only be recognised – and used to inform action – after the onset of periodic motion. However, this smooth pursuit model still fails to account for anticipatory effects that are not directly available in sensory trajectories. Empirical observations suggest that any systematic or regular structure in target motion can facilitate the accuracy of smooth pursuit, even if this information is not represented explicitly in target motion. A nice example of this rests on the use of rectified periodic motion, in which only rightward target excursions are presented. Experimentally, subjects can anticipate the periodic but abrupt onset of motion, provided they recognise the underlying periodic behaviour of the target. We can emulate this hemi-periodic motion by thresholding the hidden cause to suppress leftward deflections. Figure 6.9 shows the results of simulating smooth pursuit using the same format as Figure 6.8. The only difference here is that we replaced the sinusoidal hidden cause \( \nu(t) = \sin(2\pi f \cdot t) \) with \( \nu(t) = \exp(4(\sin(2\pi f \cdot t) - 1)) \). This essentially suppresses motion before rightward motion. This suppression completely removes the benefit of smooth pursuit after a cycle of motion – compare Figures 6.8 and 6.9. Here, the peak spatial lag at the onset of the second cycle of motion is exactly the same as the lag at the onset of motion; in other words, there is no apparent benefit of modelling the hidden causes of motion in terms of pursuit accuracy. This failure to model the anticipatory eye movements seen experimentally leads us to consider a full hierarchical model for anticipatory pursuit.

6.6 Results: anticipatory pursuit

This section presents a full hierarchical model of anticipatory smooth pursuit eye movements that tries to account for anticipatory oculomotor responses that are driven by extra-retinal beliefs about the periodic behaviour of targets. This entails adding a hierarchical level to the model that enables the agent
to recognise and remember the latent structure in target trajectories and suitably optimise its pursuit movements – which are illustrated here in terms of an improvement in the accuracy of target following after the onset of rectified target motion.

6.6.1 Anticipatory pursuit

The generative process used in these simulations is exactly the same as in the above (smooth pursuit) scheme (see Equation 6.9); however, the generative model of this process is equipped with an extra level in place of the model for the hidden cause of target motion in the generative model:

\[
\begin{align*}
\mathbf{s} &= \begin{bmatrix} s_o \\ s_t \end{bmatrix} = \begin{bmatrix} x_o \\ x_t - x_o \end{bmatrix} + \omega_\nu^{(1)} \\
\dot{x} &= \begin{bmatrix} \dot{x}_o \\ \dot{x}_t \end{bmatrix} = \begin{bmatrix} \frac{1}{t_e} (\nu^{(1)} - x_o) - \frac{1}{t_f} \dot{x}_o \\ \frac{1}{t_m} (\nu^{(1)} - x_t) \end{bmatrix} + \omega_x^{(1)} \\
\nu^{(1)} &= \sigma(x^{(2)}) + \omega_\nu^{(2)} \\
\dot{x}^{(2)} &= \begin{bmatrix} \dot{x}_1^{(2)} \\ \dot{x}_2^{(2)} \end{bmatrix} = \nu^{(2)} \begin{bmatrix} x_2^{(2)} \\ -x_1^{(2)} \end{bmatrix} + \omega_x^{(2)} \\
\nu^{(2)} &= \eta + \omega_\nu^{(3)}
\end{align*}
\]

(6.10)

The first level of the generative model is exactly the same as above. However, the hidden causes are now informed by the dynamics of hidden states at the second level. These hidden states model underlying periodic dynamics using a simple periodic attractor that produces sinusoidal fluctuations of any amplitude or phase and a frequency that is determined by a second level hidden cause with a prior expectation of a frequency of \(\eta\) (in Hz). It is somewhat similar to a control system model that attempted to achieve zero-latency target tracking by fitting the trajectory to a (known) periodic signal [BM83]. Our formulation ensures a Bayes optimal estimate of periodic motion in terms posterior beliefs about its frequency. In these simulations, we used a fixed Gaussian prior centred on the correct frequency with a period of 512 ms. This prior reproduces a typical experimental setting in which the oscillatory nature of the trajectory is known, but its amplitude and phase (onset) are unknown. Indeed, it has been shown that anticipatory responses are cofounded when randomising the inter-cycle interval [BF85]. In principle, we could have considered many other forms of generative model, such as models with prior beliefs about continuous acceleration [Ben+10].
As above, all the random fluctuations were assumed to have a log precision of four. Crucially, the mapping between the second level (latent) hidden states and the motion of first level hidden states encoding trajectories in visual (extrinsic) space is nonlinear. This means that latent periodic motion can be distorted in any arbitrary way. Here, we use a soft thresholding function \( \sigma(x) = \exp(4(x - 1)) \) to suppress negative (rightward) excursions of the target to model hemi-sinusoidal motion. This is the same function we used to generate the motion in Figure 6.9. Note that if the precision of the noise at the second level falls to zero and there is no (precise) information at this level, the generative model assumes that the random fluctuations have an infinite variance. As a consequence, the prediction at the level below in the hierarchical model simplifies to \( \nu^{(1)} = \omega^{(2)}_\nu \), and we recover Equation 6.9 describing the smooth pursuit model. As a consequence this parameter tunes the relative strength of anticipatory modulation.

Figure 6.10 shows the results of simulating active inference under this anticipatory model, using the same format as Figure 6.9. However, there is now an extra level of hidden states encoding latent periodic motion. It can be seen that expectations about hidden states attain nonzero amplitudes shortly after motion onset and are periodic thereafter. These provide predictions about the onset of rightward motion after the first (latent) cycle, enabling a more accurate oculomotor response. This is evidenced by the reduction in the spatial lag at the onset of the second cycle of motion, relative to the first (solid red lines on the upper left). This improvement in accuracy should be compared to the previous figure and reflects Bayes optimal anticipatory responses of the sort observed empirically [BBC00]. Further evidence of anticipatory inference can be seen by examining the conditional expectations about hidden causes at the second level. Note the substantial reduction in prediction error on the hidden cause (dotted red lines), when comparing the onset of the second cycle to the onset of the first. This reflects the fact that the conditional expectations about the hidden cause show a much-reduced latency at the onset of the second cycle due to top-down conditional predictions provided by the second level hidden states. This recurrent and hierarchically informed inference provides the basis for anticipatory oculomotor control and may be a useful metaphor for the hierarchical anatomy of the visual-oculomotor system.

### 6.6.2 Summary

In conclusion, to account for anticipatory pursuit movements that are not immediately available in target motion, one needs to equip generative models with a hierarchal structure that can accommodate latent dynamics – that may or may not be expressed at the sensory level. It is important to note that this model is a gross simplification of the complicated hierarchies that may exist in the brain. For instance, while some anticipation may
Figure 6.10: This figure uses the same format as the previous figure – the only difference is that the generative model has been equipped with a second hierarchical level that contains hidden states, modelling latent periodic behaviour of the (hidden) causes of target motion. With this addition, the improvement in pursuit accuracy apparent at the onset of the second cycle of motion is reinstated. This is because the model has an internal representation of latent causes of target motion that can be called upon even when these causes are not expressed explicitly in the target trajectory.
be induced in smooth pursuit eye movements, some aspects, such as the aperture problem, may not be anticipated [MSM06]. In this model, the second level hidden causes are simply driven by prediction errors and assume a constant frequency. As a consequence, prior beliefs about frequency are modelled as stationary. In the real brain one might imagine that models of increasing hierarchical depth might allow for nonstationary frequencies and other dynamics – that would better fit behavioural data. We have chosen to illustrate the basic ideas using a minimalistic example of anticipation in eye movements. Hierarchical extensions of this sort emphasise the distinction between visual motion processing and attending oculomotor control based purely upon retinal and proprioceptive input – they emphasise extra-retinal processing that is informed by prior experience and beliefs about the latent causes of visual input. We will exploit this anticipatory smooth pursuit model in future work, where visual occluders are used to disclose beliefs about latent motion.

6.7 Discussion

In this chapter, we have considered optimal motor control in the context of pursuit initiation and anticipatory smooth pursuit. In particular, we have taken a Bayesian perspective on optimality and have simulated various aspects of eye movement control using active inference. Active inference depends upon a generative model of stimulus trajectories and their active sampling through movement. This requires a careful consideration of the generative models that might be embodied by the visual-oculomotor system – and the sorts of behaviours one would expect to see under these models. The treatment in this chapter distinguishes between three levels of predictive coding with respect to oculomotor control: the first is at the lowest level of sensorimotor message passing between the sensorium and internal states representing the causes of sensory signals. Here, we examined the potentially catastrophic effects of sensorimotor delays and how they can easily render oculomotor tracking inherently unstable. This problem can be finessed – in a relatively straightforward way – by exploiting representations in generalized coordinates of motion. These can be used to offset both sensory and motor delays, using simple and neurobiologically plausible mixtures of generalized motion. We then motivated a model of smooth pursuit eye movements by noting that a simple model of target following cannot account for the improvement in visual tracking after the onset of smooth and continuous target trajectories. In this chapter, smooth pursuit was modelled in terms of hidden causes that attracted both the target and centre of gaze simultaneously – enabling the trajectory of the target to inform estimates of the hidden cause that, in turn, provide predictions about oculomotor consequences. While this extension accounted for experimentally observed tracking improvements
—under continuous trajectories — it does not account for anticipatory movements that have to accumulate information over time. This anticipatory behaviour could only be explained with a deeper hierarchical model that has an explicit representation of latent (periodic) structure causing target motion. When the generative model was equipped with a deeper structure, it was then able to produce anticipatory movements of the sort seen experimentally. Clearly, the simulations in this chapter are just heuristic and do not represent a proper simulation of neurobiological processing. However, they can be taken as proof of principle that the basic computational architecture — in terms of generalized representations and hierarchical models — can explain some important and empirical facts about eye movements. In what follows, we consider the models in this chapter in relation to other models and how modelling of this sort may have important implications for understanding the visual-oculomotor system.

6.7.1 Comparison with other models

The model that we have presented here speaks to and complements several existing models of the oculomotor system. First, it shares some properties with computer vision algorithms used for image stabilisation. Such models often use motion detection coupled with salient feature detection for the registration of successive frames [LK81]. A major difference is that these models are often applied to very specific problems or configurations for which they give an efficient, yet ad hoc solution. A more generic approach is to use - as our model does - a probabilistic method, for instance particle filtering [IB98]. Our model provides a constructive extension — as we integrate the dynamics of both sensation and action. In principle, this could improve the on-line response of feature tracking algorithms.

Second, using our modelling approach, we reproduce similar behaviours shown by other neuromimetic models of the oculomotor system. For example, the pursuit of a dot with known uncertainty can be modelled as the response of a Kalman filter [Kal60]. Both generalized Bayesian (active inference) and Kalman filtering predict the current state of the system using prior knowledge (about previous target locations) and refine these predictions using sensory data (prediction errors). This analogy with block-diagrams from control theory was first highlighted by [RGG86] and [KL89] — and has since been used widely [Gro+97]. For a recent treatment involving the neuromorphic modelling of cortical areas, see [Shi+05]. However, it should be noted that the link with Kalman filtering is rarely explicit (but see [Xiv+13]); most models have been derived heuristically, rather than as optimal solutions under a generative model. One class of such neuromimetic models uses neural networks that mimic the behaviour of the Kalman filter [Hay01]. This model was used to fit and predict the response of smooth pursuit eye movements under different experimental parameters [Mon+07b] or while
interrupting information flow [Bog+11]. Developing this methodology — and by analogy with modular control theory architectures — these building blocks can be assembled to accommodate increasingly complex behavioural tasks. This can take the form of a multi-layered model for transparency processing [RMN11] or of an interconnected graph connecting the form and motion pathways [BON08]. Such models have been used to understand adaptation to blanking periods and to tune the balance between sensory and proprioceptive inputs [MK03]. Our model is different in a key aspect: The Kalman filter is indeed the (Bayes) optimal solution under a linear generative model but a cascade of such solutions is not the optimal solution to (non-linear) hierarchical models [BF11]. The active inference approach considers the (embodied) system as a whole and furnishes an optimal solution in the form of generalized Bayesian filtering. In particular, given the delays at the sensory and motor levels, it provides an optimal solution that accommodates (or compensates for) these delays. As shown in the results, the ensuing behavior reproduces experimental results from pursuit initiation [MMI10] to anticipatory responses [Avi+06; BBC00]. The approach thus provides in inclusive framework, compared with heuristics used in neuromimetic models that focus on specific aspects of oculomotor control (see below).

The model presented here shares many features with other probabilistic models. First, representations are encoded as probability density is. This allows processing and control to be defined in terms of probabilistic inference; for instance, by specifying a prior belief that favours slow speeds [WSA02]. This approach has been successful in explaining a wide variety of physiological and psychophysical results. For example, it allows one to model spatial [PM07] or temporal [Mon+07b] integration of information, using conditional independence assumptions. Furthermore, recent developments have addressed the estimation of the shape and parameters of priors for slow speeds [SS06] and for the integration of ambiguous versus non-ambiguous information [BMM11]. The active inference scheme used here relies on generative models that entail exactly the same sorts of priors. It has also been shown that free energy minimisation extends the type of probabilistic models described above to encompass retinal stabilisation and oculomotor reflexes [Fri+10a]. A crucial difference here is that we have explicitly considered the problem of dynamics and delays. Our goal was to understand how the system could provide an optimal solution, when it knows (or can infer) the delay between sensing input (in the past) and processing information that informs action (in the future). This endeavour allowed us to build a model — using simple priors over the dynamics of the hidden causes — that reproduces the sorts of anticipatory behaviour seen empirically.
6.7.2 Limitations

Clearly, there are many aspects of oculomotor control we have ignored in this theoretical work. Foremost, we have used a limited set of stimuli to validate the model. Pursuit initiation was only simulated using a simple sweep of a dot, while smooth pursuit was studied using a sinusoidal trajectory. However, these types of stimuli are commonly used in the literature, as they best characterise the type of behaviour (following, pursuit) that we have tried to characterise: see [Bar08] for a review. We have not attempted to reproduce the oscillations at steady state as in [RGG86] or [GKL92], although this may help to optimise the parameters of our model in relation to empirical data. The hemi-sinusoidal stimulus is also a typical stimulus for studying anticipatory responses [Avi+06; BBC00]. Further validations of this model would call on a wider range of stimuli and consider and accumulated wealth of neurophysiological and behavioural data [Tla+10].

In this chapter, we have focused on inference under a series of generative models of oculomotor control. We have not considered how these models are acquired or learned. In brief, the acquisition of generative models and their subsequent optimisation in terms of their parameters (i.e., synaptic connection strengths) is an important, if distinct, issue. In the context of active inference, model acquisition and perceptual learning can be cast in terms of model selection and parameter optimisation through the minimisation of free energy. Under certain simplifying assumptions, this learning reduces to associative plasticity. A discussion of these and related issues can be found in Friston [Fri08].

The generative model used in this chapter has no explicit representation of space but only the uncertain, vectorial position of a target. We have previously studied the role of prediction in solving problems that are associated with the detection of motion using a dynamical and probabilistic model of spatial integration [PM12b]. Both that model and the current model entertain a similar problem: that of the integration of local information into a global percept, in both the temporal (this manuscript) and spatial [PM12b] domains. We have considered integrating sensory information in the spatial domain: terms of the prediction of sensory causes and their sampling by saccades [FTC12], and of the effects on smooth pursuit of reducing the precision. This manipulation can account for several abnormalities of smooth pursuit eye movements typical of schizophrenia [APF12]. In this chapter, we have limited ourselves to integrating information over time. It would be nice, in the future, to consider temporal and spatial integration simultaneously.

A final limitation of our model is the simplified modelling of the physical properties of the oculomotor system—due to the biophysics of the eyes and photoreceptors, sensory input contains motion streaks that can influence the detection of motion [BO04]. Furthermore, we have ignored delays in neuronal message passing among and within different levels of the hierarchy:
for a review of quantitative data from monkeys see [SB95]. Finally, we have not considered in any depth the finer details of how predictive coding or Bayesian filtering might be implemented neuronally. It should be noted that predictive coding in the cortex was attended by some early controversies; for example, paradoxical increases in visual evoked responses were observed when prediction error should be minimal. For example, a match between sensory signals and descending predictions can lead to the enhancement of neuronal firing [RLS98]. The neuronal implementation assumed in our work (see 6.2) finesses many of these issues. In this (hypothetical) scheme, predictions and prediction errors are encoded by the neuronal activity of deep and superficial pyramidal cells respectively (Mumford 1992; Bastos et al. 2012). In this scheme, the enhancement of evoked responses is generally thought to reflect attentional gain, which corresponds to the optimization of the expected precision (inverse variance) of prediction errors, via synaptic gain control [FF10]. Put simply, attention increases the gain of salient or precise prediction errors that the predictions are trying to suppress. Indeed, the orthogonal effects of expectations and attention in predictive coding have been established empirically using fMRI [Kok+11]. See Bastos et al. (2012) for a review of the anatomical and electrophysiological evidence that is consistent with the scheme used here.

6.7.3 Perspectives

Notwithstanding the limitations above, this approach may provide some interesting perspectives on neural computations in the oculomotor system. First, the model presented here can be compared to existing models of the oculomotor system. In particular, any commonalities of function suggest that extant neuromimetic models may be plausibly implemented using a generic predictive coding architecture. Second, the Bayes optimal control solution rests on a computational (anatomical) architecture that can be informed by electrophysiological or psychophysical studies. For example, we have considered only delays at the motor and sensory level. However, delays in axonal conduction between hierarchical levels – within the visual-oculomotor system – may have implications for intrinsic and extrinsic connectivity: in visual search, predictions generated in higher areas (say supplementary and frontal eye fields) may exploit a shorter path, by stimulating the actuator to sample more information (by making an eye movement) rather than accumulating evidence by explaining away prediction errors in lower (striate and extrastriate) cortical levels [MMI10]. By studying the structure of connections implied by theoretical considerations (see Figure 6.3), our modelling approach could provide a formal framework to test these sorts of hypotheses. A complementary approach would be to apply dynamic causal modelling [FHP03] to electrophysiological data, using predictive coding architectures, such that
transmission delays (and their compensation or modeling) among levels of
the visual-oculomotor system could be evaluated empirically.

Second, this work may provide a new perspective for experiments, in
particular for the generation of stimuli. We have previously considered
such a line of research by designing naturalistic, texture-like pseudo-random
visual stimuli to characterise spatial integration during visual motion detec-
tion [San+12]. We were able to show that the oculomotor system exhibits
an increased following gain, when stimuli have a broad spatial frequency
bandwidth. Interestingly, the velocities of these stimuli were harder to
discriminate relative to narrow bandwidth stimuli – in a two alternative
forced-choice psychophysical task [Sim+12]. In this work, the authors used
competitive dynamics based on divisive normalisation. Moreover, textured
stimuli were based on a simple forward model of motion detection [San+12].
This may call for the use of more complex generative models to generate such
textures. In addition, the use of gaze contingent eye-tracking systems allows
real-time manipulation of the configuration (position, velocity, delays) of the
stimulus, with respect to eye position and motion. By targeting different
sources of uncertainty, at the different levels of the hierarchical model, one
might be able to get a better characterisation of the oculomotor system.

The confounding influence of delays inherent in neuronal processing is
a strong biophysical constraint on neuronal dynamics. Representations in
generalized coordinates of motion provide a potential resolution that may
have enjoyed positive evolutionary pressure. However, it remains unclear
how neural information, represented in a distributed manner across the
nervous system, is integrated with exteroceptive, operational time. The
“binding” of different information, without a central clock, seems essential,
but the correlate of such a temporal representation of sensory information
(independent of delays) has never been observed explicitly in the nervous
system. Elucidating the neural representation of temporal information would
greatly enhance our understanding of both neural computations themselves
and our interpretation of measured electromagnetic (EEG and MEG) signals
that are tightly coupled to those computations.

Appendix

Appendix 1: Variational free energy

Here, we derive various formations of free-energy and show they relate to
each other. We start with the quantity we want to bound and implicitly
minimise — namely, surprise or the negative log-evidence associated with
sensory states \( \hat{s}(t) \) that have been caused by some unknown quantities \( \Psi(t) \).
These hidden causes correspond to the (generalized) motion (that is, position,
velocity, acceleration, ...) of a target that the oculomotor system is tracking.
\[-\ln p(\tilde{s}) = -\ln \int p(\tilde{s}, \Psi) d\Psi \quad (6.11)\]

We now simply add a non-negative cross-entropy or divergence between some arbitrary (conditional) density \( q(\Psi) = q(\Psi|\tilde{\mu}) \) and the posterior density \( p(\Psi|\tilde{s}) \) to create a free-energy bound on surprise

\[
F = -\ln p(\tilde{s}) + \int q(\Psi) \ln \frac{q(\Psi)}{p(\Psi|\tilde{s})} d\Psi
= -\ln p(\tilde{s}) + D(q(\Psi)||p(\Psi|\tilde{s})) \quad (6.12)
\]

The cross entropy term is non-negative by Gibb’s inequality. Because surprise depends only on sensory states, we can bring it inside the integral and use \( p(\tilde{s}, \Psi) = p(\Psi|\tilde{s})p(\tilde{s}) \) to show free-energy is a Gibb’s energy \( G = -\ln p(\tilde{s}, \Psi) \) expected under the conditional density minus the entropy of the conditional density

\[
F = \int q(\Psi) \ln \frac{q(\Psi)}{p(\Psi|\tilde{s})p(\tilde{s})} d\Psi
= \int q(\Psi) \ln \frac{q(\Psi)}{p(\Psi, \tilde{s})} d\Psi
= -\int q(\Psi) \ln p(\Psi, \tilde{s}) d\Psi + \int q(\Psi) \ln q(\Psi) d\Psi \quad (6.13)
\]

This is a useful formulation because it can be evaluated in a relatively straightforward way given a probabilistic generative model \( p(\tilde{s}, \Psi) \). A final rearrangement, using \( p(\tilde{s}, \Psi) = p(\tilde{s}|\Psi)p(\Psi) \), shows free-energy is also complexity minus accuracy, where complexity is the divergence between the recognition density \( q(\Psi) \) and the prior density \( p(\Psi) \)

\[
F = \int q(\Psi) \ln \frac{q(\Psi)}{p(\Psi|\tilde{s})p(\tilde{s})} d\Psi
= -\int q(\Psi) \ln p(\tilde{s}|\Psi)p(\tilde{s}) + D(q(\Psi)||p(\Psi)) \quad (6.14)
\]

6.7.4 Appendix 2: The maximum entropy principle and the Laplace assumption

If we admit an encoding of the conditional density up to second order moments, then the maximum entropy principle [Jay57], implicit in the definition of free energy above, requires \( q(\Psi|\tilde{\mu}) = \mathcal{N}(\tilde{\mu}, \Sigma) \) to be Gaussian. This is because a Gaussian density has the maximum entropy of all forms
that can be specified with two moments. Assuming a Gaussian form is known as the Laplace assumption and enables us to express the entropy of the conditional density in terms of its first moment or expectation. This follows because we can minimise free energy with respect to the conditional covariance as follows:

\[
F = G(\tilde{s}, \tilde{\mu}) + \frac{1}{2} \text{tr}(\Sigma \partial_{\tilde{\mu}} G) - \frac{1}{2} \ln |\Sigma|
\]

\[
G = -\ln p(\tilde{s}, \Psi)
\]

\[
\partial_{\Sigma} F = \frac{1}{2} \partial_{\tilde{\mu}} G - \frac{1}{2} \Pi
\]

so that \( \partial_{\Sigma} F = 0 \) implies

\[
\Pi = \partial_{\tilde{\mu}} G
\]

\[
F = G(\tilde{s}, \tilde{\mu}) + \frac{1}{2} \ln |\partial_{\tilde{\mu}} G|
\] (6.15)

Here, the conditional precision \( \Pi(\tilde{s}, \tilde{\mu}) \) is the inverse of the conditional covariance \( \Sigma(\tilde{s}, \tilde{\mu}) \). In short, free energy is a function of generalized conditional expectations and sensory states.

Appendix 3: Integrating or solving active inference schemes using generalized descents.

Given a generative model or its associated Gibbs energy function, one can now simulate active inference by solving the following set of ordinary differential equations for a system that includes generalized real-world states and internal states of the agent mediating (delayed) action and perception:

\[
\dot{\tilde{u}} = \begin{bmatrix}
\dot{\tilde{s}} \\
\dot{\tilde{x}} \\
\dot{\tilde{\nu}} \\
\dot{\tilde{\omega}} \\
\dot{\tilde{s}} \\
\dot{\tilde{\mu}} \\
\dot{\tilde{\eta}} \\
\dot{\tilde{\alpha}}
\end{bmatrix} = \begin{bmatrix}
\mathcal{D}\tilde{\vartheta}(\tilde{x}, \tilde{\nu}, \tilde{\alpha}) + \mathcal{D}\tilde{\omega}_\nu \\
\mathcal{f}(\tilde{x}, \tilde{\nu}, \tilde{\alpha}) + \tilde{\omega}_x \\
\mathcal{D}\tilde{\nu} \\
\mathcal{D}\tilde{\omega}_\nu \\
\mathcal{D}\tilde{\omega}_x \\
\mathcal{D}\tilde{\mu} - \partial_{\tilde{\mu}} F(T(\tau_s - \tau_\alpha)\tilde{s}, \tilde{\mu}) \\
\mathcal{D}\tilde{\eta} \\
-\partial_\alpha F(T(\tau_s - \tau_\alpha + \tau_\alpha - \tau_a)\tilde{s}, T(\tau_\alpha - \tau_a)\tilde{\mu})
\end{bmatrix}
\] (6.17)

generalized action \( \tilde{\alpha}(t) \) is approximated using discrete values of \( \alpha(t) \) from the past. Note that we have included a prior expectation \( \tilde{\eta}(t) \) of hidden causes to complete the agent’s generative model of its world. Integrating or solving Equation 6.17 corresponds to simulating active inference. The updates of the collective states over time steps of \( \Delta t \) use a local linearisation scheme \[Oza92\]:

162
\[ \Delta u = (\exp(\Delta t \cdot \partial_u \dot{u}) - I)(\partial_u \dot{u})^{-1} \]
\[ \partial_u \dot{u} = \begin{bmatrix}
0 & D \partial_x \tilde{g} & D \partial_y \tilde{g} & D & \ldots & D \partial_a \tilde{g} \\
\partial_x \tilde{f} & \partial_y \tilde{f} & I & \partial_a \tilde{f} \\
\vdots & D & \vdots & \vdots & \vdots & \vdots \\
D & \ldots & D & \ldots & \ldots & \ldots \\
-\partial_{j\eta} F & \ldots & -D \partial_{j\eta} F & -\partial_{j\eta} F & -\partial_{j\eta} F \\
-\partial_{a\eta} F & -\partial_{a\eta} F & -\partial_{a\eta} F & -\partial_{a\eta} F & -\partial_{a\eta} F
\end{bmatrix} \]

(6.18)

Details about how to compute the gradients and curvatures pertaining to the conditional expectations can be found in [Fri+10b]. These are generally cast in terms of prediction errors using straightforward linear algebra. Because action can only affect free-energy through the sensory states, its dynamics are prescribed by the following gradients and curvatures (ignoring higher-order terms):

\[ \partial_a F = (\partial_a \tilde{z}^{(1)}_\nu) \cdot \Pi^{(1)}_a T(\tau_a - \tau_a) \tilde{z}^{(1)}_\nu \]
\[ \partial_{aa} F = (\partial_a \tilde{z}^{(1)}_\nu) \cdot \Pi^{(1)}_a T(\tau_a - \tau_a)(\partial_a \tilde{z}^{(1)}_\nu) \]

(6.19)

The partial derivative of the sensory states with respect to action and is specified by the generative process. In biologically plausible instances of this scheme, this derivative would have to be computed on the basis of a mapping from action to sensory consequences. It is generally assumed that agents are equipped epigenetically, because it has a simple form. For example, contracting a muscle fibre elicits a proprioceptive stretch signal in a one-to-one fashion. The precision matrix \( \Pi^{(1)}_a \) in Equation 6.19 is specified such that only proprioceptive prediction errors with these simple forms have nonzero precision. This can be regarded as the motor gain in response to proprioceptive prediction errors. Equation 6.18 may look complicated but can be evaluated automatically using numerical derivatives for any given generative model. All the simulations in this chapter used just one routine — toolbox/DEM/spm_ADEM.m — and summarised in the script toolbox/DEM/ADEM_oculomotor_delays.m. Both are available as part of the SPM software (http://www.fil.ion.ucl.ac.uk/spm).

163
Part III

Perspectives
Chapter 7

Perspectives de recherche

Comme exposé dans l’introduction (chapitre 2) et détaillé dans les chapitres précédents, j’ai développé depuis mon intégration au CNRS une thématique étendue dans le domaine de l’étude du système oculomoteur en la focalisant sur l’axe du codage prédictif. Nous allons maintenant essayer de dégager les perspectives de recherche qui vont structurer mon programme de recherche futur.

7.1 Applications et recherche translationnelles

Comme nous l’avons vu, en parallèle de l’aspect fondamental de ces recherches, nous développons de nombreuses applications pratiques. Un premier aspect est l’utilisation d’algorithmes neuromimétiques pour le traitement de l’image et du signal. Nous l’avons vu par exemple au chapitre 4 dans lequel nous avons montré qu’un algorithme inspiré de l’architecture des connections dans l’aire visuelle primaire permet de générer un algorithme efficace de catégorisation entre des classes a priori séparées par des concepts de haut niveau (animal / non-animal). Nous avons développé —en collaboration avec Gabriel Cristobal (Instituto de Optica, Madrid)— la mise en pratique de tels algorithmes pour la détection de tumeurs pulmonaires (empyèmes) [Nav+13]. De la même façon, les algorithmes présentés dans le chapitre 3 permettent l’amélioration d’algorithmes de suivi —même perturbés par une occlusion— et nous avons proposé une implémentation rapide de cet algorithme utilisant une méthode de filtre particulière. De surcroît, nous avons proposé, grâce à une collaboration avec l’équipe d’Anders Lansner au KTH (Stockholm), une implémentation sous forme d’un réseau de neurones dans le chapitre 5. Dans le cadre du consortium BrainScaleS, en collaboration avec l’équipe de Karl-Heinz Meier (UHEI, Heidelberg), nous développons la mise en place de tels algorithmes sur des circuits intégrés qui permettront de multiplier la vitesse de tels algorithmes (et par conséquent leurs capacités de traitement et d’apprentissage) par un facteur $\times 10000$. 

165
Un second aspect de l’amélioration de nos connaissances des mécanismes computationnels du cerveau grâce à des modèles est la possibilité de concevoir des outils de diagnostic. Cette aide au diagnostic est un problème de société de premier plan face au vieillissement de la population et de la rationalisation des capacités de médicalisation. La prévention devient alors un outil essentiel pour les années futures et nous étudions plusieurs pistes pour conduire ces études. Nous avons évoqué dans la section 2.4.4 comment nous avons appliqué le principe de MEL à une meilleure compréhension de la schizophrénie [APF12]. Il s’est alors dégagé qu’une origine possible des désordres liés à cette psychopathologie sont liés à un dysfonctionnement neuro-modulateur. Des tests sont actuellement menés pour appliquer cette hypothèse sur des patients en collaboration avec le docteur Adams (Institut de neurologie de Londres). Une autre piste parallèle que nous explorons actuellement en collaboration avec Bruno Wicker (équipe Scalp, INT) est de construire des outils efficaces de diagnostic de patients autistes. En particulier, le principe de MEL permet de dégager des hypothèses sur les origines possibles de l’autisme [FLF13] et du lien possible avec le traitement des émotions [JC13]. Nous construisons des protocoles expérimentaux adaptés à la translation à des patients jeunes (sous forme de jeux vidéos) qui vont explicitement jouer sur les paramètres des modèles génératifs utilisés pour caractériser une boucle de perception/action.

7.2 Neurosciences computationnelles et échelles d’étude

Au cours des dix dernières années, le champ des neurosciences computationnelles a été profondément modifié: explosion du volume de données, de la quantité des connaissances et de modèles proposés. Afin de pouvoir prendre avantage de cette importante mutation, il a été nécessaire de créer des ponts, d’abord entre acteurs des neurosciences computationnelles. Cela a pris par exemple la forme de la création d’un langage commun de description pour la définition, la simulation et l’analyse de réseaux de neurones impulsionnels (PyNN - consortium FACETS) tel qu’il est décrit dans [Dav+08] (plus de 120 citations à ce jour). Cette entreprise a été étendue lors du consortium BrainScales grâce à l’initiative NeuralEnsemble qui regroupe des outils de développement. Ces outils sont développés en collaboration avec d’autres champs des neurosciences (notamment la physiologie). Un de ces outils, les Motion Clouds, sont exemplaires à plus d’un titre: En effet ils constituent à ma connaissance une première dans le champ de l’étude du système oculomoteur dans le sens où, à partir d’un modèle d’un mécanisme computationnel, une théorie de caractérisation du système est établie pour quantifier sa réponse de façon globale [San+12]. Cette approche n’est toutefois qu’un premier pas et nous travaillons à l’établissement de protocoles expérimentaux.
—en physiologie et psychophysique— qui étendent les Motion Clouds à l’étude de mécanismes plus complexes.

On peut donc maintenant affirmer que la classification proposée par Marr a volé en éclats et que les niveaux d’études —même s’ils existent toujours— sont plus perméables aux échanges. Un exemple d’une telle interpénétration est la construction de modèles de décodage de l’activité. Souvent basés sur des modèles computationnelles de l’activité neuronnaux, ceux-ci visent à comprendre de façon plus globale l’activité uninaire de neurones, par exemple comme le décodage au niveau d’une population de neurones. Nous développons actuellement de tels algorithmes en collaboration avec Giacomo Benvenuti et Frédéric Chavane (équipe inViBe, INT) sur des données collectées au laboratoire et chez Tony Movshon (NYU, New-York) par Frédéric Chavane. Ces modèles s’appliquent ensuite en retour à l’implémentation d’algorithmes de réseaux de neurones impulsionnels (voir chapitre 5) dans un dialogue qui vise à mieux comprendre ces systèmes complexes. Une direction de recherche à long terme est d’affiner ces algorithmes de décodage en dialogue entre modèles et neuro-physiologie.

7.3 Inférence active: de la prédiction à l’anticipation

À ce titre, le principe de MEL que nous avons exposé dans ce manuscrit constitue une base théorique essentielle pour consolider la compréhension des mécanismes neuronaux. En particulier, ce principe permet d’implanter des mécanismes de codage prédictif à différentes échelles, de l’aire corticale au système oculomoteur complet. Un objectif à moyen terme est d’unifier les différentes approches que nous avons exposées dans le domaine spatial (chapitre 4) ou spatio-temporel [PM12b] pour les étendre à une approche commune en terme de minimisation de l’énergie libre. Une telle approche nous permettra en particulier de mettre en synergie des sources d’information de modalités différentes. En effet, s’il est important de comprendre les aspects prédictifs isolément (d’un contour local en fonction du contexte voisin, d’un point le long d’un trajetatoire), il semble essentiel dans le futur de pouvoir en comprendre les interactions, comme par exemple le mouvement de contours le long de trajectoires complexes.

Un avantage du principe de MEL et de la formalisation proposée par Karl Friston est d’avoir défini une terminologie claire et univoque. Par exemple, alors que la prédiction est considéré comme un ajustement de l’état présent en fonction du contexte passé (en utilisant par exemple l’inférence statistique), le concept d’anticipation est alors défini de façon claire comme un changement prévu dans le futur de l’état du système. Le principe de MEL permet ainsi de dépasser des problèmes de dialogues entre certaines disciplines qui ont freiné le développement des approches computationnelles. En particulier, un
objectif de recherche à long terme est d’étudier la notion de temps dans les processus neuronaux et en particulier dans le cadre du système oculomoteur. En effet, nous avons vu précédemment dans le chapitre 6 qu’une difficulté majeure dans le système oculomoteur est l’existence de délais sensori-moteurs. Si nous avons considéré dans le modèle précédent des délais fixes, ceux-ci sont plus généralement variables, et de plus peuvent exister entre des niveaux de traitement différents (ce que nous avons omis dans le modèle précédemment). Prendre en compte ces délais en toute généralité permettrait de s’attaquer explicitement à un problème essentiel en neurosciences et qui est le plus souvent éluqué : l’activité que je mesure sur un neurone se réfère-t-elle au temps présent ? Au temps auquel j’ai reçu l’information sensorielle ? Au temps auquel j’anticipe de fournir l’action motrice ? Bien sûr, il semblerait que les temps auxquels se réfèrent chaque point d’activité neurale, vus de façon globale, constitue un continuum autour du temps présent et permette de situer l’agent dans un contexte temporel. Ce contexte constitue alors la base de mécanismes essentiels et pour lesquelles nous ne connaissons encore de réponse, comme les mécanismes de mémoire, ou la coordination des flux d’information dans l’ensemble du système nerveux central.
Bibliography


[Ben+07] S. J. Bennett et al. “Target acceleration can be extracted and represented within the predictive drive to ocular pursuit”. In: *Journal of Neurophysiology* 98.3 (Sept. 1, 2007), pp. 1405–1414 (cit. on p. 124).


Yoonsuck Choe and Risto Miikkulainen. “Contour integration and segmentation with self-organized lateral connections”. In: *Biological Cybernetics* 90.2 (Feb. 1, 2004), pp. 75–88 (cit. on p. 75).

Sung-Hyuk Cha and Sargur N. Srihari. “On measuring the distance between histograms”. In: *Pattern Recognition* 35.6 (June 2002), pp. 1355–1370 (cit. on pp. 78, 80, 88).


[Kre+07a] Jens Kremkow et al. “Emergence of population synchrony in a layered network of the cat visual cortex”. In: Neurocomputing 70.10-12 (June 2007), pp. 2069–2073 (cit. on p. 31).


Jean Lorenceau et al. “Different motion sensitive units are involved in recovering the direction of moving lines”. In: Vision Research 33.9 (June 1992), pp. 1207–1217 (cit. on p. 55).


[PBM06] Laurent U. Perrinet, Frédéric V. Barthélemy, and Guillaume S. Masson. “Input-output transformation in the visuo-oculomotor loop: modeling the ocular following response to center-surround stimulation in a probabilistic framework”. In: 1ère conférence francophone NEURoscienes COMPutationnelles - NeuroComp. 2006 (cit. on p. 28).


